Completeness, Correctness and Conciseness of Physician-written *versus* Large Language Model Generated Patient Summaries Integrated in Electronic Health Records

Rosanne C. Schoonbeek[1,2], MD, PhD; Jessica Workum[3], MD, MSc; Stephanie C.E. Schuit[4], MD, PhD; Job N. Doornberg[5], MD, PhD; Tom P. van der Laan[1,2], MD, PhD; Charlotte M.H.H.T. Bootsma-Robroeks[2,6], MD, PhD
On behalf of the Applied Artificial Intelligence in Healthcare Consortium

Affiliations

1. Department of Otolaryngology – Head and Neck Surgery, University Medical Center, Groningen, the Netherlands
2. Department of Medical Information Technology, University Medical Center, Groningen, the Netherlands
3. Department of Intensive Care, Elizabeth-TweeSteden Hospital, Tilburg, the Netherlands
4. Board of Directors, University Medical Center, Groningen, the Netherlands
5. Department of Trauma Surgery / Orthopedics, University Medical Center, Groningen, the Netherlands
6. Department of Pediatrics, Pediatrics Nephrology, Beatrix Children's Hospital, University Medical Center, Groningen, the Netherlands

Abstract

Background: The development of large language models (LLMs) has resulted in many applications, including their implementation in electronic health records (EHRs). Especially for complex, time-consuming cognitive tasks, such as summarizing medical charts, the use of LLM could support the clinician in reducing administrative burden. In this study, we compared physician-written summaries with LLM-generated medical summaries integrated in the EHR in a non-English clinical environment.

Methods: A non-inferiority validation study was designed to compare physician-written and LLM-generated medical summaries. A total of 400 summaries were evaluated. Comparison was performed using objective and subjective evaluation by physician evaluators on summary quality (completeness, correctness and conciseness), preference and trust.

Findings: Mean writing time for the summary for the physicians was 7 minutes and 15·7 seconds for the LLM. The completeness and correctness of the LLM-generated summaries did not differ significantly from the physician summaries. LLM-generated summaries were less concise (3·0 vs. 3·5, p=0·001). The overall evaluation scores for physician vs. LLM summaries were not different (3·4 vs. 3·3, p=0·373). There was a preference (or equal score) for the LLM-generated summary compared to the physician-written summary (57% vs. 43%). Trust in both the physician and LLM summaries was similar: 77 vs. 81% of the summaries were trusted enough to be used in clinical decision making (p=0·187). Interobserver variability showed excellent reliability (ICC 0·975).

Interpretation: This study found that LLM-generated summaries are comparable to physician-written summaries in terms of completeness and correctness, though slightly less concise. These findings suggest that LLMs might be effective in reducing clinicians' administrative burden without compromising summary quality. This supports the body of evidence that this functionality can be safely integrated and used in the EHR and reinforces the potential of LLMs to enhance clinical documentation processes.

Funding: None.

Research in context

<u>Evidence before this study:</u>

Previous research of Van der Veen and colleagues has shown that LLM-generated summaries can outperform medical experts using generative AI tools outside of the EHR in the English language. These promising results call for studies to validate actual implementation of this technology within EHRs, and in non-English clinical environments.

<u>Added value of this study:</u>

This study provides a robust prospective validation of physician-written summaries compared to LLM-generated summaries, where physicians were instructed to write summaries to set a high standard instead of using already available summaries in the EHR. This is the first study to demonstrate that LLM-generated summaries are non-inferior to physician-written summaries with regard to correctness and completeness. Additionally, LLM-generated summaries were deemed trustworthy for clinical use.

<u>Implications of all the available evidence:</u>

The quality of LLM-generated medical summaries are at the least comparable to physician-written summaries regarding correctness and completeness. This study validated LLM-generated medical summaries implemented in the EHR in a non-English clinical environment. This validation study supports the safe integration of LLMs into EHRs. Integrating LLMs into clinical workflows can significantly reduce the administrative burden on healthcare professionals, allowing them to focus more on direct patient care. Further evolution of the prompts and training of LLMs on medical texts will presumably result in performance superior to physicians. Future research should explore the long-term impacts of

LLM integration on clinical workflows, patient outcomes, and the adaptability of these models across various languages and healthcare settings.

Total word count manuscript: 3346 words

Introduction

The use of Generative Artificial Intelligence (GenAI), and in particular Large Language Models (LLMs), in healthcare has the potential to reduce the administrative burden of clinicians without compromising quality of care, thus supporting a sustainable healthcare system.[1–3] The development of these LLMs has resulted in many potential applications in healthcare, such as medical note summarizations and clinical decision-making.[3,4] LLMs use deep learning to process and interpret human language. They undergo a multi-layered training process, including pre-training and fine-tuning, and can therefore generate specific outputs based on giving inputs for particular use cases. Currently, the Generative Pretrained Transformer 4 (GPT-4) model by OpenAI (San Francisco, CA, USA) is evaluated as the best LLM.[3] LLMs were predominantly trained and tested using the English language and with a US-centric point of view.[5,6] As numerous research groups explore the potential applications of LLMs in healthcare, there is a growing demand for robust clinical validation, so that they can be safely integrated into the Electronic Health Records (EHRs) and applied in daily clinical practice. However, this integrated EHR functionality has not been scientifically tested yet in non-English clinical environments.

Healthcare professionals spend a significant amount of time on administrative duties, affecting physicians' perceptions of being able to deliver high-quality care, career satisfaction, burnout, and even the likelihood of continuing clinical practice.[7] A large part of the administrative tasks consists of ways of summarizing patient files, for example, to prepare for outpatient visits or when writing hospital discharge letters. This complex and time-consuming cognitive process is prone to inconsistencies and human errors. These issues are magnified when patients are treated by multiple healthcare providers, as this results in fragmented EHRs scattered across different healthcare providers.

If GenAI can accurately and consistently generate medical summaries, it could save a significant amount of time for healthcare professionals.[8–11] Generative AI presents a compelling opportunity to revolutionize note summarization, streamlining workflow efficiency and optimizing patient outcomes.[12] Incorporating notes from shared care centers into these summaries can provide a comprehensive understanding of the patient's health status, further enhancing patient care and safety.

In order to safely use summarization by GenAI, validation and benchmarking against current medical practice (i.e., physician written summaries) is needed to result in a trustworthy application. In this study, we compared LLM-generated medical summaries with physician-written medical summaries, using a non-inferiority study design. Summaries were compared by independent physicians as well as objective protocols. Furthermore, we assessed clinicians' preference and level of trust for each summary. To the best of our knowledge, this is the first clinical trial to validate LLM-generated medical summaries embedded in the EHR and in a non-English clinical environment.

Methods

Physician-written Summaries

For this non-inferiority validation study, 60 Dutch physicians across 10 departments in a large Dutch academic hospital were recruited to participate in this study in February 2024. Their experience was defined as the numbers of years of clinical practice since attaining the medical license. For each department, 5 patients were selected (*Figure 1*), to a total of 50 discrete patient records. Their corresponding anonymized, dummy EHRs were frozen. The participating physicians (physician writers) were instructed to write a summary of the patient files, as if they would prepare for an outpatient visit. The instructions were to write the summaries in a similar timeframe as they would normally use for outpatient clinic preparation. Additionally, they were asked to time the duration of these preparations for each summary.

A total of 42 physicians (70%) completed all 5 summaries (n=210 summaries). Baseline characteristics of the participating physician population were collected.

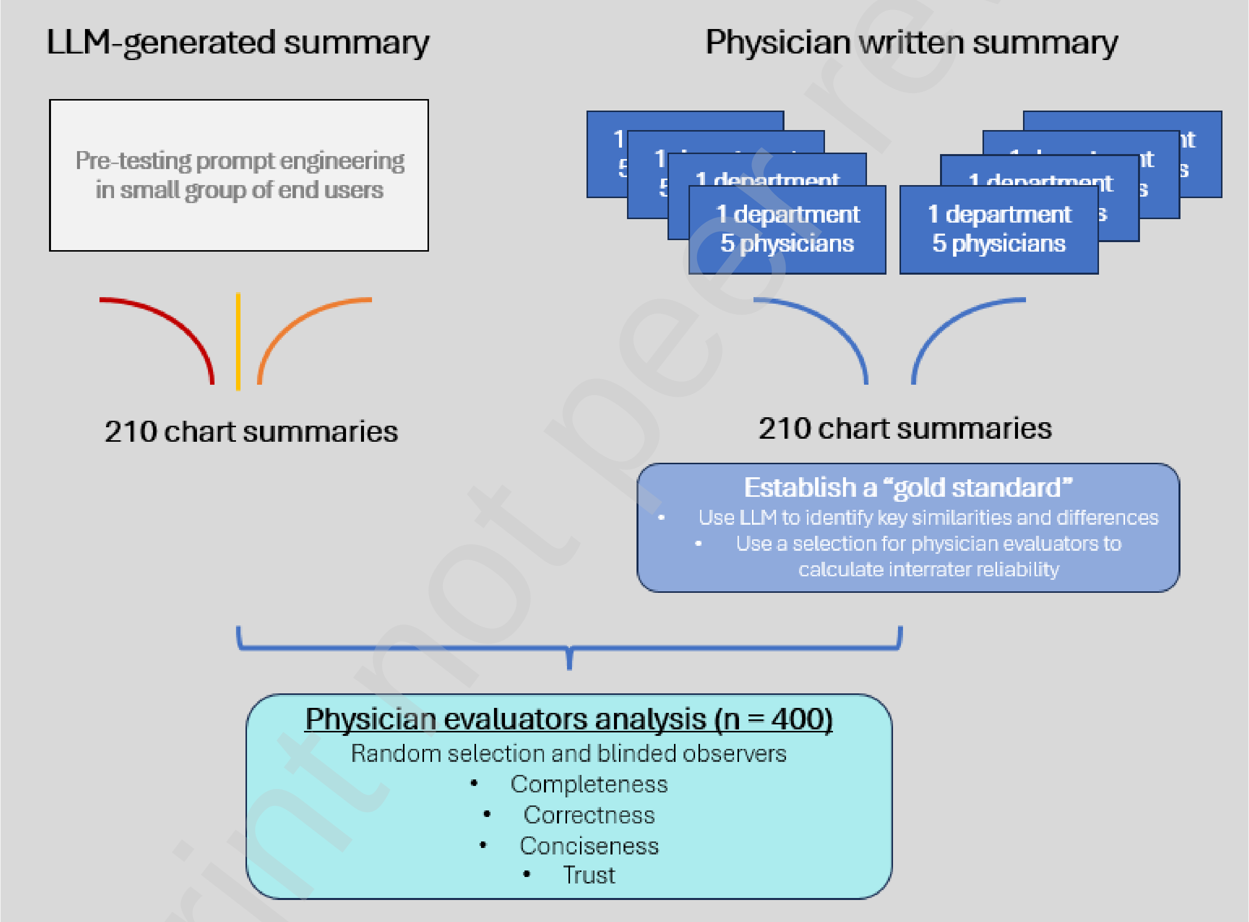

LLM-generated summary
Physician written summary
Pre-testing prompt engineering in small group of end users
1 department 5 physicians
1 department 5 physicians
210 chart summaries
210 chart summaries
Establish a "gold standard"
Use LLM to identify key similarities and differences
Use a selection for physician evaluators to calculate interrater reliability
Physician evaluators analysis (n = 400)
Random selection and blinded observers
Completeness
Correctness
Conciseness
Trust

Figure 1. The non-inferiority study design.

LLM Generated Summaries - Large Language Models within the EHR

LLM-generated summaries were generated for the selected patients via Microsoft's Azure OpenAI, using the GPT-4 model, through the EHR (Epic Systems Corporation, Verona, WI, USA).

Prompt engineering (the act of writing sound instructions to the model) was performed by a team of multicenter medical and technical experts in an iterative manner. The final prompt used in this study is added in the Supplementary Information.

Comparing Physician-written versus LLM-Generated Summaries

Physician-written summaries were compared with LLM-generated summaries using objective and subjective measures. For objective measures, the validated ROUGE and BLEU scores for analysis of computational linguistics and natural language processing were used[13–15]. A higher ROUGE and BLEU score reflect higher textual similarity between the summaries. The ROUGE score measures the overlap of n-grams (contiguous sequence of n items, such as words, from a given sample of text) between the LLM-generated summary and the physician-written (reference) summary. For this study, an n-gram of 1 was used. The BLEU score assesses summary quality by measuring the precision of n-grams, which means the comparison was done at the level of individual words.

Since automated metrics do not directly reflect summary quality, readability and accuracy, a selection of the summaries (total: n=400) was evaluated by ten physician evaluators. These physician evaluators reviewed paired LLM-generated summaries and physician-written summaries, presented in a blinded and randomized order to a total of 40 summaries per evaluator. Physicians' evaluation was measured using a 5-point Likert scale of three domains (derived from Van Veen et al. 2024[3]):

- o *Completeness*: captures recall, amount of relevant clinical details. "Which summary more completely captures important information?"
- o *Correctness*: captures precision, a summary without errors. "Which summary includes less false information?"
- o *Conciseness*: decreasing the amount of irrelevant information. "Which summary contains less non-important information?"

Furthermore, these physician evaluators were asked which summary they believed was the LLM-generated summary, which summary they would prefer, and which summary they would trust during clinical practice.

Statistical analysis

SPSS ® Statistics version 28·0 (Armonk, NY: IBM Corp.) was used to analyze data. Descriptive statistics for baseline characteristics are presented depending on their distribution. For comparison, Student's *t* tests, the Mann-Whitney *U* test or the χ^2 test were used, depending on the type of variable studied. A p<0·05 was considered statistically significant.

For the natural language processing analyses (baseline characteristics, ROUGE, BLEU scores), Python (version 3·12·2) was used (Natural Language Toolkit (NLTK) packages). NLTK scores are presented as a percentage and are designed to compare two sets of texts. The ROUGE-1 recall represents the percentage of words that match between the LLM-generated summary vs. the physician-written summary (where 100 represents two equal texts). The BLEU score reflects the number of similar words divided by the total words (as percentage), again, with 100 representing two equal texts.

The combinations of Likert-scale scores were calculated for each summary for each observer, and paired t-tests were used for statistical analyses. Two-sided p-values are used for comparisons of paired data. Preference for either the physician-written summary, the LLM-generated summary or equal assessment was established for each pair of summaries for each observer. Inter-rater reliability was calculated using the intraclass correlation coefficient (ICC).

Ethical considerations

The functionality of generating chart summaries using GenAI is intended as a tool to reduce the administrative burden, in line with European regulation. The AI-application as described in this study does not fall within the Medical Device Regulation scope. It is not used as a medical device, as it does not provide clinical decision support, and the generated output is always revised by the responsible clinician, using the 'human in the loop' principles.

The study was prospectively registered (no. 19035). Our institutional review board granted permission for this study according to the declaration of Helsinki (ref. M24·328217).

The generated output is kept within the secured environment of the hospital and was not shared with the EHR provider or OpenAI. The privacy officers were closely involved in the setup of this study.

Results

Baseline characteristics

A total of 420 summaries were written (n=210) or generated (n=210). The mean age and years of experience for the physicians writers (n=42) and the physician evaluators (n=10) were comparable (39·9 vs. 39·1 years of age, and 12·5 vs. 13·7 years of expertise, p=0·981 and p=0·422, respectively, *Table 1*). The mean writing time for the summary for the physicians was 7 minutes (± 5). The mean generating time for the LLM-summaries was 15·7 seconds (95% confidence interval 2·1).

Objective measures

Physician-written summaries were significantly shorter compared to LLM-generated summaries in both word count (60 vs. 100 words, p<0·001) and characters (463 vs. 696 characters, p<0·001). The overall ROUGE recall score was 24·8, and the overall BLEU score was 14·2 (*Table 2*).
Hallucinations were not observed.

Performance of LLM-generated summaries vs. physician-written summaries

The combined scores of completeness, correctness, and conciseness for LLM vs. physician summaries were not significantly different (3·3 vs. 3·4, p=0·373). Completeness and correctness of the LLM-generated summaries did not significantly differ compared to the physician-written summaries (*Table 3*). LLM-generated summaries were significantly less concise (3·0 vs. 3·5, p=0·001).
An example of the scoring system is provided in *Figure 2*. Overall, there was a preference (or equal score) for the LLM-generated summary compared to the physician-written summary (57% vs. 43%, *Figure 3*). Evaluators were able to correctly identify the LLM-generated summaries in the majority of the summaries (84%). Trust in both the physician and LLM summaries was similar: 77 vs. 81% of the summaries were trusted enough to be used in clinical decision making (p=0·187). Interobserver variability showed excellent reliability (ICC 0·975).

Table 1. Baseline characteristics of the physician writers and the physician evaluators.

Baseline characteristics	**Physician writers** ***n = 42***	**Physician evaluators** ***n = 10***
Age (mean ± SD)	39·9 ± 8·9	39·1 ± 6·7
Departments		
ENT Surgery	5	1
Radiotherapy	4	
Orthopedics	5	1
Pediatrics		
General	15	4
Cardiology	5	
Intensive care		3
Gynecology	2	
Urology	3	
Internal medicine	3	1
Years of experience	12·5 ± 8·6	13·7 ± 6·7
Time to write summary		
Mean (min) ± SD	7·2 ± 5·0	

Figure 2. Example of summaries (upper) and their evaluation as used for the physician evaluators (lower). (in red: mistake by physician, in green: additional valuable information in physician summary).

LLM-generated summary

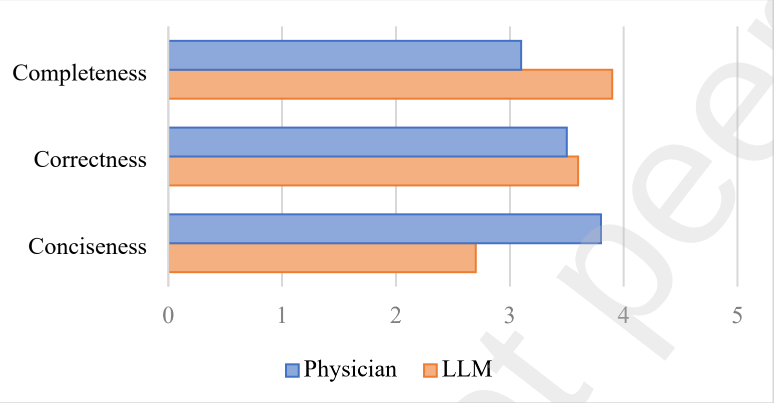
The patient has recurrent idiopathic subglottic stenosis, for which she has undergone multiple therapeutic microlaryngoscopies with balloon dilation. On [date], she underwent an uncomplicated therapeutic microlaryngoscopy, during which four incisions were made and a balloon dilation to a diameter of 15mm was performed. After the procedure, she experienced mild airway symptoms but showed improved airflow four weeks after the seventh dilation. The plan was to wait for the recovery from a possible viral cause of the symptoms. If the symptoms persisted, a check-up with flexible endoscopy would be performed to rule out other causes of reduced efficacy.

Physician-written summary I

The patient has been known to have idiopathic subglottic stenosis since [month][year], having undergone multiple (7 times) endoscopic treatments with steric incision and dilation between [month][year] and [month][year]. No other relevant medical history.

Physician-written summary II

Patient has been experiencing shortness of breath since [year] due to subglottic stenosis of unknown etiology, for which they have undergone more than 10 dilation procedures under anesthesia since that time, with the latest procedure in [month][year]. At the last check-up, peak flow was 320 L/min.



	LLM summary	Physician summary
Total	3·4	3·5
Trust	80%	70%
Guessed for AI	90%	10%

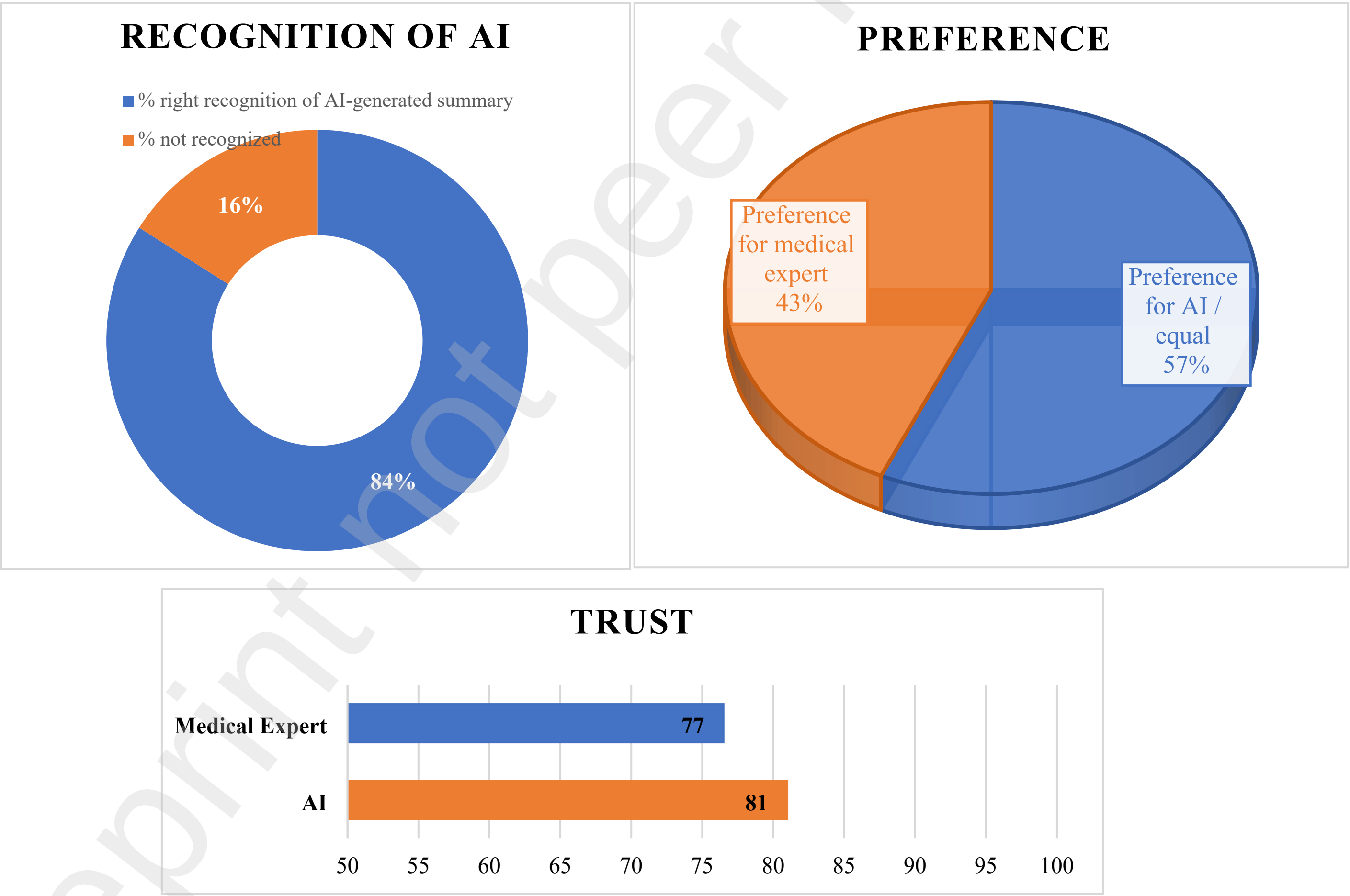
Table 2. Objective measurements.

Baseline		ROUGE-1		BLEU
Mean word count (n)	Mean characters (n)	Recall	Precision	BLEU-1
80	580	24·8	14·7	14·2

Table 3. Performance of LLM vs. Physician using a 5-point Likert scale of three domains.

	LLM	Physician	p-value
Completeness	3·5 ± 0·5	3·3 ± 0·6	0·517
Correctness	3·3 ± 0·3	3·3 ± 0·4	0·938
Conciseness	3·0 ± 0·4	3·5 ± 0·5	**0·001**
Overall scores	**3·3 ± 0·3**	**3·4 ± 0·3**	0·373

Figure 3. Recognition, preference and trust infographic.



Discussion

This non-inferiority validation study, conducted in a large academic hospital, is the first to prospectively compare LLM-generated summaries with physician-written summaries. The tool used for generating LLM-summaries is embedded in the EHR. Furthermore, this is the first study to validate LLM-generated medical summaries in a non-English clinical environment.

Our main findings show that LLM-generated medical summaries are non-inferior to physician-written medical summaries, particularly in terms of completeness and correctness. While the LLM-generated summaries were less concise they were trusted as much as the physician-written summaries. Notably, physicians were able to discern whether a summary was created by an LLM or a human in most cases.

These results suggest that integrating LLMs into EHR systems might be an effective strategy to reduce the administrative burden on clinicians without compromising the quality of patient care documentation. This study lays the groundwork for further research into optimizing LLM prompts and refining the technology. The results from this study allow us to test the functionality in real-world scenarios by providing evidence of its performance and reliability, enabling us to achieve even greater accuracy and efficiency through feedback loops. Future studies should explore the long-term impacts of LLM integration on clinical workflow and patient outcomes, as well as the adaptability of these models across different languages and healthcare settings.

Objective and subjective comparison of physician-written vs. LLM-generated summaries

The relatively low ROUGE and BLEU scores indicate that LLM-generated summaries and physician-written summaries differ significantly on word-level matching. However, as these scores primarily measure the overlap of words and phrases (n-grams) between two texts, they do not deeply consider the semantic meaning or importance of the words. Although widely used in NLTK analyses, they do not optimally account for synonyms, paraphrases, or different word forms that may convey the same meaning. Therefore, in our study physician evaluators were asked to review both the physician-written summary and the LLM-generated summary on completeness, correctness, and conciseness.

Our findings show that LLM-generated summaries are non-inferior to the physician-written summaries on completeness and correctness, the most clinically important parameters. Physicians wrote the summaries in a realistic timeframe, but with the knowledge that their summaries would be used as gold standard for the LLM-generated summaries. Therefore, the level of comparison was set considerably higher compared to earlier studies[3,4], where already existing data was used as reference. The LLM-generated summaries were significantly less concise. This was also reflected in the word count, with the LLM-summaries using on average 100 words per summary versus 60 for the physician-written summaries. However, as the average read time is approximately 238 words per minute, this results in 25 seconds reading time for the LLM-generated summaries versus 15 seconds for the physician-written summaries.[16] Although this finding is statistically significant, it is therefore unlikely that this will translate to any clinical relevancy. Therefore, conciseness seems less clinically important than completeness and correctness.

LLM-generated summaries were generated 42 times faster than physician written ones, suggesting significant potential for reducing clinical hours. Although not all clinicians typically produce written summaries in preparation for outpatient consultations, they are obliged to read up. Its impact may vary depending on individual clinical practices.

The majority of physician evaluators could identify the LLM-generated summaries. This suggests that the characteristics of LLM-generated summaries are distinct enough to be noticed by professionals, potentially influencing their trust, acceptance and reliance on these tools. Our results furthermore show a high level of trust for the LLM-generated summaries to be used in clinical practice. Their trustworthiness paralleled that of physician-written summaries, with no significant differences. The majority of physician evaluators either preferred the LMM-generated summaries or found them equivalent to physician-written summaries. These findings suggest that LLM-generated summaries not only hold up in terms of trust but are also viewed favorably by a significant proportion of clinicians with respect to usability and quality.

Comparison with Previous Studies
Only one other study has been published that evaluated LLM summarization on clinical notes[3]. Van Veen and colleagues recently performed a comprehensive study comparing the performance of different LLMs on various medical summarization tasks, one of which was summarizing clinical notes, and found an overall preference or equivalent score of 81% for summaries generated by the best LLM versus medical experts. They used readily available progress notes from the MIMIC-III dataset and used the problem list notes as the clinical summary. In contrast, we performed a prospective study and asked physicians to summarize medical files specifically for the purpose of our study as if they were preparing a patient visit. It is therefore likely that the physician-written summaries are of superior quality than those already available in the notes. Furthermore, this is the first study to evaluate LLM-generated medical summaries in a non-English clinical environment and the Dutch language. Note that the Dutch language is only spoken by less than 0.5% of the world population and the Dutch word count in the GPT-model is only 0.34%.[6] This study shows that foundation models can be used in a non-English clinical setting without further linguistic adaptations, underscoring the language agnostic characteristics of LLMs.

Limitations
This study has several limitations. First, physicians were used as a gold standard for evaluating the LLM-generated summaries. While physicians are highly skilled and knowledgeable, their summaries may inherently vary due to individual differences in experience, expertise, and subjective interpretation of clinical data. This variability can introduce inconsistencies in the benchmark against which AI-generated summaries are measured, potentially skewing the assessment of AI performance. Due to the low temperature settings used for the GPT-model, the LLM-generated summaries showed more consistency. Additionally, prompt engineering plays a crucial role in generating high-quality summaries. The effectiveness of LLM-generated summaries heavily depends on the design and specificity of the prompts given to the LLM.

Furthermore, medical summaries are inherently context specific, as, for example, an ENT surgeon would have a different focus and preference for the content of the summary than an orthopedic surgeon. In this study, this effect will be reflected in the physician-written summaries but not in the LLM-generated summaries. We aimed to find a prompt which serves every specialty, or only would need minor fine-tuning for end-users.

Implications for Clinical Practice
The findings of this study have significant implications for clinical practice. By demonstrating that LLM-generated summaries are non-inferior to physician-written summaries in terms of completeness and correctness, our study supports the integration of LLMs into EHR systems to reduce the administrative burden on clinicians. This can lead to more efficient use of clinical time, allowing healthcare providers to focus more on direct patient care. Furthermore, the potential of differentiated prompting—customizing LLM prompts to cater to specific medical specialties and disciplines such as nursing—can enhance the relevance and accuracy of the summaries. For instance, prompts tailored for an ENT surgeon can emphasize otolaryngological details, while those for a pediatrician can highlight developmental history. This customization ensures that the summaries meet the precise needs of different medical fields, improving the overall quality of patient documentation and care. As a result, integrating LLMs with differentiated prompting into clinical workflows can enhance the usability and effectiveness of EHR systems, ultimately leading to better patient outcomes and increased clinician satisfaction.

Future ambitions

Based on these results, further development should focus on prompt engineering to increase conciseness of the LLM-generated summaries without loss in quality (completeness and correctness), for example by reducing word count. The collected set of summaries can function as a validation set for future prompts. The scoring system used by the physician evaluators is quite laborious and future endeavors may focus on automating this process through LLMs and validating this approach.
This LLM was trained on English language and ingested and generated output in Dutch, even though there is a discrepancy between the amount of text available on the internet in English vs. in Dutch (estimated 55% vs. 1%)[14]. This could have influenced model performance; however, results were still non-inferior to the golden standard of physicians, suggesting even more room for improvement for the future.

In conclusion, the findings in our study indicate that LLM-generated medical summaries are a viable alternative to physician-written summaries. Additionally, our study shows that LLM-generated summaries are deemed trustworthy for clinical practice. This suggests that this functionality can be safely integrated and used in the EHR with significant potential for reducing administrative burdens and enhancing clinical efficiency. As LLM technology continues to evolve, its role in healthcare is likely to expand, offering new opportunities to improve patient care and streamline clinical workflows.

Acknowledgements
We thank the IT-team that supported this research and all physician participants, listed in the AAIH Consortium (Supplementary Information).

References

1 Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. NPJ Digit Med. 2024; **7**. DOI:10.1038/s41746-023-00988-4.

2 Yu P, Xu H, Hu X, Deng C. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Basel)* 2023; **11**. DOI:10.3390/HEALTHCARE11202776.

3 Van Veen D, Van Uden C, Blankemeier L, *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024; published online April 1. DOI:10.1038/s41591-024-02855-5.

4 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; **29**: 1930–40.

5 Li Z, Shi Y, Liu Z, Yang F, Liu N, Du M. Quantifying Multilingual Performance of Large Language Models Across Languages. 2024; published online April 17. http://arxiv.org/abs/2404.11553.

6 Tom B Brown. OpenAI - Dataset Language Statistics. Github. 2020. https://github.com/openai/gpt-3/tree/master/dataset_statistics (accessed May 17, 2024).

7 Rao SK, Kimball AB, Lehrhoff SR, *et al.* The Impact of Administrative Burden on Academic Physicians: Results of a Hospital-Wide Physician Survey. *Acad Med* 2017; **92**: 237–43.

8 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; **1**: 18.

9 Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review. *JMIR Med Inform* 2024; **12**: e52073.

10 The Lancet Regional Health – Europe. Embracing generative AI in health care. The Lancet Regional Health - Europe. 2023; **30**. DOI:10.1016/j.lanepe.2023.100677.

11 Duffourc M, Gerke S. Generative AI in Health Care and Liability Risks for Physicians and Safety Concerns for Patients. *JAMA* 2023; **330**: 313–4.

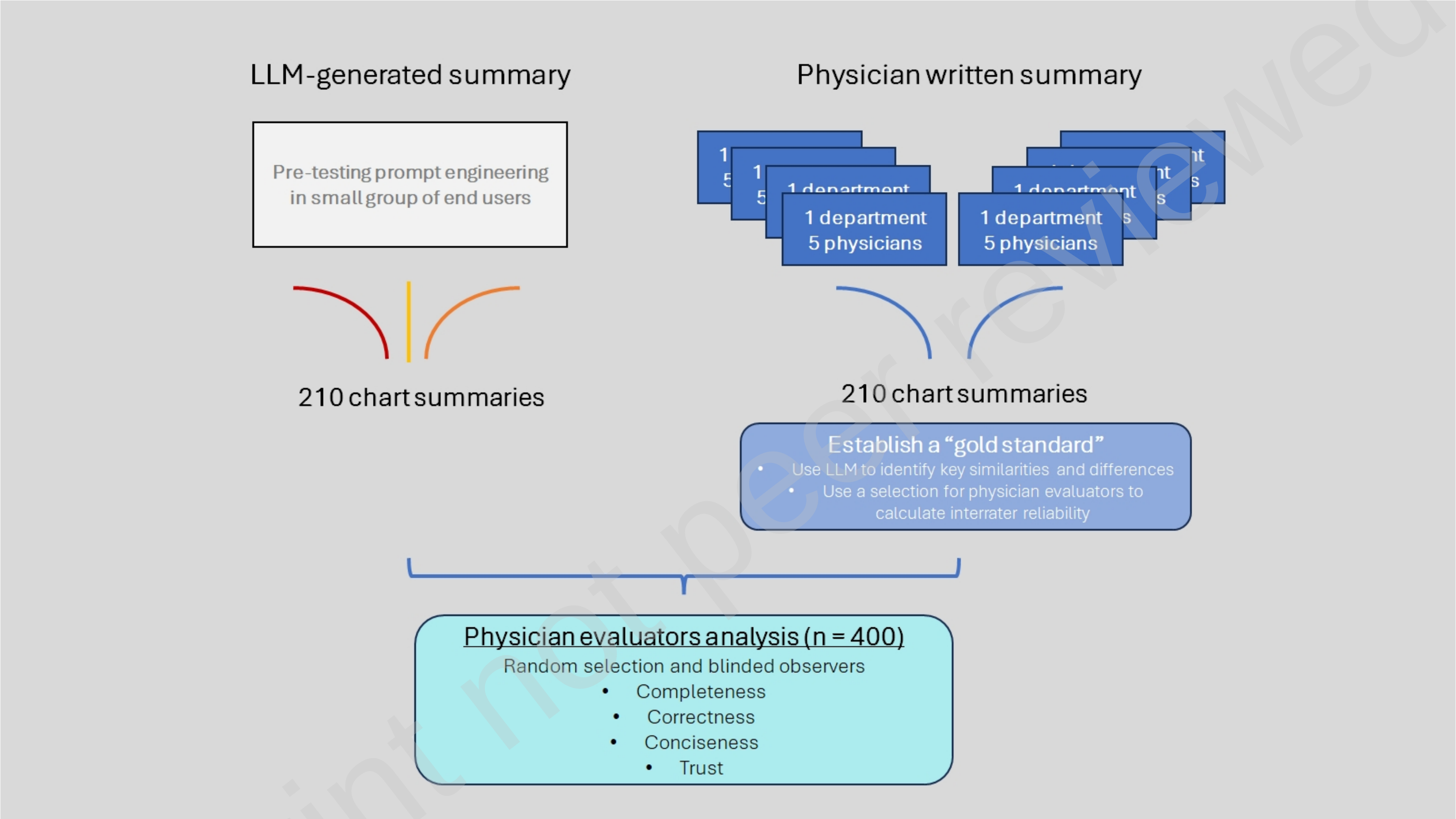
12 Meskó B. The Impact of Multimodal Large Language Models on Health Care's Future. *J Med Internet Res* 2023; **25**. DOI:10.2196/52865.

13 Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a Method for Automatic Evaluation of Machine Translation. .

14 Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. .

15 Pimienta D, Prado D, Blanco Disclaimer Á. Twelve years of measuring linguistic diversity in the Internet: balance and perspectives. 2009.

16 Brysbaert M. How many words do we read per minute? A review and meta-analysis of reading rate. *J Mem Lang* 2019; **109**: 104047.



LLM-generated summary

Pre-testing prompt engineering in small group of end users

210 chart summaries

Physician written summary

1 department
5 physicians

1 department
5 physicians

210 chart summaries

Establish a "gold standard"

- Use LLM to identify key similarities and differences
- Use a selection for physician evaluators to calculate interrater reliability

Physician evaluators analysis (n = 400)

Random selection and blinded observers

- Completeness
- Correctness
- Conciseness
- Trust

LLM-generated summary

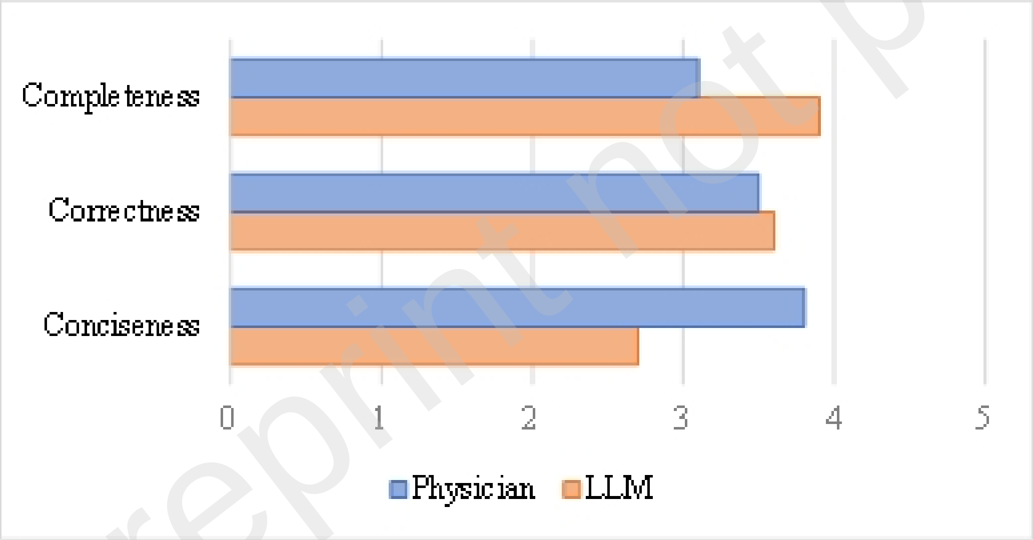
The patient has recurrent idiopathic subglottic stenosis, for which she has undergone multiple therapeutic microlaryngoscopies with balloon dilation. On [date], she underwent an uncomplicated therapeutic microlaryngoscopy, during which four incisions were made and a balloon dilation to a diameter of 15mm was performed. After the procedure, she experienced mild airway symptoms but showed improved airflow four weeks after the seventh dilation. The plan was to wait for the recovery from a possible viral cause of the symptoms. If the symptoms persisted, a check-up with flexible endoscopy would be performed to rule out other causes of reduced efficacy.

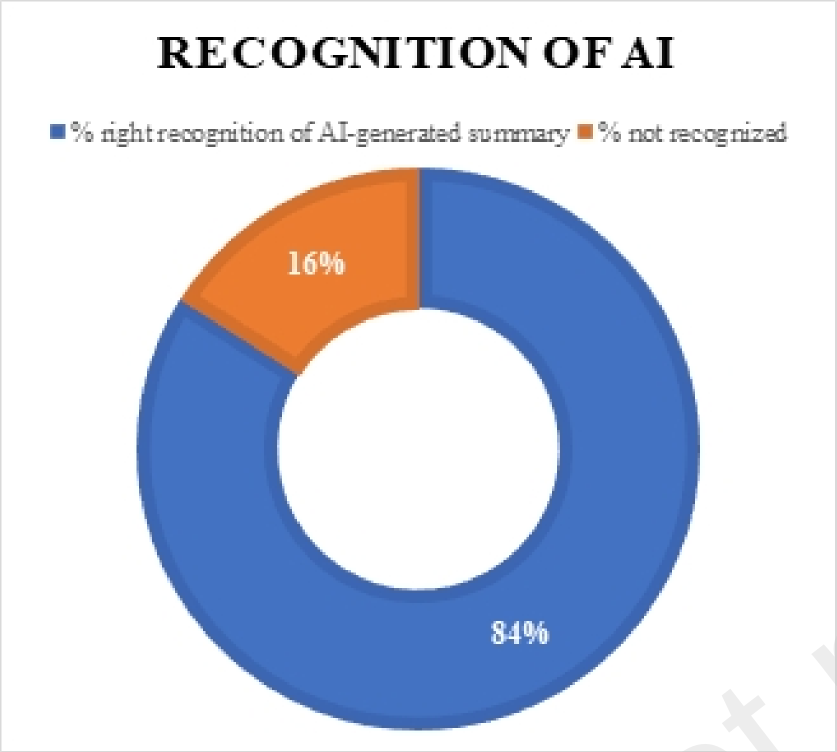
Physician-written summary I

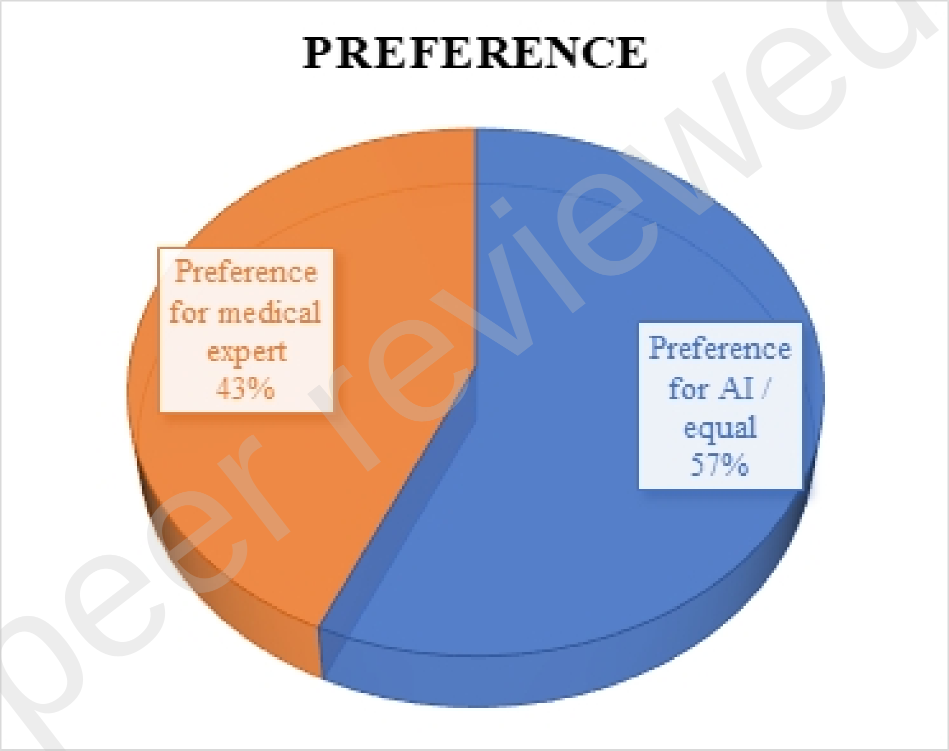
The patient has been known to have idiopathic subglottic stenosis since [month] [year], having undergone multiple (7 times) endoscopic treatments with steric incision and dilation between [month] [year] and [month] [year]. No other relevant medical history.

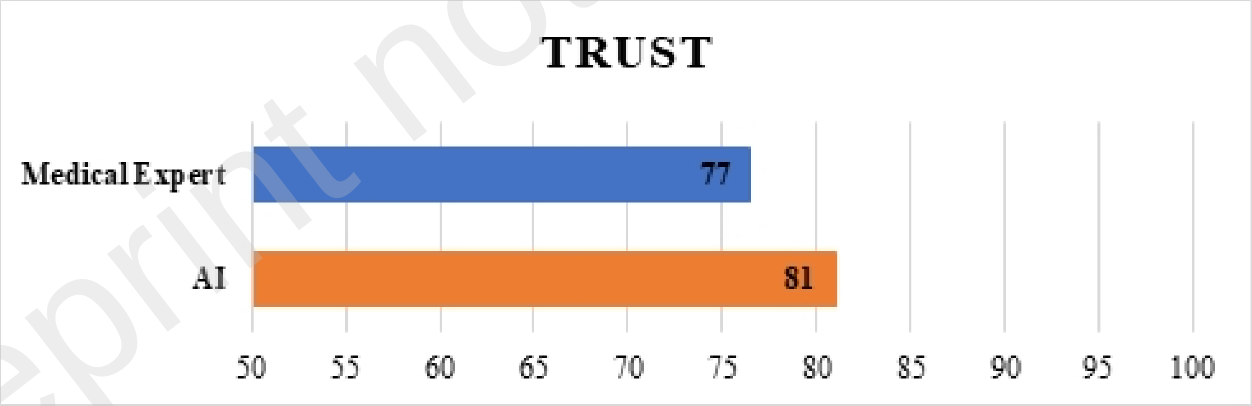
Physician-written summary II

Patient has been experiencing shortness of breath since [year] due to subglottic stenosis of unknown etiology, for which they have undergone more than 10 dilation procedures under anesthesia since that time, with the latest procedure in [month] [year]. At the last check-up, peak flow was 320 L/min.




RECOGNITION OF AI
■ % right recognition of AI-generated summary
■ % not recognized
16%
84%


PREFERENCE
Preference for medical expert 43%
Preference for AI / equal 57%


TRUST
Medical Expert
AI
77
81
50
55
60
65
70
75
80
85
90
95
100