Evaluating AI Speech and Text Technologies to Reduce the Administrative Burden in Occupational Health: An Exploratory Study on Whisper and ChatGPT-4

by

Youssef Al Ghouch

in partial fulfilment of the requirements of

Master of Science in Biomedical Engineering

Track: Medical Devices

at the Delft University of Technology,

to be defended publicly on Thursday, November 14, 2024, at 10:45 AM.

Student Number:	5654920	
Project Duration:	March, 2024 – November, 2024	
Faculty:	Mechanical Engineering	
Thesis advisors:	A. H. A. Stienen	TU Delft
	O. Scharenborg	TU Delft
	A. Saarig	Arbomaatschap B.V.
	A. R. Abdelgabar	Arbomaatschap B.V.
Thesis Committee:	A. H. A. Stienen	TU Delft
	O. Scharenborg	TU Delft
	H. E. J. Veeger	TU Delft

This thesis is confidential and cannot be made public until October 30, 2024

Evaluating AI Speech and Text Technologies to Reduce the Administrative Burden in Occupational Health: An Exploratory Study on Whisper and ChatGPT-4.

Youssef Al Ghouch
Faculty of Mechanical Engineering
(of Technical University Delft)
Delft, The Netherlands
& Arbomaatschap B.V. Amsterdam, The Netherlands

Odette Scharenborg
Faculty of Electrical Engineering, Mathematics and Computer Science
(of Technical University Delft)
Delft, The Netherlands

Arno H. A. Stienen
Faculty of Mechanical Engineering
(of Technical University Delft)
Delft, The Netherlands

Aimane Saarig
Arbomaatschap B.V.
Amsterdam, The Netherland

Abdel-Rahman Abdelgabar
Arbomaatschap B.V.
Amsterdam, The Netherlands

***Abstract— Objective*: The objective of this exploratory study is to investigate how AI speech and text technologies, specifically Whisper and ChatGPT-4, can help reduce the administrative burden in occupational health consultations, with a focus on accuracy, efficiency, and user satisfaction.**

***Methods*: A quantitative research approach was employed, utilizing a controlled trial design. Fourteen occupational health doctors participated in simulations using Whisper for transcription and customized ChatGPT-4 modules for generating medical summaries (Medical Summarizer), letters (Letter Generator), and documents (Document Generator), and providing medical protocol-based replies on participants' questions (Medical Protocol Assistant). The accuracy, efficiency, and user satisfaction of these technologies were compared against traditional administrative methods, with descriptive statistics and paired samples t-tests conducted to assess performance differences.**

***Findings*: The study revealed that Whisper transcriptions had a word error rate (WER) of 13.1%, with an average transcription time of 18.4 minutes, which was 6.6 minutes faster than human transcription. ChatGPT-4's Medical Summarizer generated summaries 29 times faster than human participants, with an average generation time of 0.5 minutes but had a 38.6% error rate in element generation. The Letter Generator and Document Generator exhibited error rates of 90.4% and 17.5%, respectively, although both were significantly more efficient than manual processes, with average generation times of 0.4 and 3.9 minutes, respectively. The Medical Protocol Assistant provided protocol-based replies with an 86.7% accuracy, achieving the highest user satisfaction score (4.4 out of 5) among all modules.**

***Conclusion*: AI speech and text technologies show potential in reducing administrative tasks in occupational health settings, particularly in terms of efficiency. However, the moderate accuracy and varying satisfaction rates indicate that further refinement is necessary to enhance their applicability in clinical practice. Future research should focus on improving accuracy, evaluation of the technologies in actual patient-physician consultations and developing robust privacy safeguards.**

Keywords— Artificial Intelligence, Occupational Medicine, Administrative Burden, Speech Recognition, Text Generation, User Satisfaction, Exploratory Study.

Funding— No funding was received in the context of this research.

Conflict of interest— YaG, AS and AR are employed by Arbomaatschap B.V. While the research was conducted independently, the authors acknowledge a potential conflict of interest due to their affiliation with the company. The authors declare no other conflicts of interest.

Précis— This exploratory study evaluates AI speech and text technologies, specifically Whisper and ChatGPT-4, in reducing administrative burdens in occupational health, finding that while these technologies increase efficiency, they require further refinement to enhance accuracy and user satisfaction.

Word count— 8,960 words (excluding Abstract, References and Appendices)

Number of pages— 21 pages

Number of tables— 4

Number of figures— 1

Appendices— 5

Supplemental materials— 2

Author Contributions— Concept and design: YaG, AHAS, OS, AR; Data collection: YaG; Analysis and interpretation of data: YaG; Drafting of the manuscript: YaG, AS; Critical revision of the paper for important intellectual content: YaG, AHAS, OS, AS, AR; Supervision: AHAS, OS, AR

I. Introduction

Healthcare workers are increasingly confronted with administrative demands which significantly contributes to their workload (Herd & Moynihan, 2021). Specifically, studies have found that 16.6% to 24% of medical doctors' working hours are spend on administrative tasks (Rao et al., 2017; Woolhandler & Himmelstein, 2014). Despite the substantial usage of time for these tasks, only 36% of the administrative work is considered beneficial for the quality of care (Zegers et al., 2022). This is particularly concerning since administrative burdens form a major factor that contributes to burnout and the turnover of healthcare workers (Swensen, Shanafelt & Mohta, 2016). Moreover, as the burden of administrative tasks grows, the time allocated to direct patient care diminishes, which may lead to a decline in the quality of medical services and reduced patient satisfaction (Wu, 2023).

This issue is especially relevant in the field of occupational medicine, where occupational physicians have identified administrative burdens as a significant source of their job dissatisfaction (Plomp & Van Der Beek, 2014). In the current Dutch context, this issue is exacerbated by the rising demand for occupational medicine services due to an aging population, which places an increasing burden on a relatively smaller workforce (Oude Mulders, Henkens & van Dalen, 2020). Additionally, while the demand for medical professionals remains high, fewer are being trained (Hingstman, Velden & Schepman, 2009). The combination of these factors underscores a growing need for solutions that can alleviate the administrative workload in occupational medicine.

To address the growing administrative burden within healthcare, increasing attention is being given to artificial intelligence (AI) as a potential solution, driven by its ability to automate processes and enhance efficiency (Davenport & Kalakota, 2019). This focus on AI depends on its capacity to handle huge quantities of data, execute daily tasks, and supporting decision-making, which can significantly reduce the time healthcare professionals spent on administrative duties. As healthcare providers seek to optimize their operations amidst rising demands and workforce constraints, AI is seen as a promising set of technologies that can enhance productivity and allow clinicians to focus more on patient care (Davenport & Kalakota, 2019; Spear, Ehrenfeld & Miller, 2023). Especially, speech and text technologies, including natural language processing (NLP) which, for instance, can facilitate real-time transcription of a conversation, can potentially be used to reduce the administrative burden. This can be achieved not only by enhancing the efficiency of administrative tasks through automation but also by improving accuracy through the reduction of human error (Kumar & Gond, 2023).

While AI speech and text technologies can potentially reduce administrative burdens within healthcare, much of the existing research is predominantly focused on its applications in clinical decision-making, with relatively little attention paid to its potential for administrative use (Al Ghouch & Stienen, 2024). In addition, most of the studies have solely focused on accuracy and efficiency as measurement of the technology's performance, while more user focused parameters, such as user's satisfaction, are infrequently investigated (Al Ghouch & Stienen, 2024). However, user satisfaction is a critical factor that significantly influences the actual adoption and effective utilization of these technologies (Wixom & Todd, 2005). Neglecting this aspect can undermine the potential benefits, as even highly accurate and efficient systems may fail to gain traction if they do not align with user needs and preferences (Wixom & Todd, 2005).

This article, therefore, sets out to investigate AI speech and text technologies with regards to the accuracy, efficiency, and user satisfaction in an exploratory study involving medical consultations between occupational physicians and employees experiencing health issues, i.e. patients. Specifically, this study explores the state-of-the-art AI speech and text technologies Whisper, an automatic speech recognition tool that can transcribe conversations real-time, and ChatGPT-4, a large language model (LLM) that can generate medical documentation and provide medical information, to address these challenges (Alto, 2023). Here the accuracy, efficiency and user satisfaction among users is evaluated by examining the application of Whisper for transcription tasks and ChatGPT-4 for generating medical documentations and providing document-based information. The following research questions guide this investigation:

Main Research Question: How do AI speech and text technologies, specifically Whisper and ChatGPT-4, perform across accuracy, efficiency, and user satisfaction in an exploratory study involving medical consultations between occupational physicians and employees?
Sub-question A: How effective and efficient is Whisper in transcribing consultations between occupational physicians and employees, and how satisfied are occupational physicians with this technology?
Sub-question B: How effective and efficient is ChatGPT-4 in generating medical reports, letters, and filling out forms during consultations between occupational physicians and employees, and how satisfied are users with this technology?
Sub-question C: How effective and efficient is ChatGPT-4 in providing protocol- and guideline-based information during consultations between occupational physicians and employees, and how satisfied are users with this technology?

II. Methodology

A. Study design overview

This study utilized a quantitative research approach with a controlled trial design. As displayed in Figure 1, medical doctors have conducted occupational health consultations with simulated patients based on ChatGPT-4 generated cases. Transcriptions of these consultations have been made using Whisper. The transcriptions were subsequently processed through customized ChatGPT-4 models, each developed to perform specific language processing tasks within the workflow. The first customized model, the 'Medical Summarizer', created a medical summary based on the transcription. The second model, the 'Letter Generator', generated a letter with the medical summary as its input. The third model, the 'Document Generator', created a document called the Functional Capabilities List (Dutch: Functionele Mogelijkheden Lijst; FML) based on the medical summaries and letters. The final model, the ‘Medical Protocol Assistant’, was employed to respond to case-related questions posed by the medical doctors. All these outputs, that is, the transcription, medical summary, letter, FML-document, and responses were compared with control data, generated by the medical doctor and the main researcher. The code and parameters used for applying Whisper and the end-result instructions for each customized ChatGPT-4 model can be found in Supplementary file 1. All models have been evaluated based on their accuracy, efficiency, and user satisfaction. The complete list of variables considered can be found in Table 1. Subsequently to the data evaluation, the dataset was inspected for any missing inputs. Descriptive statistics were then conducted to summarize the data, including means, standard deviations (SD), and percentages. Several statistical methods, including descriptive statistics and paired t-tests, were used to describe and test for differences in accuracy, efficiency and satisfaction variables between the control data and the AI speech and text technologies output. In all cases, a p-value of less than 0.05 was considered statistically significant. In the following paragraphs more details can be found regarding the participants, cases, Whisper, custom ChatGPT-4 models, and subgroup analyses.

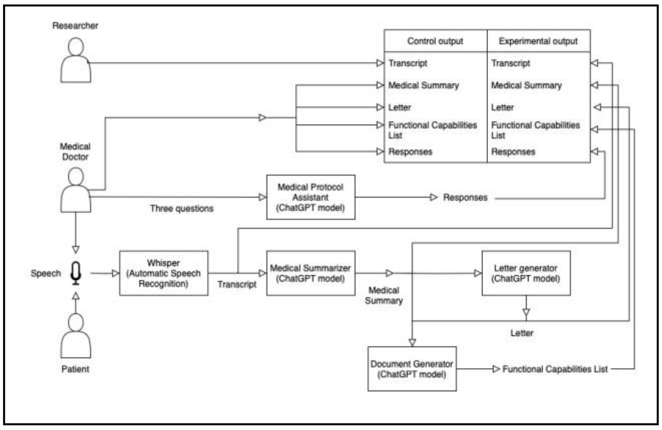
Researcher
Control output
Experimental output
Transcript
Medical Summary
Letter
Functional Capabilities List
Responses
Transcript
Medical Summary
Letter
Functional Capabilities List
Responses
Medical Doctor
Medical Protocol Assistant (ChatGPT model)
Three questions
Responses
Speech
Whisper (Automatic Speech Recognition)
Transcript
Medical Summarizer (ChatGPT model)
Medical Summary
Letter generator (ChatGPT model)
Letter
Patient
Document Generator (ChatGPT model)
Functional Capabilities List

Fig. 1. Flowchart of the AI-speech and text models and data processing

B. Participants

Simulated patients and actual occupational health doctors were used to create realistic medical scenarios. In total, 14 doctors participated in the study. Each participant interacted with all AI-models, and the doctors also acted as simulated patients. To mitigate the risk of bias due to familiarity with the cases, the doctors were given the subject of each case before participation. They were asked whether they had previously simulated this case, either as a doctor or a patient. If a doctor was familiar with a particular case, they were assigned a different case to ensure that no doctor assessed a scenario they had previously portrayed. All participants were given detailed instructions on how to conduct the consultations, use the AI models, and complete the tasks for control data. These instructions can be found in Appendix 2.

C. Cases

Cases were generated by ChatGPT 4 using a specific prompt (Appendix 3) to ensure details for both the patient and the doctor. For the patient, the prompt generated information including name, date of birth, age, gender, job role, weekly working hours, first day of sick leave, anamnesis, current symptoms, medication, treatment, medical history, exercise, stress factors, sleep and energy, daily routine, reintegration, and work factors. For the doctor, the generated information included the patient's name, date of birth, age, gender, and advice. These information categories were selected based on expert opinions from three physicians working in occupational medicine, who identified the categories they consider essential for conducting consultations. These categories reflect the information these doctors assess in their daily practice. In addition, the generation of advice was chosen to ensure that the doctor was always able to provide recommendations without needing to consult a supervisor.

TABLE I. LIST OF VARIABLES

	Variable
Speech-to-text converter	Word Error Rate
	Time needed to generate
	Doctor’s satisfaction rate
Medical report generator	Number of incorrectly generated elements
	Number of total elements in control report / letter
	Accuracy in generating elements
	Number of incorrectly categorized elements
	Accuracy in categorizing elements
	Readability-score
	Time needed to generate
	Doctor’s satisfaction rate
Letter generator	Number of incorrectly generated elements
	Number of total elements in control letter
	Accuracy in generating elements
	Number of incorrectly categorized elements
	Accuracy in categorizing elements
	Readability-score
	Time needed to generate
	Doctor’s satisfaction rate
Document generator	Number of incorrectly generated elements
	Number of total elements in control document
	Accuracy in generating elements
	Time needed to generate
Medical Protocol Assistant	Number of responses from an actual protocol
	Number of prompts needed
	Number of questions answered as desired
	Doctor’s relevancy score
	Time needed to generate
	Doctor’s satisfaction rate
Subgroup analyses	Whisper model used
	Microphone type

Diversity was sought in generating cases to ensure a varied set of scenarios, differing in aspects such as disease, gender, age, name, and job role. The content of the prompts was derived from the occupational medicine doctors who identified key variables that reflect real-world scenarios they encounter. The goal was to represent each available Dutch Association for Occupational Medicine guideline in at least one case (Dutch: Nederlandse Vereniging voor Arbeids- en

Bedrijfsgeneeskunde, NVAB) (Nederlandse Vereniging voor Arbeids- en Bedrijfsgeneeskunde, n.d.). Therefore, 28 cases were generated based on diseases covered by NVAB guidelines, and 2 additional cases were generated for diseases not covered by NVAB guidelines, resulting in a total of 30 cases, each simulated once. These two cases were utilized to assess how the final model, the Medical Protocol Assistant, would respond when no specific protocol is available. It was anticipated that the model would either be unable to answer the questions or refer to other protocols that might provide insights for addressing the questions.

D. *Whisper*

Whisper has been used to transcribe the simulated consultations. Whisper is an automatic speech recognition (ASR) model built by OpenAI that takes advantage of the enormous size of weakly supervised training on a dataset of 680,000 hours of multilingual audio from the internet (OpenAI, 2022; Radford et al., 2023). This enables the model to identify the spoken language, including Dutch, which was used in the simulations. As opposed to traditional speech recognition systems, which depend significantly on completely supervised data and require tailoring for individual tasks, Whisper is intended to generalize well among processes and languages without a requirement for tailoring. It accomplishes this by estimating transcripts directly from raw audio employing an encoder-decoder transformer model trained on various and noisy datasets, including languages other than English and translation tasks (Radford et al., 2023). In addition, Whisper has several model versions, with the medium model, while generally more efficient, exhibiting reduced accuracy compared to the large model (OpenAI, 2022). To compare the transcription of Whisper, control data has been created by the researcher through manually transcribing the consultations. The accuracy was assessed using the word-error-rate (WER), calculated as the sum of substitutions, insertions, and deletions divided by the number of words in the reference, multiplied by 100 (Yakubovskyi & Morozov, 2023) (Formula 1). Efficiency was measured by the time required to generate the transcription, and the doctor's satisfaction was recorded on a scale of 1 to 5, where 1 indicates very dissatisfied and 5 indicates very satisfied. Paired t-tests were conducted to determine whether the Whisper model produced transcriptions significantly faster or slower compared to human transcription.

$$\text{WER} = \frac{\text{Substitutions+Insertions+Deletions}}{\text{Number of Words in Reference}} * 100 \qquad (1)$$

E. *ChatCPT-4 Custom Model: Medical Summarizer*

ChatGPT-4 is a large language model which also uses an encoder-decoder method, in which the encoder interprets the input text to build a context representation, which is then used by the decoder to generate the matching output (Abdullah, Madain & Jararweh, 2022; Kernberg, Gold, & Mohan, 2024). ChatGPT-4 has been trained using a large amount of text data, which allows the model to acquire knowledge of language patterns, grammar, and semantics. The first training phase was unsupervised, during which the model learned to predict upcoming words in a sequence by evaluating vast amounts of text. After pre-training, the model was adjusted to add Reinforcement Learning from Human Feedback (RLHF). During this process, human testers offered input on the model's outputs, directing it to develop replies that tend to be accurate, contextually relevant, and consistent with human communication standards. The integration of unsupervised pre-training and supervised tailoring enabled ChatGPT-4 to perform NLP tasks, including having a dialogue, and it can be customized by adding documents to its knowledge base. This customization allows it to generate responses based on the content of the added documents during dialogues. However, it is notable that, to date, this functionality has not been extensively researched in health care applications (Abdullah, Madain & Jararweh, 2022; Kernberg, Gold, & Mohan, 2024). Therefore, in this research, ChatGPT-4 models were created by imputing instructions into the standard ChatGPT-4 model, such that it performs a specific administrative task. These instructions were created through trial-and-error until the point the main researcher did not perceive improvements of the administrative task results.

The first ChatGPT-4 custom model, the ‘Medical Summarizer’, has been designed to generate structured medical summaries from the consultation transcriptions. This model extracts and categorizes information into predefined categories, namely the reason for reporting sick, first day of sick leave, anamnesis, current symptoms, treatment, medication, medical history, lifestyle factors (including exercise, stress, sleep, substance use, and daily routine), work history (focusing on reintegration and work factors), conclusion, and policy. Importantly, the model is strictly instructed to solely utilize the information provided in the transcripts without adding any new data. An example: a transcript detailing patient symptoms and anamnesis would result in a summary with categories such as "Reason for Sick Leave: Pneumonia," "First Day of Sick Leave: 15-07-2024," and "Anamnesis: Patient reports coughing in the last three weeks," among others. The medical doctors have written medical summaries that form the baseline for comparison with the model’s summary output.

Accuracy was measured by two metrics: the percentage of correct elements or sentences in the entire text and the categorization accuracy, defined as the percentage of correct elements or sentences under the correct subheading. An element was considered correct if it was also included in the control, i.e., the medical summary manually written by the doctor. Similarly, an element was considered correctly categorized if it was placed under the same category in the control. The total number of elements was therefore based on the control. Examples of such elements include the patient's first name, last name, date of the consultation, diagnosis or main symptom under the conclusion, and the different aspects of a complaint if described. For instance, if a patient describes a headache, the nature, location, duration, and intensity are all considered individual elements. In the treatment section, the type of treatment, the treating professional, and the frequency of the treatment are all individual elements. For medication, the medication name, dosage, and frequency are each considered individual elements. For other categories, the elements were determined in a similar manner. Efficiency was assessed by the speed of document generation. Moreover, user satisfaction was also gauged by a scale from 1 to 5. An additional metric, readability, was measured using the Flesch-Douma formula, which is specifically designed for assessing the readability of Dutch text, calculating a readability score based on average sentence length and word length (Formula 2). Here, a higher score indicates better readability with a maximum score of 206.83 (Vandeghinste & Bulté, 2019). In the case of the medical summarizer, paired t-tests were applied to compare the readability scores of the summaries with the human-written summaries, testing the hypothesis that the model-generated text would be equally readable. Efficiency, measured by the time required to generate summaries, was similarly evaluated with paired t-tests to test whether the summarizer was significantly faster or slower. Here, two different human-generated times were considered, as some physicians begin writing their summaries during the consultation, while others complete them afterward. Therefore, the time taken by the model to generate summaries was compared using paired t-tests against both human-generated times: one that includes the consultation time and one that does not.

$$\text{Flesch} - \text{Douma} - \text{formule: score} = 206.835 - (0.77 \times \text{average sentence length}) - (100 \times \text{average word length}) \qquad (2)$$

F. ChatCPT-4 Custom Model: Letter Generator

The second custom model, the ‘Letter Generator’, drafts formal letters for employers and employees based on the medical summaries with the intention to ensure confidentiality by not reporting any medical aspects. This model uses the structured summaries to compose letters with several predefined sentences organized into the following sections: Introduction, Limitations regarding work, Advice, Prognosis, and Follow-up Appointment. For instance, an introduction might read, "I spoke with Mr. A. during a physical consultation. The individual has limitations due to disease and receives adequate support. I have given the individual several recommendations." The limitations section might state, "The individual has limitations in the areas of dynamic action: limited with respect to frequent bending, pushing, pulling, lifting and carrying." These limitations are determined by evaluating the medical summaries through the ‘Claim Assessment and Assurance System’-criteria (Dutch: Claim Beoordelings- en Borgingssysteem-criteria; CBBS-criteria) (UWV, 2024). The CBBS-criteria serve as a reference framework, which includes the assessment of functional capabilities and limitations, and ensures compliance with Dutch regulations. To assess the letters generated by the model, the medical doctors also created handwritten letters.

The evaluation metrics used for the letter generator were similar to those used for the medical summarizer. Accuracy was assessed by comparing the generated letters to a control letter manually written by the doctor. Elements within the letter, such as the patient's condition, limitations, and treatment recommendations, were considered correct if they matched the corresponding elements in the control. For example, a specific limitation in dynamic functioning, such as not being able to walk for a certain time is considered as an element. Categorization accuracy was evaluated by ensuring that each element was correctly placed under the same section of the control letter, such as ‘Restrictions’ or ‘Recommendations’. The total number of elements was, therefore, based on the control letter. Efficiency, user satisfaction, and readability were measured using the same methods as for the medical summarizer. For the letter generator, the analysis focused on determining with paired t-tests whether the model could produce letters with readability scores comparable to those manually drafted by doctors, and with greater efficiency in terms of generation time.

G. ChatCPT-4 Custom Model: Document Generator

The third custom model, known as the ‘Document Generator’, is tasked with creating a document called the Functional Capabilities List (Dutch: Functionele Mogelijkheden Lijst; FML) based on the medical summaries and letters. This model also evaluates the content according to the CBBS-criteria and fills out a Word-document FML-template, which includes 64 elements, representing the capabilities or limitations of an individual (Appendix 1). The document includes the sections: Personal Details, Personal Functioning, Social Functioning, Adjustment to Environmental Conditions, Dynamic Actions, Static Postures and Working Hours. The model is also instructed to not share any medical information within the produced document. An example of a produced element might include an entry, such as: "Section 4: Dynamic Actions: 4.11 Frequent bending during work: 2 Limited, can bend approximately 150 times per hour during the workday if necessary."

The documents were compared with FML forms that were filled in by the medical doctors. The accuracy was measured by comparing the generated FML to a control version manually prepared by the doctor. Each of the 64 elements, ranging from physical capabilities to mental and emotional functioning, was assessed for correctness by determining whether it matched the corresponding element in the control. Efficiency and user satisfaction were also measured similar to the other models, while readability was not assessed in this context, as the document is structured in a list format with standardized options. The document generator was assessed with paired t-tests to determine if the model could efficiently generate an FML by comparing the time required to generate these documents with the time needed for the doctors to generate the document

H. ChatCPT-4 Custom Model: Medical Protocol Assistant

Finally, the ‘Medical Protocol Assistant’ is a model designed to answer questions based on 28 medical protocols, established by the NVAB. It is instructed to provide responses strictly based on these guidelines. It cites the exact location of information within the protocols to support its answers. If a question cannot be answered due to a lack of information in the protocols, the model clearly states the absence of relevant information and refrains from providing information from other sources. To evaluate this model, doctors formulated the input for the model, namely three random medical questions related to the given case (e.g., concerning symptoms, diagnosis, treatment, or reintegration advice). The output of the model, i.e. the responses of the model, were displayed to the participants for evaluation. Additionally, the doctors attempted to find the answers for the questions formulated for the evaluation of the Medical Protocol Assistant by themselves in the available protocols.

This data served as the ground truth and provided a means to evaluate the responses of the ChatGPT-4 model.

When evaluating the medical protocol assistant, accuracy was determined by assessing whether the responses provided by the assistant to the user's questions were based on the protocols accessible to the model, specifically the NVAB protocols. If a response did not originate from these protocols, it was considered fabricated and thus incorrect. Efficiency was evaluated by measuring the time taken to generate answers to three questions posed by the doctor, with this duration compared to the time the doctor would take to independently locate the answers within the protocols. Additionally, the doctor evaluated the relevance of the responses as a dichotomous variable (relevant or not relevant) and rated overall satisfaction with the assistant's performance on a 5-point scale. In this context, relevant means that the assistant's answer appropriately addresses the given question and is considered useful and applicable. The medical protocol assistant was evaluated with paired t-tests to see if the model could efficiently generate responses compared to the manual searches.

I. Subgroup analyses

Subgroup analyses of the large and medium models of the Whisper speech-to-text converter were conducted. This analysis was essential to evaluate the trade-offs between efficiency and accuracy in the specific context of this application. As the subgroups were relatively small, Mann-Whitney U-tests were utilized on the original data to compare the accuracy and efficiency variables of the speech-to-text converter and of the medical summarizer, letter generator and document generator. Readability scores for the medical summarizer and letter generator were also compared for these subgroups. Kruskal-Wallis H-tests were used on the original data to compare the satisfaction variables. Initially, an internal microphone was used for transcribing; however, it was noted that the audio quality was not sufficiently clear, leading to the decision to switch to an external microphone. Equivalent to the subgroup analyses of the large and medium Whisper models, subgroup analyses were performed to compare simulations that used an external microphone versus an internal microphone. The subgroup analyses results can be found in Appendix 4.

III. RESULTS

A total of 30 simulations, each representing a different case, were conducted with 14 participants, all of whom are medical doctors that were working in the field of occupational medicine. All data is made available in Supplementary file 2. These participants engaged in the simulations in two roles: as medical doctors and as simulation patients. 25 out of 30 simulations applied an external microphone (Table 2). 19 simulations applied the Whisper Medium model, while the remaining simulations applied the Large v3 model. The descriptive statistics for all the models are summarized In Table 3 and the results of the paired samples t-tests for efficiency and readability variables are presented in Table 4. These tests compare the performance of the AI-based speech and text models against human performance in terms of generating time and readability scores. In Appendix 5 a summary table of the most important findings is displayed.

Of all simulations three cases missed some data. In the simulation corresponding to case ID 14, the audio file was corrupted and therefore unplayable, which precluded the possibility of obtaining a human transcription. As a result, this case lacked both a speech-to-text generation time data point and a WER data point. The simulation with case ID 9 was missing the human-generated time data for the medical summarizer and letter generator, because the participant did not comply with the instructions. Additionally, the simulation with case ID 26 lacked an accuracy score and satisfaction rate of the document generator due to a corrupted generated Word file, which could not be opened and read.

TABLE II. DISTRIBUTION OF WHISPER MODEL USAGE BY MICROPHONE TYPE

Whisper Model	**Microphone**		
	Internal	***External***	***Total***
Medium	1	18	19
Large	4	7	11
Total	5	25	30

TABLE III. DESCRIPTIVE STATISTICS

Variable	**N (Percentage)**	**Mean (SD)**
Speech-to-text converter		
Substitutions	29 (96.7%)	98.7 (76.9)
Insertions	29 (96.7%)	61.1 (48.1)
Deletions	29 (96.7%)	56.1 (37.6)
Total number of words in reference	29 (96.7%)	1,676.0 (728.2)
Words-error-rate	29 (96.7%)	13.1 (6.6)
Generating time	30 (100%)	18.4 (11.6)
Human generating time	29 (96.7%)	25.0 (10.9)
Satisfaction rate	30 (100%)	3.2 (0.8)
Medical Summarizer		
Percentage of incorrect elements	30 (100%)	38.6 (15.2)
Total number of elements	30 (100%)	65.0 (21.8)
Percentage of incorrectly categorized elements	30 (100%)	4.8 (5.8)
Readability score	30 (100%)	191.5 (4.2)
Human readability score	30 (100%)	189.7 (6.7)
Generating time	30 (100%)	0.5 (0.6)
Human generating time (excluding consultation time)	29 (96.7%)	2.4 (2.0)

Human generating time	29 (96.7%)	14.7 (5.9)
Satisfaction rate	30 (100%)	4.2 (0.9)
Letter Generator		
Percentage of incorrect elements	30 (100%)	90.4 (63.4)
Total number of elements	30 (100%)	18.9 (8.6)
Percentage of incorrectly categorized elements	29 (96.7%)	4.8 (9.3)
Readability score	30 (100%)	177.5 (10.6)
Human readability score	30 (100%)	175.9 (12.6)
Generating time	29 (96.7%)	0.4 (0.6)
Human generating time	29 (96.7%)	3.4 (2.1)
Satisfaction rate	30 (100%)	3.1 (0.9)
Document Generator		
Percentage of incorrect elements	29 (96.7%)	17.5 (8.5)
Total number of elements	30 (100%)	68 (0.0)
Generating time	30 (100%)	3.9 (5.0)
Human generating time	30 (100%)	5.3 (2.7)
Satisfaction rate	29 (96.7%)	2.5 (1.1)
Medical Protocol Assistant		
Number of protocol based replies	29 (96.7%)	2.6 (0.6)
Number of desired replies	29 (96.7%)	2.5 (0.7)
Relevancy score	30 (100%)	4.4 (0.8)
Generating time	30 (100%)	1.3 (0.6)
Human generating time	30 (100%)	5.0 (2.3)
Satisfaction rate	30 (100%)	4.4 (0.8)

A. *Performance of Whisper*

The accuracy of the speech-to-text converter was evaluated using the WER, with a mean of 13.1% (SD = 6.6). Errors were primarily due to substitutions (mean of 98.7 words, SD = 76.9), followed by insertions (mean of 61.1 words, SD = 48.1) and deletions (mean of 56.1 words, SD = 37.6), out of an average total amount of words of 1,676.0 (SD = 728.2) in the human-generated transcription. The efficiency of the speech-to-text converter was assessed by measuring the transcription time of the recorded audio files. The conversation part of the consultation took on average 12.0 minutes (SD = 5.0 minutes). The Whisper model generated the transcription with a mean of 18.4 minutes (SD = 11.6), while the main researcher transcribed the audio files with a mean of 25.0 minutes (SD = 10.9). The generating time was significantly faster than human generating time, with a mean difference of 6.6 minutes (t = -3.06, p = .002). This corresponds to a moderate effect size estimate (Cohen's d = -.541, 95% CI: -.927 – -.147). The medical doctors rated their satisfaction with the Whisper model's generated transcriptions, resulting in a mean satisfaction rate of 3.2 out of 5 (SD = 0.8).

TABLE IV. PAIRED SAMPLES T-TESTS OF EFFICIENCY AND READABILITY VARIABLES

Variable 1 - Variable 2	**Mean difference**	**t**	**p**	**Cohen's d effect size estimate (95% CI-interval)**
Speech-to-text converter				
Generating time - Human generating time	-6.6	-3.06	**.002**	-.541 (-.927 – -.147)
Medical summarizer				
Readability score - Human readability score	1.8	1.22	.224	.222 (-.142 – .582)
Generating time - Human generating time (excluding consultation time)	-1.9	-4.51	**<.001**	-.937 (-1.370 – -.493)
Generating time - Human generating time	-14.2	-13.62	**<.001**	-2.456 (-3.187 – -1.714)
Letter generator				
Readability score - Human readability score	1.6	0.60	.551	.109 (-2.51 – .467)
Generating time - Human generating time	-3.0	-7.17	**<.001**	-1.425 (-1.939 – -.899)
Document generator				
Generating time - Human generating time	-1.4	-1.33	.183	-.243 (-.604 – .122)
Medical Protocol assistant				
Generating time - Human generating time	-3.7	-8.67	**<.001**	-1.584 (2.118 – -1.037)

B. *Performance of the Medical Summarizer*

The medical summarizer model's accuracy was assessed by the percentage of incorrect elements, which had a mean of 38.6% (SD = 15.2%) out of an average total of 65.0 elements (SD = 21.8). An example of such an incorrect element is seen in the simulation with case ID 23, where the summarizer did not mention that the employee had visited her general practitioner in the treatment section, while the medical doctor did in his summary. Additionally, some elements were also categorized incorrectly, with a mean of 4.8% (SD = 5.8%). The simulation with case ID 26 illustrates such incorrect categorizations, where quitting the usage of the herbal supplement valerian was recommended by the medical doctor to improve the employee's sleep and thus should be placed in the policy section, yet the summarizer placed it in the treatment section. Readability was another metric that we considered, which was determined with the earlier stated Flesch-Douma-formula for both the summary produced by the medical summarizer and for the human generated summary. The medical summarizer achieved a mean readability score of 92.6% (SD = 2.0%), slightly higher than the human readability score of 91.7% (SD = 3.2%). The readability score showed no significant difference between the model and human-generated summaries, with a mean difference of 1.8 (t = 1.22, p = .224) and a small effect size estimate (Cohen's d = .222, 95% CI: -.142 – .582). The summarizer's efficiency, measured by generating time, had a mean of 0.5 minutes (SD = 0.6), compared to the mean of 14.7 minutes (SD = 5.9), which is the time needed for the participants to generate the summary. This indicates that the summarizer is approximately 29 times faster than the human participants in producing a summary. The generating time for the summarizer was significantly shorter than human generating time, both excluding consultation time (-1.9 minutes, t = -4.51, p < .001, Cohen's d = -.937, 95% CI: -1.370 – -.493) and including consultation time (-14.2 minutes, t = -13.62, p < .001, Cohen's d = -2.456, 95% CI: -3.187 – -1.714), indicating a large effect size. The participants scored their satisfaction with the medical summarizer a mean of 4.2 out of 5 (SD = 0.9).

C. *Performance of the Letter Generator*

Similar metrics were used to evaluate the letter generator model. Accuracy was evaluated by the percentage of elements, with a mean of 90.4% (SD = 63.4%) out of an average total of 18.9 elements (SD = 8.6). Case 26 depicts such an incorrect element, where the medical doctor solely restricted the employee with regards to his energy capacity, while the model added limitations regarding personal functioning, stating that the employee could only perform concrete, singular tasks in an environment with few distractions and no frequent deadlines or peaks. Additionally, the model incorrectly imposed restrictions on working hours, specifying that the employee was limited in performing night shifts and preferred day shifts to facilitate recovery. Moreover, the percentage of incorrectly categorized elements had a mean of 4.8% (SD = 9.3%). Case 26 also illustrates this erroneous categorization, where the advice of the medical doctor to abstain from valerian usage is placed in the advice section of the letter, while this is perceived as medical information and should thus not be shared explicitly in the letter, such as the medical doctor did by mentioning in the introduction section that the employee was given various advice. The readability mean score of the model-generated letter was 85.8% (SD = 5.1%), compared to the readability score of the human-generated letter of 85.1% (SD = 6.1%). The readability score did not differ significantly from the human-generated letters, with a mean difference of 1.6 (t = 0.60, p = .551) and a small effect size (Cohen's d = .109, 95% CI: -.251 – .467). The efficiency of the letter generator, measured by generating time, had a mean of 0.4 minutes (SD = 0.6), while the human-generated time averaged 3.4 minutes (SD = 2.1). The generating time for the letter generator was significantly shorter than that of humans, with a mean difference of -3.0 minutes (t = -7.17, p < .001), reflecting a large effect size (Cohen's d = -1.425, 95% CI: -1.939 – -.899). The participants rated their satisfaction with the letter generator with a mean of 3.1 out of 5 (SD = 0.9).

D. *Performance of the Document Generator*

The document generator's accuracy was also measured by the number of incorrect elements, which had a mean of 12.0 elements (SD = 5.8). As the document is standardized the total amount of elements were consistent, namely 68 elements; the mean percentage of incorrect elements was thus 17.5% (SD = 8.5%). As an example, in the output of the simulation with case ID 10 the model generated a restriction in pushing, while the doctor determined that the restriction was not applicable, and the capability was normal. The efficiency of the document generator, as indicated by the generating time, had a mean of 3.9 minutes (SD = 5.0), while the human-generated time averaged 5.3 minutes (SD = 2.7). The generating time was not significantly different from the human-generated time, with a mean difference of -1.4 minutes (t = -1.33, p = .183) and a small effect size (Cohen's d = -.243, 95% CI: -.604 – .122). The satisfaction rate for the document generator was the lowest among the models, with a mean of 2.5 out of 5 (SD = 1.1).

E. *Performance of the Medical Protocol Assistant*

The medical protocol assistant was used by the participants to answer three questions that were defined by the participants themselves. The model was evaluated based on the number of protocol-based replies, with a mean of 2.6 (SD = 0.6) out of every 3 questions, and the number of replies that were deemed as desired according to the participants, with a mean of 2.5 replies (SD = 0.7). An example from the simulation with case ID 11 illustrates a situation where the model generated an answer that was not supported by any existing protocol. In this case, the question was, "When should I refer an employee to a medical specialist?" The model provided a detailed response, including various scenarios such as when a diagnosis cannot be established, in cases of complex issues requiring multidisciplinary coordination, or when there is stagnation in the recovery process. The specific guidance given by the model did not directly align with the established protocol used by the participant. Moreover, the participants rated whether the answer was relevant on a 5-point scale. The relevancy score according to the participants was high, with a mean of 4.4 (SD = 0.8). Efficiency was measured by the generating time, which had a mean of 1.3 minutes (SD = 0.6) when the model was used, compared to participants attempting to acquire the answers on their questions by searching the protocols themselves, which took 5.0 minutes on average (SD = 2.3). The generating time was significantly shorter than that of the human-generated responses, with a mean difference of -3.7 minutes (t = -8.67, p < .001), indicating a large effect size (Cohen's d = -1.584, 95% CI: -2.118 – -1.037). The satisfaction rate for the medical protocol assistant was also high, with a mean of 4.4 out of 5 (SD = 0.8).

IV. Discussion

The objective of this exploratory study was to investigate the potential of AI speech and text technologies, specifically Whisper and ChatGPT-4, in decreasing the administrative burden of occupational physicians. The increasing administrative burden has been noted as a major factor to burnout and in reducing the time availability for and quality of care of patients (Swensen et al., 2016; Wu, 2023). By investigating Whisper for transcription and ChatGPT-4 for various text generation tasks we have estimated the accuracy, efficiency and user satisfaction of these applications in the context of occupational health consultations.

A. Discussion of Whisper

When utilizing Whisper for transcription we observed a WER of 13.1% in our application. This is comparable to the study of Adedeji, Joshi & Doohan (2024), which displayed a WER of 12.6% when applying Whisper (version 1) in primary care consultations. Their study demonstrated that Whisper could cope with medical terminology and diverse English speaker accents. In contrast, the study by Thuestad & Grutle (2023) reported a higher WER of 33.5% when using Whisper (version 2) for Norwegian emergency calls, potentially highlighting the challenges Whisper faces when applied to low-resource languages and the more demanding audio conditions in emergency scenarios. This variation displays the importance of context and language when assessing ASR performance, which was also recognized by the developers of Whisper, OpenAI, where a high variety in WER can be observed in various languages, depending on factors such as the amount of training data available for each language, the complexity of the spoken language, and the acoustic environment in which the model is applied (Radford et al., 2023).

Next to Whisper's accuracy parameter, we also observed that transcriptions were created more efficiently in our experiment, namely on average 6.6 minutes faster. It should be noted that these transcriptions were made by a single researcher, which may limit the generalizability of the efficiency gains across different transcribers. However, the experiment involved 30 cases with different doctors, and in all of these cases we observed a difference in time. Moreover, utilizing Whisper for transcription tasks can be advantageous, as transcribing tasks can be experienced as tedious by transcribers (Point & Baruch, 2023). However, it is also needed to consider that, in direct medical practice, the standalone use of transcriptions is of limited relevance. Medical professionals typically rely more on summaries than full transcriptions of consultations, especially when making clinical decisions (Clough et al., 2024). This context diminishes the direct applicability of Whisper's transcription efficiency in routine medical workflows, where a concise and accurate summary holds greater value than a verbatim transcript. This consideration might also explain the moderate user satisfaction rate of 3.2 out of 5 observed in our study. While Whisper's efficiency is an asset, the need for integration with other applications, such as the Medical Summarizer, is necessary to maximize its utility in clinical settings.

In addition, in assessing the performance of the medium and large Whisper models, we observed that while both models deliver similar accuracy, the medium model offers a distinct advantage in efficiency, completing transcriptions notably faster than the large model. The large model, with its higher parameter count, requires more computational resources and takes longer to process transcriptions. In contrast, the medium model, despite having fewer parameters, achieves similar levels of accuracy, with only minimal differences in word error rate (WER) compared to the large model (OpenAI, 2022). This makes the medium model an attractive choice for scenarios where both accuracy and time efficiency are crucial.

B. Discussion of the Medical Summarizer

The Medical Summarizer, as one of the ChatGPT-4 modules, converted transcriptions from consultations into structured medical summaries. The accuracy of this module was assessed by the percentage of correct elements generated, revealing that the Medical Summarizer had a significant error rate with 38.6% of elements being incorrectly generated. These inaccuracies can be attributed to several factors. First, doctors used their own familiar medical summary formats, which reflected their realistic practice but differed from the standardized format that the Medical Summarizer used, which was created by the researcher in consultation with an occupational doctor and a trainee occupational doctor. This discrepancy inherently led to variations in what doctors deemed important to document. Moreover, one respondent noted that the ChatGPT-4 module generated summaries that occasionally included elements that had not initially been heard during the consultation by the medical doctor, but which were indeed spoken when the recording was replayed by the researcher. Despite being accurate in capturing these details, these elements were still considered errors because the doctor's summary was treated as the 'ground truth.' Furthermore, it is important to consider that this is currently a standardized 'medical summarizer,' and the tool can be further fine-tuned to align with each doctor's specific format and style. Tang et al. (2023) also displayed that summarizing medical information, in their case medical evidence, with LLMs lacks accuracy and can lead to harm due to misinformation as LLMs can generate factually contradictory summaries with unclear or excessively convincing statements. Despite the inaccuracies, the Medical Summarizer demonstrated an evident efficiency, generating summaries 29 times faster than human participants, with an average time of 0.5 minutes compared to the 14.7 minutes required by the doctors. This efficiency, coupled with a slightly higher readability score than that of human-generated summaries, was reflected in a relatively high user satisfaction score of 4.2 out of 5. However, the inaccuracies suggest that while the tool is timesaving, its outputs still require careful review by medical professionals to ensure accuracy, which is comparable to other applications of LLMs (Tang et al., 2023).

C. Discussion of the Letter Generator

The Letter Generator, the second customized ChatGPT-4 module, has been designed to create formal letters based on the generated medical summaries while ensuring confidentiality by excluding explicit medical information. The study found that the Letter Generator had a notable error rate, with 90.4% of the elements being incorrectly generated. Several factors contributed to this high error rate. Firstly, a significant portion of the errors involved the module incorrectly mentioning medical information, despite being explicitly instructed not to do so. This suggests a challenge in the model's ability to consistently follow the confidentiality constraints, possibly due to the complex nature of distinguishing between what constitutes medical information and what does not in varied contexts. This aspect is critical because LLMs in the medical context must handle sensitive data with utmost care to prevent breaches and ensure the strict confidentiality of patient information (Wang et al., 2023). Secondly, the module had difficulty with using information from the medical summary to accurately apply the CBBS criteria, which are essential for determining functional capabilities and limitations in occupational health (Geiger et al., 2018). The difficulty here may stem from the model's lack of a sophisticated mechanism to integrate the nuanced requirements of the CBBS criteria with the information extracted from the medical summaries. This integration requires a level of judgment and context-awareness that the current module may not fully possess, leading to misapplications or omissions in the final letter. Moreover, the Letter Generator works with a smaller number of elements (18.9 on average) in comparison to the other text-generating ChatGPT-4 modules, the Medical Summarizer and Document Generator, which handle a larger number of elements (65 and 68 elements on average, respectively). This smaller element base means that each error has a proportionally greater impact on the overall accuracy percentage, making the module appear more prone to higher error rates. Despite these issues, the Letter Generator was significantly more efficient than manual letter drafting, completing its tasks 8.5 times faster. The satisfaction score for this module was moderate, at 3.1 out of 5, reflecting a balance between its efficiency and the inaccuracies that might hinder its full adoption in clinical settings.

D. Discussion of the Document Generator

The Document Generator, designed to create FML's, demonstrated an error rate of 17.7% of elements being incorrectly generated. One contributing factor to this error rate is the reliance of the FML on the letter produced by the Letter Generator as an input. If errors were made in determining the employee's restrictions in the letter, these errors were inevitably propagated into the FML, which is commonly observed in NLP (Lê & Fokkens, 2017). This dependency means that any inaccuracies in the letter's content were directly reflected in the generated FML, compounding the potential for errors. Moreover, the FML has a standardized structure with a fixed number of elements, most of which were not restricted in the majority of cases, whether generated by the Document Generator or determined by the medical doctor. This inherent structure likely contributed to the better overall error rate compared to the Letter Generator, as many of the elements did not require modification or restriction, reducing the likelihood of errors in those areas. However, this also suggests that the error rate might be understated when considering only the elements that were actually restricted. If the analysis were limited to the elements where restrictions were applied, the error rate could potentially be higher, revealing more about the module's limitations in accurately reflecting the nuanced restrictions required for occupational health documentation. Nonetheless, the Document Generator was slightly more efficient than manual document generation, with an average time of 3.9 minutes compared to 5.3 minutes for human-generated FMLs. However, the relatively modest time savings did not outweigh the accuracy concerns, which possibly contributed to the lower satisfaction score of 2.5 out of 5 among users. This score reflects users' apprehensions regarding the reliability and accuracy of the tool, which are crucial when dealing with standardized documents that have significant legal and clinical implications. Therefore, while the Document Generator shows potential in improving efficiency, its current level of accuracy poses a substantial barrier to its effective adoption in real-world applications, highlighting the need for further refinement and customization to meet the meticulous requirements of occupational health documentation.

E. Discussion of the Medical Protocol Assistant

The Medical Protocol Assistant, the final ChatGPT-4 module evaluated in this study, was designed to provide protocol-based answers to medical questions during occupational health consultations. The accuracy of this module was assessed by the number of correct, protocol-based replies it generated. The study found that the Medical Protocol Assistant provided accurate responses for 86.7% of the questions, with 2.6 out of 3 answers on average being based directly on the incorporated protocols. However, the module did occasionally generate responses that were not fully aligned with the protocols, leading to some inaccuracies. These errors likely stem from the model's current limitations in strictly adhering to protocol content and its potential to fabricate answers when faced with ambiguous or insufficient information from the protocols. Similarly, a recent study assessing ChatGPT's responses to a variety of medical questions found that while the model demonstrated a high level of accuracy, consistent with the performance of the Medical Protocol Assistant, it also highlighted the need for careful consideration of the generated advice in clinical settings, especially in complex or ambiguous situations that require a deep understanding of context and patient-specific nuances (Johnson et al., 2023).

Despite these accuracy challenges, the Medical Protocol Assistant demonstrated impressive efficiency, generating responses significantly faster than the manual process of searching through the protocols. On average, the assistant was on average almost four times faster in generating the responses compared to doctors manually searching for the answers. This efficiency might be reflected in the high satisfaction score of 4.4 out of 5, the highest among all the modules evaluated in this study. The high satisfaction rating indicates that, despite occasional inaccuracies, users found the Medical Protocol Assistant to be a valuable tool in streamlining the consultation process by quickly providing relevant and protocol-based information. However, for the module to be more widely adopted in clinical practice, further refinement is necessary to ensure that its responses are consistently accurate and fully aligned with medical protocols, especially in cases where precise guidance is critical.

F. Limitations

This exploratory study, although delivering useful insights into the potential of AI speech and text technologies to reduce the administrative burden on occupational physicians, carries certain limitations that should be noted when interpreting the findings.

Firstly, the study was an exploratory study with a small sample size, involving only 14 physicians who participated as both doctors and patients, though not concurrently, which may limit the generalizability of the findings. The results may be influenced by the specific characteristics of these participating doctors, such as their familiarity with AI speech and text technologies, as well as their professional background in occupational health. Because these doctors are well-versed in the field, they might be considered "ideal" patients during the simulations. They know precisely what information is expected from them, such as providing the exact drug name and dosage when asked about medication. This level of detail and accuracy might not be as easily obtained from a broader, more diverse patient population. Additionally, the fact that the doctors are all university-educated professionals means that they do not represent the full spectrum of the general population, which could further impact the applicability of the study's findings to a more diverse patient group. Moreover, the simulated nature of the consultations, while designed to reflect realistic scenarios, may not fully capture the complexities and variabilities of actual clinical practice. This limitation could affect the applicability of the findings to real-world settings, where factors such as patient interaction, time pressures, and varied case complexities might influence the performance of the technologies.

Second, the study relied on control data provided by the physicians themselves, which served as the 'ground truth' against which the modules' generated outputs were compared. However, this assumption may add bias since it assumes that the physicians' manual summaries documents are entirely accurate and error-free. Human error is inevitable and the control data may not always reflect the absolute truth, possibly skewing the evaluation of the models.

Another limitation is the study's exclusive emphasis on a single context, namely occupational health consultations in the Netherlands. The findings may not be generalizable to other medical specializations or healthcare systems, as consultations, documents requirements, and administrative burdens might vary greatly. Furthermore, the Medical Summarizer was evaluated based on its suitability with a standardized format developed by the researcher in collaboration with a limited number of occupational doctors. This standardization may not adequately represent the varying documentation styles and preferences of a wider variety of practitioners, thus affecting perceived accuracy and user satisfaction in other contexts. The standardization might also not fully capture the potential of the module when tailored to specific individual user's needs. Furthermore, it is important to note that when using Whisper and ChatGPT-4, OpenAI currently collects the data inputted, to improve their performance. This study did not evaluate the privacy implications of this data collection, highlighting the need for a solution that ensures medical data is not shared with the provider of such technologies to protect patient confidentiality.

Moreover, the definitions of key terms used in the study, such as "element" and "satisfaction," could be interpreted differently by various participants. For instance, the study defined an "element" as a distinct piece of information in the medical summaries or documents, but participants might have had different understandings of what constitutes an element, leading to potential inconsistencies in the evaluation. Similarly, "satisfaction" was not explicitly defined for the participants, which could result in varying interpretations of what satisfaction entails, potentially affecting the accuracy of the satisfaction ratings.

G. Future studies

Given the limitations identified in this exploratory study, several avenues for future research can be explored. First, expanding the study to different healthcare settings and including a larger and more diverse sample of participants, including non-medical professionals and patients from various backgrounds in real-life consultations, could provide more insight in the applicability of these technologies. Additionally, future studies should consider longitudinal designs to assess the long-term impact of these technologies. Moreover, refining the definitions of key terms including "element", “relevance” and "satisfaction" can help to reduce variability in future evaluations. In addition, to reduce bias from physician-provided control data, future studies could use a consensus-based ground truth, where multiple physicians independently validate the data, or employ a blinded review process with independent reviewers assessing the outputs. Lastly, addressing the ethical and privacy concerns related to data collection by exploring secure, on-premises solutions for deployment of the technologies in healthcare would be essential to safeguard patient confidentiality and build trust in these emerging technologies.

H. Conclusion

This study assessed the potential of AI speech and text technologies, particularly Whisper and customized ChatGPT-4 models, to alleviate the administrative workload for occupational physicians. The use of Whisper for transcription demonstrated notable efficiency gains, however, the moderate word error rate and user satisfaction indicate that further refinements are needed to enhance accuracy and user acceptance in clinical settings. The ChatGPT-4 Medical Summarizer showed significant time-saving potential, generating summaries 29 times faster than human participants. Despite these efficiencies, the error rate highlights the need for the continued involvement of physicians and customization to align with individual medical professionals' documentation styles. The Letter Generator and Document Generator, while also improving efficiency, exhibited higher error rates and lower satisfaction scores, particularly in accurately applying and assessing criteria and maintaining confidentiality. Overall, while AI technologies like Whisper and ChatGPT-4 offer promising avenues for reducing administrative tasks in occupational medicine, their successful implementation will require ongoing adjustments to improve accuracy and user satisfaction. Future research and development should focus on enhancing these technologies to ensure they meet the high standards required in healthcare settings, particularly in handling sensitive and legally significant documentation. Overall, while AI technologies like Whisper and ChatGPT-4 show potential to reduce administrative tasks in occupational medicine, their current limitations in accuracy indicate the need for further research.

V. References

[1] Abdullah, M., Madain, A., & Jararweh, Y. (2022, November). ChatGPT: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-8). Ieee.

[2] Adedeji, A., Joshi, S., & Doohan, B. (2024). The Sound of Healthcare: Improving Medical Transcription ASR Accuracy with Large Language Models. *arXiv preprint arXiv:2402.07658*.

[3] Al Ghouch, Y., & Stienen, A. H. A. (2024). *A systematic review of AI recommendation generating technologies applied real-time in medical consultations*. Faculty of Mechanical Engineering, Technical University Delft, Delft, Netherlands.

[4] Alto, V. (2023). *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4*. Packt Publishing Ltd.

[5] Clough, R. A. J., Sparkes, W. A., Clough, O. T., Sykes, J. T., Steventon, A. T., & King, K. (2024). Transforming healthcare documentation: harnessing the potential of AI to generate discharge summaries. *BJGP open*, *8*(1).

[6] Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94-98.

[7] Functionele mogelijkhedenlijst. (2013). In UWV, *UWV*. https://www.steungroep.nl/images/her_keuring_WIA_of_WAO/Wetten_en_regels_bij_her_keuring/Functionele_mogelijkhedenlijst_UWV_2013.pdf

[8] Geiger, B. B., Garthwaite, K., Warren, J., & Bambra, C. (2018). Assessing work disability for social security benefits: international models for the direct assessment of work capacity. *Disability and Rehabilitation*, *40*(24), 2962-2970.

[9] Herd, P., & Moynihan, D. (2021). Health care administrative burdens: Centering patient experiences. *Health services research*, *56*(5), 751.

[10] Hingstman, L., Velden, L. F. J., & Schepman, S. M. (2009). *Mobiliteit van bedrijfsartsen*. NIVEL.

[11] Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., ... & Wheless, L. (2023). Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square*.

[12] Kernberg, A., Gold, J. A., & Mohan, V. (2024). Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. *Journal of Medical Internet Research*, *26*, e54419.

[13] Kumar, A., & Gond, A. (2023). NATURAL LANGUAGE PROCESSING: HEALTHCARE ACHIEVING BENEFITS VIA NLP. *ScienceOpen Preprints*.

[14] Lê, M., & Fokkens, A. (2017). Tackling error propagation through reinforcement learning: A case of greedy dependency parsing. *arXiv preprint arXiv:1702.06794*.

[15] Nederlandse Vereniging voor Arbeids- en Bedrijfsgeneeskunde. (n.d.). *NVAB Richtlijnen*. https://nvab-online.nl/richtlijnen/richtlijnen-nvab

[16] OpenAI. (2022, September 21). *Introducing Whisper*. https://openai.com/index/whisper/

[17] Oude Mulders, J., Henkens, K., & van Dalen, H. P. (2020). How do employers respond to an aging workforce? Evidence from surveys among employers, 2009–2017. *Current and emerging trends in aging and work*, 281-296.

[18] Plomp, H. N., & Van Der Beek, A. J. (2014). Job satisfaction of occupational physicians in commercial and other delivery settings: A comparative and explorative study. *International journal of occupational medicine and environmental health*, *27*, 672-682.

[19] Point, S., & Baruch, Y. (2023). (Re) thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management*, *39*(2), 101272.

[20] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492-28518). PMLR.

[21] Rao, S. K., Kimball, A. B., Lehrhoff, S. R., Hidrue, M. K., Colton, D. G., Ferris, T. G., & Torchiana, D. F. (2017). The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. *Academic Medicine*, *92*(2), 237-243.

[22] Spear, J., Ehrenfeld, J. M., & Miller, B. J. (2023). Applications of artificial intelligence in health care delivery. *Journal of medical systems*, *47*(1), 121.

[23] Swensen, S., Shanafelt, T., & Mohta, N. S. (2016). Leadership survey: Why physician burnout is endemic, and how health care must respond. *NEJM Catalyst*, *2*(6).

[24] Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., ... & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, *6*(1), 158.

[25] Thuestad, J. A., & Grutle, Ø. (2023). *Speech-to-text models to transcribe emergency calls* (Master's thesis, The University of Bergen).

[26] UWV. (2024). *Basisinformatie CBBS*. https://www.uwv.nl/overuwv/Images/bijlage-1-basisinformatie-cbbs-versie-maart-2024.pdf

[27] Vandeghinste, V., & Bulté, B. (2019). Linguistic proxies of readability: Comparing easy-to-read and regular newspaper Dutch. *Computational Linguistics in the Netherlands Journal*, *9*, 81-100.

[28] Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. *Journal of Medical Internet Research*, *25*, e48009.

[29] Wixom, B. H., & Todd, P. A. (2005). A theoretical integration of user satisfaction and technology acceptance. *Information systems research*, *16*(1), 85-102.

[30] Woolhandler, S., & Himmelstein, D. U. (2014). Administrative work consumes one-sixth of US physicians' working hours and lowers their career satisfaction. *International Journal of Health Services*, *44*(4), 635-642.

[31] Wu, J.-S. (2023). Healthcare Service Efficiency: An Empirical Study on Healthcare Capacity in Various Counties and Cities in Taiwan. Healthcare, 11(11), 1656. https://doi.org/10.3390/healthcare11111656

[32] Yakubovskyi, R., & Morozov, Y. (2023). Speech Models Training Technologies Comparison Using Word Error Rate. *Advances in Cyber-Physical Systems*, *8*(1), 74-80.

[33] Zegers, M., Veenstra, G. L., Gerritsen, G., Verhage, R., Van Der Hoeven, H. J., & Welker, G. A. (2022). Perceived burden due to registrations for quality monitoring and improvement in hospitals: a mixed methods study. *International Journal of Health Policy and Management*, *11*(2), 183.

VI. Appendices

Appendix 1. FML-template

(“Functionele Mogelijkhedenlijst”, 2013)
FUNCTIONELE MOGELIJKHEDENLIJST

Onderstaande versie van de Functionele Mogelijkhedenlijst van UWV is aangepast aan de Basisinformatie CBBS, mei 2013.

Enige uitleg voor degenen die gekeurd/beoordeeld worden in verband met ziekteverzuim en arbeidsongeschiktheid is te vinden in:

- Werk en inkomen bij ziekte. Een praktische gids. Bijlage 3. Steungroep ME en Arbeidsongeschiktheid, 2012

- Handleiding voor de (her)keuring. Met 60 praktische tips. Bijlage 3.
Steungroep ME en Arbeidsongeschiktheid. 2005, herziene uitgave verwacht.
Achtergrondinformatie en uitleg voor professionals is te vinden in 'Basisinformatie CBBS', UWV.

Functionele mogelijkheden en voorwaarden voor het verrichten van arbeid
---instrument voor verzekeringsarts en arbeidsdeskundige---

bron: Lisv/UWV
datum: 2001/2010/2013

FUNCTIONELE MOGELIJKHEDENLIJST

Deze lijst geeft een overzicht van mogelijkheden om in het algemeen gedurende een hele werkdag (tenminste 8 uur) te functioneren. Beperkingen van deze mogelijkheden ten opzichte van normale waarden worden in aparte rubrieken weergegeven, voor zover deze naar het oordeel van de verzekeringsarts uitingen zijn van ziekten, gebreken of ongevallen.
Als normale waarden zijn die niveaus van functioneren gekozen die het dagelijks leven regelmatig vereist. Tenzij uitdrukkelijk anders vermeld, zijn incidentele piekbelastingen boven de aangegeven niveaus van functioneren eveneens mogelijk.
Deze lijst is niet geschikt voor toepassing los van een verzekeringsgeneeskundige rapportage waarin de mogelijkheden en beperkingen aan de hand van een probleemanalyse in hun onderlinge samenhang beoordeeld, gemotiveerd en beschreven zijn.

Klant en beoordelingsgegevens
BSN : Arts :
Achternaam klant: Datum vastgelegd :
Voorletters klant : FML geldig vanaf :
Geslacht: Geldig t/m :
Geboortedatum : Criterium :
Type beoordeling:
Sjabloonversie : November 2002

Conclusie:

O De cliënt beschikt over duurzaam benutbare mogelijkheden
O De cliënt beschikt niet over duurzaam benutbare mogelijkheden Toelichting:
O De cliënt is in staat om het eigen werk volledig uit te voeren O De cliënt is in staat tot normaal functioneren (zie rubrieken)
O De cliënt heeft beperkingen ten opzichte van normaal functioneren (zie rubrieken)
O Anders, zie rapportage verzekeringsarts

O De cliënt is sterk beperkt in het persoonlijk en/ of sociaal functioneren (zie rubrieken I, II)
O De cliënt is opgenomen in ziekenhuis of AWBZ-erkende instelling O De cliënt is bedlegerig (grootste deel van de dag en langdurig)
O De cliënt is in grote mate ADL-afhankelijk
O De cliënt heeft sterk wisselende mogelijkheden/ verlies van mogelijkheden
< 3 maanden -1 jaar Duurzaamheid arbeidsbeperking: Algemene opmerking:

RUBRIEK 1: PERSOONLIJK FUNCTIONEREN

1.1 Vasthouden van de aandacht

0 normaal, kan de aandacht tenminste een half uur richten op één informatiebron.
1 beperkt, kan de aandacht niet langer dan een half uur richten op één informatiebron.
2 sterk beperkt, kan de aandacht niet langer dan 5 minuten richten op één informatiebron.

1.2 Verdelen van de aandacht

0 normaal, kan de aandacht alternerend richten op meerdere uiteenlopende informatiebronnen (autorijden in druk stadsverkeer).
1 beperkt, kan de aandacht alternerend richten op een beperkt aantal uiteenlopende informatiebronnen (het zelfstandig reizen per openbaar vervoer incl. overstappen).
2 sterk beperkt, kan niet of nauwelijks de aandacht alternerend richten op uiteenlopende informatiebronnen (kan niet zelfstandig reizen met openbaar vervoer).

1.3 Herinneren

0 normaal, kan zich meestal tijdig, zonder ongebruikelijke hulpmiddelen, relevante zaken herinneren.
1 beperkt, moet regelmatig dingen apart opschrijven als geheugensteun om de continuïteit van het handelen te waarborgen.
2 sterk beperkt, weet zich onontbeerlijke alledaagse gegevens (tijd, plaats, persoon, onderwerp) niet te herinneren en kan dit niet compenseren met hulpmiddelen.

1.4 Inzicht in eigen kunnen

0 normaal, schat meestal de eigen mogelijkheden en beperkingen realistisch in.
1 beperkt, overschat meestal ernstig de eigen mogelijkheden.
2 beperkt, overschat meestal ernstig de eigen beperkingen.

1.5 Doelmatig handelen (taakuitvoering) (gecoördineerd handelen, eigen activiteiten afstemmen op het realiseren van een doel)

0 normaal, geen specifieke beperkingen in de routine van het dagelijks leven (staat op tijd op, wast zich, kleedt zich aan, maakt ontbijt klaar, ontbijt, sluit de huisdeur af en verschijnt op tijd op afspraken)
1 beperkt, start niet tijdig activiteiten om het gestelde doel te bereiken
2 beperkt, voert de benodigde activiteiten niet in een logische volgorde uit

3 beperkt, controleert het verloop van de activiteiten niet
4 beperkt, beëindigt de activiteiten niet als het gestelde doel bereikt is, of niet bereikt kan worden
5 anderszins beperkt in doelmatig handelen, namelijk......................

1.6 Zelfstandig handelen (zelfstandige taakuitvoering)

0 normaal, geen specifieke beperkingen in het zelfstandig handelen in het dagelijks leven
1 beperkt, neemt meestal niet uit zich zelf het initiatief tot handelen
2 beperkt, stelt zich zelf meestal geen doelen
3 beperkt, ontwerpt meestal zelf geen handelingsvarianten
4 beperkt, besluit meestal zelf niet welke aanpak de meest geëigende is
5 beperkt, onderkent meestal zelf niet wanneer de gevolgde aanpak te kort schiet
6 beperkt, kiest in dat geval meestal niet zelf voor een alternatieve aanpak of een ander doel
7 beperkt, gaat uit zich zelf meestal niet door totdat het doel bereikt is
8 beperkt, doet meestal niet zelf tijdig een beroep op hulp van anderen, wanneer de situatie dat gebiedt
9 anderszins beperkt in het zelfstandig handelen, namelijk.................

1.7 Handelingstempo

0 normaal, er zijn geen specifieke beperkingen in het handelingstempo in het dagelijks leven
1 beperkt, het handelingstempo is aanmerkelijk vertraagd

1.8 Overige beperkingen in het persoonlijk functioneren

0 normaal, geen specifieke overige beperkingen in persoonlijk functioneren in het dagelijks leven
1 beperkt, specifieke overige beperkingen, namelijk......................

1.9 Specifieke voorwaarden voor het persoonlijk functioneren in arbeid (is het functioneren in arbeid door de genoemde beperkingen, of het daarop gerichte compensatiegedrag, afhankelijk van specifieke voorwaarden?)

0 nee, er gelden geen specifieke voorwaarden voor het persoonlijk functioneren in arbeid
1 ja, de cliënt is aangewezen op volledig voorgestructureerd werk: concrete enkelvoudige opdrachten (wat, wanneer, hoelang; één taak per opdracht) en voorgeschreven uitvoeringswijzen (hoe)
2 ja, de cliënt is aangewezen op vaste, bekende werkwijzen (routine-afhankelijk)
3 ja, de cliënt is aangewezen op werk dat onder rechtstreeks toezicht (veelvuldig feedback) en/of onder intensieve begeleiding wordt uitgevoerd
4 ja, de cliënt is aangewezen op werk waarbij hij niet wordt afgeleid door activiteiten van anderen
5 ja, de cliënt is aangewezen op een voorspelbare werksituatie, kan niet flexibel inspelen op sterk wisselende uitvoeringsomstandigheden en/ of taakinhoud
6 ja, de cliënt is aangewezen op een werksituatie zonder veelvuldige storingen en onderbrekingen
7 ja, de cliënt is aangewezen op werk zonder veelvuldige deadlines of productiepieken
8 ja, de cliënt is aangewezen op werk waarin geen hoog handelingstempo vereist is
9 ja, de cliënt is aangewezen op werk zonder verhoogd persoonlijk risico
10 ja, er gelden overige specifieke voorwaarden, namelijk................

RUBRIEK 2: SOCIAAL FUNCTIONEREN

2.1 Zien

0 normaal, geen specifieke beperking in het dagelijks functioneren
1 beperkt, namelijk...

2.2 Horen

0 normaal, geen specifieke beperking in het dagelijks functioneren
1 beperkt, namelijk...

2.3 Spreken

0 normaal, geen specifieke beperking in het dagelijks functioneren
1 beperkt, namelijk...

2.4 Schrijven

0 normaal, geen specifieke beperking in het dagelijks functioneren
1 beperkt, namelijk...

2.5 Lezen

0 normaal, geen specifieke beperking in het dagelijks functioneren
1 beperkt, namelijk...

2.6 Emotionele problemen van anderen hanteren

0 normaal, kan zich doorgaans inleven in problemen van anderen, maar kan daarvan afstand nemen in gedrag en beleving
1 beperkt, trekt zich meestal problemen van anderen erg aan, kan desondanks wel voldoende afstand nemen in gedrag, echter niet in beleving
2 sterk beperkt, identificeert zich meestal met problemen van anderen en kan daarvan noch in gedrag, noch in beleving afstand nemen

2.7 Eigen gevoelens uiten

0 normaal, kan meestal persoonlijke gevoelens op een voor anderen duidelijke en acceptabele manier in woord en gedrag tot uiting brengen.
1 beperkt, brengt anderen in verwarring door onduidelijke, onvoorspelbare of onconventionele wijze van gevoelsuitingen.
2 sterk beperkt, is doorgaans niet in staat gevoelens te uiten (blokkeert zichzelf) of uit deze ongecontroleerd (ongeremd), ongeacht de reacties van anderen.

2.8 Omgaan met conflicten

0 normaal, kan een conflict met agressieve of onredelijke mensen in rechtstreeks contact hanteren;
1 beperkt, kan een conflict met agressieve of onredelijke mensen uitsluitend in telefonisch of schriftelijk contact hanteren;
2 sterk beperkt, kan meestal geen conflicten hanteren.

2.9 Samenwerken

0 normaal, kan in onderlinge afstemming met anderen een taak gezamenlijk uitvoeren (werken in teamverband);

1 beperkt, kan met anderen werken, maar met een eigen, van tevoren afgebakende deeltaak;
2 sterk beperkt, kan in de regel niet met anderen werken.
2.10 Vervoer

0 normaal, kan autorijden of fietsen, of zelfstandig gebruik maken van het openbaar vervoer;
1 beperkt, is voor vervoer aangewezen op hulp van anderen.

2.11 Overige beperkingen in het sociaal functioneren

0 normaal, geen specifieke overige beperkingen in sociaal functioneren in het dagelijks leven;
1 beperkt, specifieke overige beperkingen, namelijk......................

2.12 Specifieke voorwaarden voor het sociaal functioneren in arbeid (is het sociaal functioneren in arbeid door de genoemde beperkingen, of het daarop gerichte compensatiegedrag, afhankelijk van specifieke voorwaarden?)

0 nee, er gelden geen specifieke voorwaarden voor het sociaal functioneren in arbeid.
1 ja, de cliënt is aangewezen op werk waarin meestal weinig of geen rechtstreeks contact met klanten vereist is (sommige beroepen in de dienstverlening).
2 ja, de cliënt is aangewezen op werk waarin meestal weinig of geen direct contact met patiënten of hulpbehoevenden vereist is (sommige beroepen in zorg- en hulpverlening).
3 ja, de cliënt is aangewezen op werk waarin zo nodig kan worden teruggevallen op directe collega's of leidinggevenden (géén solitaire functie).
4 ja, de cliënt is aangewezen op werk waarin doorgaans geen direct contact met collega's vereist is.
5 ja, de cliënt is aangewezen op werk dat geen leidinggevende aspecten bevat.
6 ja, er gelden overige specifieke voorwaarden voor het sociaal functioneren in arbeid, namelijk...................

RUBRIEK 3: AANPASSING AAN FYSIEKE OMGEVINGSEISEN

3.1 Hitte

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.2 Koude

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.3 Tocht

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..
3.4 Huidcontact

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.5 Beschermende middelen

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.6 Stof, rook, gassen en dampen

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.7 Geluidsbelasting

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.8 Trillingsbelasting

0 normaal, geen specifieke beperkingen
1 beperkt, namelijk..

3.9 Overige beperkingen van de fysieke aanpassingsmogelijkheden

0 normaal, geen specifieke overige beperkingen in fysieke aanpassingsmogelijkheden;
1 allergie, namelijk...
2 verhoogde vatbaarheid voor infecties, namelijk...........................
3 verzwakte huidbarrière, namelijk.......................................
4 andere beperkingen, namelijk..

3.10 Specifieke voorwaarden voor de aanpassing aan de fysieke arbeidsomgeving (is de aanpassing aan de arbeidsomgeving door de genoemde beperkingen, of het daarop gerichte compensatiegedrag, afhankelijk van specifieke voorwaarden?)

0 nee, er gelden geen specifieke voorwaarden voor de aanpassing aan de fysieke arbeidsomgeving
1 ja, er gelden specifieke voorwaarden voor de aanpassing aan de fysieke arbeidsomgeving, namelijk...

RUBRIEK 4: DYNAMISCHE HANDELINGEN

4.1 Dominantie

0 niet van toepassing
1 rechts
2 links

4.2 Lokalisatie beperkingen
0 noch rechts, noch links
1 rechts
2 links
3 tweezijdig

4.3 Hand- en vingergebruik

0 normaal, geen specifieke beperkingen bij het gebruik van handen en vingers in het dagelijks leven.
1 de bolgreep is beperkt.
2 de pengreep is beperkt.
3 de pincetgreep is beperkt.
4 de sleutelgreep is beperkt.
5 de cilindergreep is beperkt.
6 knijp/grijpkracht is beperkt.
7 fijn-motorische hand/vingerbewegingen zijn beperkt.
8 repetitieve hand/vingerbewegingen zijn beperkt.

4.4 Tastzin

0 normaal, geen specifieke beperkingen in het dagelijks leven
1 beperkt, namelijk...

4.5 Toetsenbord bedienen en muis hanteren

0 normaal, kan alle hiervoor benodigde bewegingen uitvoeren
1 beperkt, namelijk..

4.6 Werken met toetsenbord en muis

0 normaal, kan zo nodig gedurende het merendeel van de werkdag met toetsenbord en muis werken (professioneel tekstverwerken, cad/cam werk en elektronische verkoop)
1 licht beperkt, kan zo nodig gedurende de helft van de werkdag (ongeveer 4 uur) met toetsenbord en muis werken (beleidsmedewerker)
2 beperkt, kan zo nodig gedurende een beperkt deel van de werkdag (ongeveer 1 uur) met toetsenbord en muis werken (e-mailen)
3 sterk beperkt, kan gedurende minder dan een half uur per werkdag met toetsenbord en muis werken

4.7 Schroefbewegingen met hand en arm

0 normaal, kan alle hiervoor benodigde bewegingen uitvoeren.
1 beperkt, namelijk..

4.8 Reiken

0 normaal, kan met gestrekte arm reiken (afstand schouder-hand: ongeveer 70 cm)
1 licht beperkt, kan met licht gebogen arm reiken (afstand schouder-hand: ongeveer 60 cm)
2 beperkt, kan met sterk gebogen arm reiken (afstand schouder-hand niet meer dan 50 cm)

4.9 Frequent reiken tijdens het werk

0 normaal, kan zo nodig tijdens elk uur van de werkdag ongeveer 1.200 keer reiken (kassawerk in grootwinkelbedrijf, inpakwerk)
1 licht beperkt, kan zo nodig elk uur van de werkdag ongeveer 600 keer reiken
2 beperkt, kan zo nodig elk uur van de werkdag ongeveer 450 keer reiken
3 sterk beperkt, kan zo nodig elk uur van de werkdag ongeveer 300 keer reiken

4.10 Buigen

0 normaal, kan ongeveer 90 graden buigen (papiertje van de grond oprapen)
1 beperkt, kan ongeveer 60 graden buigen (tas van de grond oppakken)
2 sterk beperkt, kan ongeveer 45 graden buigen (kruimels uit stoelzitting oprapen)

4.11 Frequent buigen tijdens het werk

0 normaal, kan zo nodig elk uur van de werkdag ongeveer 600 keer buigen
1 licht beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 300 keer buigen
2 beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 150 keer buigen

3 sterk beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 50 keer buigen

4.12 Torderen

0 normaal, kan de romp tenminste 45 graden draaien (achterom kijken op de fiets; vóórin zittend een tas van de achterbank van de auto pakken)
1 beperkt, namelijk..

4.13 Duwen of trekken

0 normaal, kan ongeveer 15 kgf duwen of trekken (klemmende deur openen)
1 beperkt, kan ongeveer 10 kgf duwen of trekken (volle vuilniscontainer)
2 sterk beperkt, kan ongeveer 5 kgf duwen of trekken (deur met dranger openen)

4.14 Tillen of dragen

0 normaal, kan ongeveer 15 kg tillen of dragen (kleuter)
1 licht beperkt, kan ongeveer 10 kg tillen of dragen (peuter)
2 beperkt, kan ongeveer 5 kg tillen of dragen (zak aardappelen)
3 sterk beperkt, kan ongeveer 1 kg tillen of dragen (literpak melk)

4.15 Frequent lichte voorwerpen hanteren tijdens het werk

0 normaal, kan zo nodig tijdens elk uur van de werkdag ongeveer 600 keer voorwerpen van ruim 1 kg hanteren (orderverzamelaar)
1 licht beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 300 keer voorwerpen van ruim 1 kg hanteren

2 beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 150 keer voorwerpen van ruim 1 kg hanteren
3 sterk beperkt, kan zo nodig tijdens elk uur van de werkdag ongeveer 50 keer voorwerpen van ruim 1 kg hanteren

4.16 Frequent zware lasten hanteren tijdens het werk (ongeveer 10 keer per uur)

0 normaal, kan zo nodig tijdens ongeveer een uur per werkdag frequent lasten van ongeveer 15 kg hanteren
1 beperkt, kan niet tijdens ongeveer een uur per werkdag frequent lasten van ongeveer 15 kg hanteren

4.17 Hoofdbewegingen maken

0 normaal, kan het hoofd ongehinderd bewegen
1 beperkt, kan het hoofd beperkt bewegen
2 sterk beperkt, kan het hoofd niet of nauwelijks zijwaarts draaien
3 sterk beperkt, kan het hoofd niet of nauwelijks op en neer bewegen

4.18 Lopen

0 normaal, kan ongeveer een uur achtereen lopen (wandeling)
1 licht beperkt, kan ongeveer een half uur achtereen lopen (ommetje)
2 beperkt, kan ongeveer een kwartier achtereen lopen (naar de brievenbus)
3 sterk beperkt, kan minder dan 5 minuten achtereen lopen (binnenshuis)

4.19 Lopen tijdens het werk

0 normaal, kan zo nodig gedurende het merendeel van de werkdag lopen (postbode)
1 licht beperkt, kan zo nodig gedurende de helft van de werkdag (ongeveer 4 uur) lopen
2 beperkt, kan zo nodig gedurende een beperkt deel van de werkdag (ongeveer 1 uur) lopen
3 sterk beperkt, kan gedurende minder dan een half uur per werkdag lopen

4.20 Trappenlopen

0 normaal, kan tenminste in één keer twee trappen op en af (2 verdiepingen woonhuis)
1 licht beperkt, kan tenminste in één keer een trap op en af (1 verdieping woonhuis)
2 beperkt, kan tenminste in één keer een trap op óf af (1 verdieping woonhuis)
3 sterk beperkt, kan in één keer slechts een bordestrapje op- of aflopen

4.21 Klimmen

0 normaal, kan tenminste een ladder op en af (1 verdieping)
1 licht beperkt, kan tenminste een huishoudtrap op en af
2 beperkt, kan tenminste een opstapje op en af
3 sterk beperkt, kan geen opstap maken

4.22 Knielen of hurken

0 normaal, kan knielend of hurkend met de handen de grond bereiken (een muntstuk oprapen).
1 beperkt, kan niet of nauwelijks knielend of hurkend met de handen de grond bereiken.

4.23 Overige beperkingen van het dynamisch handelen

0 normaal, geen specifieke overige beperkingen van het dynamisch handelen in het dagelijks leven
1 specifieke overige beperkingen, namelijk.................................

4.24 Specifieke voorwaarden voor het dynamisch handelen in arbeid (is het dynamisch handelen in arbeid door de genoemde beperkingen, of het daarop gerichte compensatiegedrag, afhankelijk van specifieke voorwaarden?)

0 nee, er gelden geen specifieke voorwaarden voor het dynamisch handelen in arbeid
1 ja, er gelden specifieke voorwaarden voor het dynamisch handelen, namelijk..........

RUBRIEK 5: STATISCHE HOUDINGEN

5.1 Zitten

0 normaal, kan ongeveer 2 uur achtereen zitten (autorit)
1 licht beperkt, kan ongeveer een uur achtereen zitten (film)
2 beperkt, kan ongeveer een half uur achtereen zitten (maaltijd)
3 sterk beperkt, kan minder dan een kwartier achtereen zitten (tv-journaal)

5.2 Zitten tijdens het werk

0 normaal, kan zo nodig gedurende vrijwel de gehele werkdag zitten (assemblagewerk, kassawerk, uitvoerend administratief werk)
1 licht beperkt, kan zo nodig gedurende het grootste deel van de werkdag zitten (niet meer dan 8 uur)
2 beperkt, kan zo nodig gedurende de helft van de werkdag zitten (ongeveer 4 uur)
3 sterk beperkt, kan gedurende minder dan 4 uur per werkdag zitten

5.3 Staan

0 normaal, kan ongeveer 1 uur achtereen staan (toeschouwer bij sportwedstrijd)
1 licht beperkt, kan ongeveer een half uur achtereen staan (wachttijd voor attractie in pretpark)
2 beperkt, kan ongeveer een kwartier achtereen staan (afwassen)
3 sterk beperkt, kan minder dan ongeveer 5 minuten achtereen staan (tanden poetsen)

5.4 Staan tijdens het werk

0 normaal, kan zo nodig gedurende het merendeel van de werkdag staan (verkoopfuncties, productiefuncties)
1 licht beperkt, kan zo nodig gedurende de helft van de werkdag staan (ongeveer 4 uur)
2 beperkt, kan zo nodig gedurende een beperkt deel van de werkdag staan (ongeveer 1 uur)
3 sterk beperkt, kan gedurende minder dan een half uur per werkdag staan

5.5 Geknield of gehurkt actief zijn

0 normaal, tenminste 5 minuten achtereen (tuinieren)
1 beperkt, minder dan 5 minuten achtereen (deur aanrechtkastje afnemen)

5.6 Gebogen en/ of getordeerd actief zijn

0 normaal, kan tenminste 5 minuten achtereen gebogen en/ of getordeerd actief zijn (stoep vegen)
1 beperkt, kan minder dan 5 minuten achtereen gebogen en/ of getordeerd actief zijn (schoenveters strikken)

5.7 Boven schouderhoogte actief zijn

0 normaal, tenminste 5 minuten achtereen (gordijnen ophangen)
1 beperkt, minder dan 5 minuten achtereen (gloeilamp verwisselen)

5.8 Het hoofd in een bepaalde stand houden tijdens het werk

0 normaal, kan zo nodig gedurende het merendeel van de werkdag het hoofd in een bepaalde stand houden (beeldschermwerk, kwaliteitscontrole)
1 licht beperkt, kan zo nodig gedurende de helft van de werkdag het hoofd in een bepaalde stand houden (ongeveer 4 uur)
2 beperkt, kan zo nodig gedurende een beperkt deel van de werkdag het hoofd in een bepaalde stand houden (ongeveer 1 uur)
3 sterk beperkt, kan gedurende minder dan ongeveer een half uur per werkdag het hoofd in een bepaalde stand houden
5.9 Afwisseling van houding

0 normaal, geen specifieke opeenvolging van verschillende houdingen vereist
1 specifieke eisen aan afwisseling van houdingen, namelijk................

5.10 Overige beperkingen van statische houdingen

0 normaal, geen specifieke overige beperkingen van statische houdingen in het dagelijks leven
1 specifieke overige beperkingen, namelijk................................

5.11 Specifieke voorwaarden voor statische houdingen in arbeid (zijn statische houdingen in arbeid door de genoemde beperkingen, of het daarop gerichte compensatiegedrag, afhankelijk van specifieke voorwaarden?)

0 nee, er gelden geen specifieke voorwaarden voor statische houdingen in arbeid
1 ja, er gelden specifieke voorwaarden voor statische houdingen, namelijk.................

RUBRIEK 6: WERKTIJDEN

6.1 Perioden van het etmaal

0 normaal, kan zo nodig op elk uur van het etmaal werken, ook 's nachts
1 beperkt, kan 's nachts niet werken (00.00 - 06.00 uur)
2 beperkt, kan 's avonds niet werken (18.00 - 24.00 uur)

6.2 Uren per dag

0 normaal, kan gemiddeld tenminste 8 uur per dag werken
1 enigszins beperkt, kan gemiddeld ongeveer 8 uur per dag werken
2 licht beperkt, kan gemiddeld ongeveer 6 uur per dag werken
3 beperkt, kan gemiddeld ongeveer 4 uur per dag werken
4 zeer beperkt, kan gemiddeld ongeveer 2 uur per dag werken

6.3 Uren per week

0 normaal, kan gemiddeld tenminste 40 uur per week werken
1 enigszins beperkt, kan gemiddeld ongeveer 40 uur per week werken
2 licht beperkt, kan gemiddeld ongeveer 30 uur per week werken
3 beperkt, kan gemiddeld ongeveer 20 uur per week werken
4 zeer beperkt, kan gemiddeld ongeveer 10 uur per week werken

6.4 Overige beperkingen ten aanzien van werktijden

0 normaal, er zijn geen specifieke overige beperkingen ten aanzien van werktijden
1 specifieke overige beperkingen, namelijk.............................

RUBRIEK 7: OVERIGE BELASTINGPUNTEN

7.1 Probleem oplossen
Geen beoordelingspunt op FML, wel bij beschrijving functiebelasting

7.2 Kruipen
Geen beoordelingspunt op FML, wel bij beschrijving functiebelasting

7.3 Getordeerd actief zijn
Zie 5.6

Appendix 2. Instructions for Doctor and Patient Simulation

Doctor Simulation Instructions

1. Preparation:
   - Read all instructions thoroughly before beginning the simulation.
   - You have a case in front of you containing some personal details of a simulation patient.
   - A number of recommendations are provided to assist you (if desired). Consider this as information you have previously prepared with your supervisor. You can choose whether or not to use this information.
2. Consultation:
   - Conduct an occupational health consultation with the simulation patient as you normally would.
   - During or after the conversation, write a medical report as you are accustomed to.
3. Documentation:
   - Write a letter to the employer/employee.
   - Fill out an FML-form.
4. Review and Feedback:
   - Review the transcription created by the Whisper-module and rate the result on a scale of 1 to 5, where 1 indicates very dissatisfied and 5 indicates very satisfied.
   - Similarly, rate I) the medical report II) the letter, and III) the FML generated by ChatGPT-4 modules.
5. Questions and Answers:
   - Write down three questions you would like to have answered about your case (e.g., questions about etiology, diagnosis, treatment, prognosis, etc.). You are free to choose the questions.
   - Independently find the answers in the NVAB guidelines and note these down.
   - Ask the module of ChatGPT-4 the same questions.
   - Determine if the answers match your answers from the protocol.
   - Assess if you perceive the answer as relevant (yes/no).
   - Rate the module's performance on a scale of 1 to 5, where 1 indicates very dissatisfied and 5 indicates very satisfied.

Patient Simulation Instructions

1. Preparation:
   - Read all instructions thoroughly before beginning the simulation.
   - You have a case in front of you with some details of a simulation patient.
2. Role Playing:
   - You play an employee/patient who is ill and has an appointment with the occupational physician / A(N)IOS occupational health physician.
   - Use the information from the case as a guide for the consultation. Try to stick to the main points of the case, but you are free to mention other or additional information during the conversation.

Appendix 3. Case generation prompt

Generate a medical patient case about … experiencing … for a simulation between a doctor and a patient. Provide an overview for both the patient and the doctor. For the patient, generate information on the following aspects: First and last name, date of birth, age, gender, job position, number of working hours per week, first day of sick leave, anamnesis, current symptoms, medication, treatment, medical history, exercise and physical activity, stress factors, sleep and energy levels, daily narrative, reintegration, and work factors. For the doctor, generate the following information: First and last name of the patient, date of birth, age, gender, and advice.

Appendix 4. Subgroup analyses

The subgroup analyses comparing the large and medium Whisper models, as well as the use of external and internal microphones, are summarized in Appendix 4.1, 4.2, 4.3 and

4.4. Mann-Whitney U-tests were used to compare efficiency and accuracy variables, while Kruskal-Wallis H-tests were employed for satisfaction variables.

1. Whisper-subgroups analyses: Whisper:

No significant difference was found in the WER between the medium and large Whisper models with a mean of 13.3% (SD = 5.8%) and 12.9% (SD = 8.1%) respectively (U = 88.0, Z = -.494, p = .642). However, the generating time was significantly, shorter for the medium Whisper model compared to the large Whisper model; the mean difference was namely 13.1 minutes (U = 30.5, Z = -3.194, p < .001). The satisfaction rates did not differ significantly between the models (large: mean = 3.1, SD = 0.9; medium: mean = 3.3, SD = 0.7; H = .269, p = .604).

2. Whisper-subgroups analyses: Medical Summarizer:

The percentage of correct elements of the Medical Summarizer did not significantly differ, with a mean of 62.3% (SD =14.7%) and 59.8% (SD = 16.7%) for the medium and large model respectively (U = 93.0, Z = -.495, p = .641). No significant difference was found with regards to the readability scores either, with the medium model having a mean score of 191.1 (SD = 4.2) and the large model a mean score of 192.1 (SD = 4.5) (U = 82.5, Z = -.957, p = .350). Nevertheless, the number of correctly categorized elements was significantly higher for the large model, with a mean of 98.5% (SD = 1.9%) compared to a mean of 93.3% (SD = 6.4%) (U = 46.5, Z = -2.252, p = .011). The generating time for the medical summarizer did not show a significant difference between the models (U = 68.0, Z = -1.808, p = .123), with the medium model needing 0.3 minutes (SD = 0.5 minutes) to generate the summary and the large model 0.7 minutes (SD = 0.6 minutes) on average. The satisfaction rates did not significantly differ between the models (large: mean = 4.2, SD = 1.0; medium: mean = 4.3, SD = 0.9; H = .035, p = .852)

3. Whisper-subgroups analyses: Letter Generator:

With regards to the Letter Generator, the readability score was significantly higher when utilizing the large Whisper model, with a mean readability of 181.1 (SD = 10.5) compared to the medium model, which had a mean readability score of 175.4 (SD = 10.4) (U = 53.0, Z = -2.218, p = .026). There was no significant difference in the number of correct elements between the models, with the large model achieving 32.8% correct elements (SD = 25.2%) and the medium model 24.0% correct elements (SD = 23.1%) (U = 80.0, Z = -1.069, p = .307). The generating time also did not significantly differ, with the large model averaging 0.6 minutes (SD = 0.7 minutes) and the medium model averaging 0.3 minutes (SD = 0.5 minutes) (U = 70.0, Z = -1.381, p = .266). The satisfaction rates did not significantly differ between the models (large: mean = 3.1, SD = 0.9; medium: mean = 3.1, SD = 0.8; H = .013, p = .909).

4. Whisper-subgroups analyses: Document Generator:

Moreover, there were no significant differences in the number of correct elements of the Document Generator when using the large Whisper model, which had a mean of 80.1% (SD = 8.8%) correct elements, and when using the medium Whisper model, which had 83.9% (SD = 8.2%) (U = 82.0, Z = -.975, p = .350). The generating time also did not significantly differ, with the large model taking 5.0 minutes (SD = 6.0 minutes) and the medium model taking 3.2 minutes (SD = 4.4 minutes) (U = 72.5, Z = -1.556, p = .171). The satisfaction rates did not significantly differ between the models (large: mean = 2.3, SD = 0.8; medium: mean = 2.6, SD = 2.2; H = .590, p = .442).

APPENDIX 4.1. MANN-WHITNEY U-TESTS BETWEEN LARGE AND MEDIUM WHISPER MODELS OF EFFICIENCY AND ACCURACY VARIABLES

Variable	Whisper model	N	Mean rank	U	Z	p (exact)
Speech-to-text converter						
Words-error-rate	Medium	18	15.61	88.0	-.494	.642
	Large	11	14.00			
Generating time	Medium	19	11.61	30.5	-3.194	**<.001**
	Large	11	22.23			
Medical summarizer						
Number of correct elements	Medium	19	16.11	93.0	-.495	.641
	Large	11	14.45			
Number of correctly categorized elements	Medium	19	12.45	46.5	-2.252	**.011**
	Large	11	20.77			
Readability score	Medium	19	14.34	82.5	-.957	.350
	Large	11	17.50			
Generating time	Medium	19	13.58	68.0	-1.808	.123
	Large	11	18.82			
Letter generator						
Number of correct elements	Medium	19	14.21	80.0	-1.069	.307
	Large	11	17.73			
Number of correctly categorized elements	Medium	19	14.58	87.0	-.898	.471
	Large	11	17.09			
Readability score	Medium	19	12.79	53.0	-2.218	**.026**
	Large	11	20.18			
Generating time	Medium	19	13.68	70.0	-1.381	.266
	Large	10	17.50			
Document generator						
Number of correct elements	Medium	19	16.68	82.0	-.975	.350
	Large	11	13.45			
Generating time	Medium	19	13.82	72.5	-1.556	.171
	Large	11	18.41			

APPENDIX 4.2. KRUSKAL-WALLIS H-TESTS BETWEEN LARGE AND MEDIUM WHISPER MODELS OF SATISFACTION VARIABLES

Variable	Whisper model	N	Mean rank	H	p
Speech-to-text converter satisfaction rate	Medium	19	16.08	.269	.604
	Large	11	14.50		
Medical summarizer satisfaction rate	Medium	19	15.71	.035	.852
	Large	11	15.14		

Letter generator satisfaction rate	Medium	19	15.37	.013	.909
	Large	11	15.73		
Document generator satisfaction rate	Medium	18	15.92	.590	.442
	Large	11	13.50		

5. Microphone-subgroups analyses

In the comparison between external and internal microphones, significant differences were observed in several metrics. The WER for the speech-to-text converter was significantly lower (U = 16.0, Z = -2.150, p = .030), and the generating time was significantly shorter (U = 14.0, Z = -2.706, p = .005) when using an external microphone. For the medical summarizer, the readability score was significantly higher with the external microphone (U = 25.0, Z = -2.109, p = .037), though the generating time showed only a trend towards significance (U = 30.5, Z = -2.050, p = .074). Similarly, for the letter generator, the readability score approached significance (U = 28.0, Z = -1.921, p = .057), and the generating time showed a trend towards being shorter with the external microphone (U = 29.0, Z = -2.155, p = .078). However, for the document generator, neither the number of correct elements (U = 61.0, Z = -.084, p = .957) nor the generating time (U = 33.0, Z = -1.855, p = .108) differed significantly between microphone types. The satisfaction rates also highlighted significant differences. The external microphone received higher satisfaction rates for both the speech-to-text converter (H = 8.375, p = .004) and the letter generator (H = 4.143, p = .042). No significant differences in satisfaction rates were observed for the medical summarizer (H = .767, p = .381) and document generator (H = .394, p = .530) between the microphone types.

APPENDIX 4.3. MANN-WHITNEY U-TESTS BETWEEN EXTERNAL AND INTERNAL MICROPHONES OF EFFICIENCY AND ACCURACY VARIABLES

Variable	Microphone	N	Mean rank	U	Z	p (exact)
Speech-to-text converter						
Words-error-rate	Internal	4	23.50	16.0	-2.150	**.030**
	External	25	13.64			
Generating time	Internal	5	25.20	14.0	-2.706	**.005**
	External	25	13.56			
Medical summarizer						
Number of correct elements	Internal	5	13.80	54.0	-.473	.666
	External	25	15.84			
Number of correctly categorized elements	Internal	5	16.20	59.0	-.197	.872
	External	25	15.36			
Readability score	Internal	5	23.00	25.0	-2.109	**.037**
	External	25	14.00			
Generating time	Internal	5	21.90	30.5	-2.050	.074
	External	25	14.22			
Letter generator						
Number of correct elements	Internal	5	17.40	53.0	-.536	.627
	External	25	15.12			
Number of correctly categorized elements	Internal	5	17.20	54.0	-.564	.666
	External	25	15.16			
Readability score	Internal	5	22.40	28.0	-1.921	.057
	External	25	14.12			
Generating time	Internal	5	21.20	29.0	-2.155	.078
	External	24	13.71			
Document generator						
Number of correct elements	Internal	5	15.20	61.0	-.084	.957
	External	25	15.56			
Generating time	Internal	5	21.40	33.0	-1.855	.108
	External	25	14.32			

APPENDIX 4.4. KRUSKAL-WALLIS H-TESTS BETWEEN EXTERNAL AND INTERNAL MICROPHONES OF SATISFACTION VARIABLES

Variable	Microphone	N	Mean rank	H	p
Speech-to-text converter satisfaction rate	Internal	5	6.00	8.375	**.004**
	External	25	17.40		
Medical summarizer satisfaction rate	Internal	5	12.60	.767	.381
	External	25	16.08		
Letter generator satisfaction rate	Internal	5	8.60	4.143	**.042**
	External	25	16.88		
Document generator satisfaction rate	Internal	5	12.90	.394	.530
	External	25	15.44		

Appendix 5. Summary Table

Task	Model	Accuracy (Mean ± SD)	Efficiency (Mean ± SD) of AI Speech and Text model	Efficiency (Mean ± SD) of Human	Satisfaction Rate (Mean ± SD) on 5-point scale
Speech-to-text transcription	Whisper	WER = 13.1 ± 6.6%	18.4 ± 11.6 minutes	25.0 ± 10.9 minutes	3.2 ± 0.8
Writing a medical summary	Medical Summarizer	Incorrect elements = 38.6 ± 15.2%	0.5 ± 0.6 minutes	14.7 ± 5.9 minutes	4.2 ± 0.9
Writing a letter	Letter Generator	Incorrect elements = 90.4 ± 63.4%	0.4 ± 0.6 minutes	3.4 ± 2.1 minutes	3.1 ± 0.9
Filling in functional capabilities list	Document Generator	Incorrect elements = 17.5 ± 8.5%	3.9 ± 5.0 minutes	5.3 ± 2.7 minutes	2.5 ± 1.1
Answering questions	Medical Protocol Assistant	Protocol-based replies = 2.6 ± 0.6%	1.3 ± 0.6 minutes	5.0 ± 2.3 minutes	4.4 ± 0.8