scientific reports

Check for updates

OPEN

AI-generated neurology consultation summaries improve efficiency and reduce documentation burden in the emergency department

Alon Gorenshtein[1,2,3], Shay Perek[4,5], Yona Vaisbuch[5,6] & Shahar Shelly[1,3,7]✉

Physicians face a significant documentation burden, spending twice as much time on electronic health records (EHRs) as on direct patient care. Consultation summary reports from the emergency department (ED) are critical for continuity of care and clinical decision-making. This study aims to evaluate the quality and utility of automatically generated neurological consultation reports with clear recommendations, while reducing neurologists' documentation burden. We used neurological consultation reports (n = 250) from the ED as reference outputs. For each case, we fed the report's constituent components into the large language model (LLM). Using prompt engineering and retrieval-augmented generation (RAG) to generate auto-summarized reports, which were then compared against the original consultation reports. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and semantic embedding (Clinical-BioBert) were used as performance metrics. The LLM-generated report exhibited high semantic similarity with the neurologist's report (0.89 ± 0.03). However, significant differences in report length were observed, with LLM-generated reports being more concise than those written by attending neurologists (61.56 vs. 94.75 words, p < 0.001). Additionally, LLM-generated reports were written in a more straightforward and accessible style (FKGL = 11.3 vs. 12.22, p < 0.001). Despite these strengths, the LLM-generated reports exhibited substantial divergence in writing style from neurologists' reports (ROUGE-1 F1 = 0.25, ROUGE-2 F1 = 0.09, ROUGE-L F1 = 0.19). LLM-generated neurological consultation reports demonstrate strong semantic alignment with human-authored reports while offering a more concise and accessible format. Notable differences in writing style suggest a standardized approach that, while effective in conveying clinical content, may lack the personalization of neurologist-written reports.

Keywords Large language models, Neurology, Emergency department, Artificial intelligence

Neurology is a specialty which is highly susceptible to burnout among physicians[1]. The growing prevalence of chronic neurological diseases[2,3], shortage of neurologists[4] and lower salaries compared to other medical fields, increases the likelihood of burnout among neurologists. Consequently, a substantial proportion of neurologists worldwide report experiencing burnout, with prevalence rates ranging from 18.1% up to 94%[5]. One contributing factor to this burnout, which is not exclusive to neurologists, is the significant documentation burden[6], particularly pronounced in high-intensity settings such as emergency departments (ED)[7].

The role of neurologists in the ED is crucial for providing high-quality consultations on neurological cases thereby preventing misdiagnosis[8,9]. Further, a significant portion of the responsibility of the physicians is to document patient information for subsequent healthcare providers. Currently, this is accomplished through the manual writing of reports in the EHR system. However, this documentation process is known to be time-

[1]AI in Neurology Laboratory, Ruth and Bruce Rapaport Faculty of Medicine, Technion Institute of Technology , 3525408 Haifa, Israel. [2]Azrieli Faculty of Medicine , Bar-Ilan University , Safed, Israel. [3]Department of Neurology , Rambam Health Care Campus , Haifa, Israel. [4]Department of Emergency Medicine , Rambam Health Care Campus , Haifa, Israel. [5]Ruth and Bruce Rapaport Faculty of Medicine , Technion Institute of Technology , 3525408 Haifa, Israel. [6]Department of Otolaryngology - Head and Neck Surgery , Rambam Health Care Campus , Haifa, Israel. [7]Department of Neurology , Mayo Clinic , Rochester, MN, USA. ✉email: s_shelly@rmc.gov.il

consuming, with estimates indicating physicians devote twice as much time to EHR documentation as they do to direct patient care[10]. Such a task can be regarded as a low mental task, which doesn't necessary require the skills honed by the long training of a physicians. Nevertheless, accuracy in these records is paramount to avoid future medical errors. Documentation errors occur at alarming rates, ranging from 13% to 40%[11,12], usually due to physicians fatigue and cognitive biases[13,14]. Due to these factors a better solution than manually written notes is necessary, to reduce both physician work burden and medical errors.

A suitable solution could be to develop a tool that can assist neurologists by working as either first providing a draft followed by physician review, or as a tool which overlook the physician report. While both sound plausible, it's usually safer to allow the human to be the last judgement and not artificial intelligence (AI). The common framework for documentation and language tasks typically centers around large language models (LLMs)[15]. While a range of tools that utilize LLMs has been extensively explored in literature for automating medical report generation[15], most studies focus on broad topics and fail to address the nuanced and complex needs of the neurology field. This is especially true in high-intensity emergency room consultations. This study aims to investigate whether LLMs can generate consultation reports in the emergency room that not only summarize patient information, but also offer tailored recommendations to guide neurologists in determining the most appropriate next steps for patient management.

Materials and methods

Standard protocol approvals, registrations, and patient consents

The study was conducted with institutional research board (IRB) approval. Due to the retrospective nature of the study, Rambam healthcare campus IRB waived the need of obtaining informed consent. All methods were carried out in accordance with relevant guidelines and regulations.

Cohort identification

This retrospective study comprised 250 consecutive cases from the ED at Rambam Healthcare Campus. Clinical information was uniformly extracted using an electronic record retrieval system capable of accessing all clinical and laboratory results. We identified all patients who underwent neurological consultation in the ED from 01/01/2024 to 29/02/2024, with follow-up concluding on 16/08/2024. Inclusion criteria included patients above 18 years old with a medical history. Exclusion criteria included lack of complete consultation history, lack of follow-up data until 16/08/2024, and erroneous ICD-9 code at discharge in the electronic records. All consultation reports were manually translated into English from Hebrew and subsequently reviewed by a professional translator to ensure accuracy. This was done to facilitate an evaluation between AI generated report and original consultation report.

LLM implementation

The framework is based on an LLM, Gemini 1.5-pro API securely hosted within the Vertex AI platform, provided by Google Cloud services. In addition, our engagement with Gemini API is underpinned by stringent data management agreements with Google. These agreements guarantee that patient data are strictly confined to the intended research objectives and that no training of the Gemini model can occur, thereby maintaining confidentiality and integrity throughout the study. The LLM was set temperature = 0 to avoid hallucination focusing only on the input data. The LLM inputs are neurological examination, patient medical history, radiological findings, and laboratory results extracted from EHRs (Fig. 1). The output was the consultation report with recommended next step (admission vs. discharge, refer to different consultation, return to emergency department physician). All model inputs were restricted to information time-stamped on or before the moment the neurologist opened the consult note. These inputs comprised (i) demographic details and chief-complaint history recorded by the triage nurse, (ii) an unformatted "Initial Neuro Exam" scratch pad typed by the neurologist immediately after bedside assessment, and (iii) laboratory and radiology results that had been done in the EHR after the neurological examination. The final structured consult note written only after the neurologist had reviewed all subsequent results was withheld from the model to avoid circularity. To enhance the relevance and accuracy of the LLM's output, we employed a retrieval-augmented generation (RAG) technique, presenting five analogous cases that illustrate both the input parameters and the resultant neurological consultation reports.The RAG is based on a a hybrid similarity search that was built on full historical consult records. For every encounter we concatenated: (1) demographic and triage data, (2) nurse-recorded history and neurological examination, (3) laboratory and imaging results, and (4) the neurologist's final free-text note. Each composite string was embedded with BioClinicalBERT (BioBERT) and stored in a FAISS IndexFlatL2. At runtime the current case inputs was embedded with the same model; the five nearest neighbours (highest cosine similarity, patient-ID excluded) were retrieved and appended. This approach aims to mitigate common limitations observed in LLMs, such as the lack of empathy, contextual relevance, and tendencies toward verbosity or excessive informality[16].

Prompt disclosure and reproducibility

The exact system prompt supplied to the Gemini 1.5-pro model is reproduced in Supplementary File 1. It includes role definition, output schema, length constraints, and an instruction to refuse if insufficient data are provided. No patient-specific identifiers were used.

Performance metrics

To rigorously assess the quality and clinical applicability of LLM-generated neurology consult summaries, we employed a multi-faceted evaluation framework incorporating semantic similarity, readability indices. The primary objective was to ensure that AI-generated summaries preserved critical neurological details while enhancing efficiency and reducing documentation burden. Cosine similarity, calculated using Clinical-BioBERT

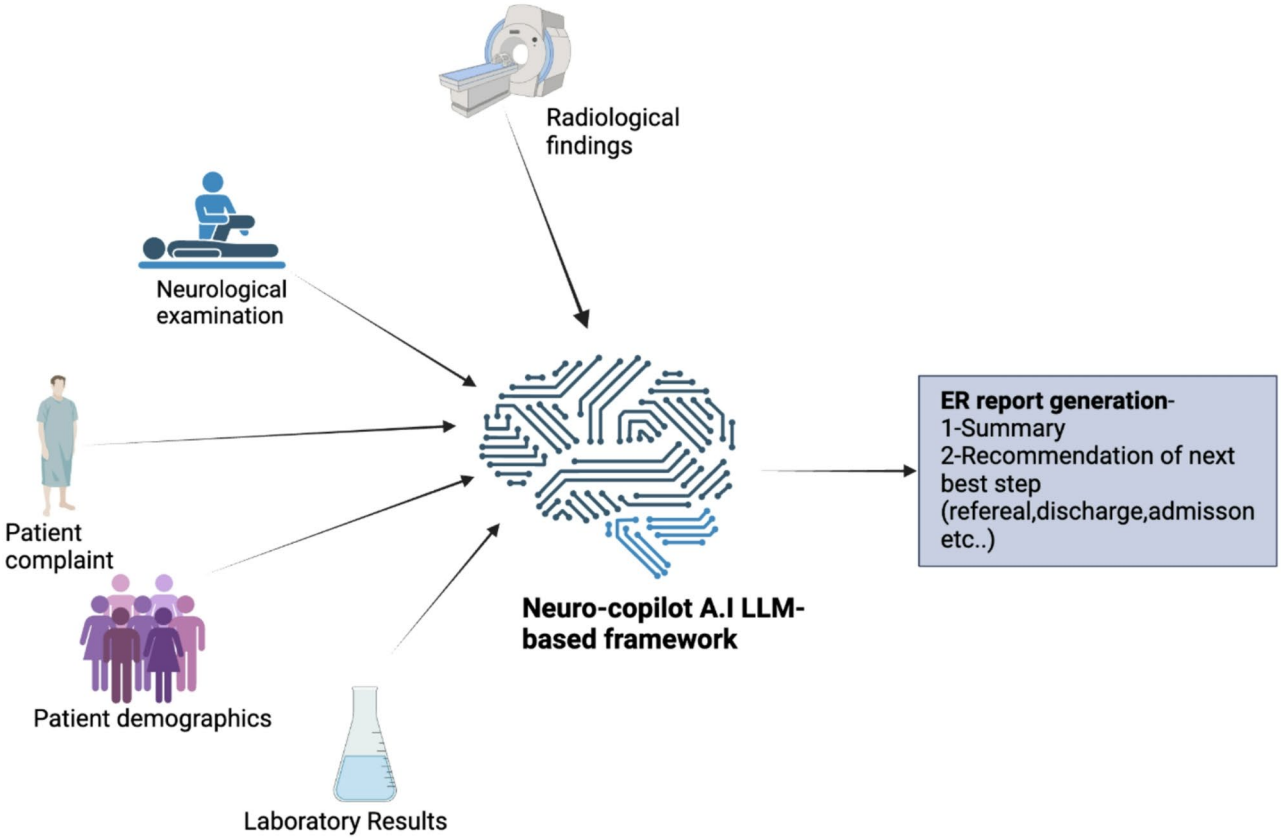

Fig. 1. Inputs and outputs of the LLM. The model receive patient anamnesis (medical history), findings from the neurological examination, patient demographic (age and gender), radiological findings, laboratory findings.

embeddings, provided a quantitative measure of semantic alignment between LLM-generated and physician-authored summaries, ensuring that generated texts retained meaningful medical context beyond superficial word overlap. Additionally, ROUGE scores (ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1) assessed lexical similarity, capturing both unigram and bigram coherence as well as syntactic structure. Readability was evaluated using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES), ensuring that summaries remained accessible for clinicians while maintaining necessary medical precision.

Capturing biases for similarity differences

For potential biases we examined the hospitalization status and temporal trends, both of which were external to the input data for the language model. Consequently, the model is unaware of these factors, even in the context of the physician's report. The reason for these bias speculations stems from the nature of hospitalized case are assumed to be more complex and nuanced suggesting that the LLM might struggle to provide high quality consultation report. For time analysis was due the fact that it may reflect human factors such as physician fatigue during night shifts (23:00–06:00) or increased patient load during peak hours (09:00–11:00), which could compromise report quality. Report timing was determined using EHR timestamps (when physicians finalized reports).

Statistical analysis

For comparisons between groups, qualitative variables were analyzed using Fisher's exact test and chi-square test. Continuous variables that followed a parametric distribution were analyzed by Student's t-test, and nonparametric variables were analyzed by the Mann–Whitney U test. The threshold for significance was set at $p < 0.05$.

Ethics approval. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Consent to participate. Approval for this study was obtained by the Institutional Review Board at the Rambam health care campus. All methods were carried out in accordance with relevant guidelines and regulations.

Results

Cohort

We identified 1,368 consecutive cases of patients who underwent neurological consults in the emergency department (ED). From this group, 250 consultation reports were selected for comparison with the AI-generated reports. The rest of the consultation reports retierved from the EHR ($n = 1118$) lacked input parts of the diagnostic components relevant to neurological consultations, including detailed patient histories, neurological examinations, radiographic findings, and laboratory results. If we Include incomplete input into the LLM this can result in inaccurate comparisons with the human-written notes, as the neurologist had access to the missing information during the consulations. Therefore, we exclude incomplete reports to ensure accuracy in our analyses.

The most prevalent neurological conditions observed were stroke and cerebrovascular diseases ($n = 35$, 14%), headache disorders ($n = 32$, 12.8%), and seizure disorders ($n = 28$, 11.2%). A total of 86 patients (34.4%) were hospitalized, with 49 (19.6%) admitted to the neurology department. Among the total, 232 patients (92.8%) had blood lab results, 182 (72.8%) underwent computed tomography (CT) scans, 148 (59.2%) had electrocardiograms (ECG), and only 12 patients (4.8%) had lumbar punctures performed (Table 1.)

AI generated report similarity performance

Cosine similarity (clinical-BioBERT)

To assess the semantic similarity between AI-generated and true summaries, we employed Clinical-BioBERT embeddings. The mean cosine similarity score was 0.89 ± 0.03. These findings indicate a high degree of semantic alignment, suggesting that the AI-generated summaries preserved the core clinical meaning of physician-authored reports. This strong semantic similarity demonstrates the model's effectiveness in capturing essential medical information without considering if the wording and phrasing differ significantly. (also see supplementray Fig. 1)

ROUGE scores

While cosine similarity confirmed the semantic alignment, ROUGE F1 evaluation provided insights into textual overlap. The mean ROUGE-1 F1 score was 0.28, indicating limited unigram-level similarity, while ROUGE-2 F1 and ROUGE-L F1 scores were 0.09 and 0.19, respectively. These results suggest that although the generated summaries contained key clinical terms, their phrasing and structure varied significantly from physician-authored reports.

Hospitalization-based differences

The mean cosine similarity score for hospitalized patients was 0.89, compared to 0.88 for non-hospitalized patients ($p = 0.45$), showing no statistically significant difference between these groups. Testing this bias was essential to ensure the tool's reliability across diverse clinical scenarios (Fig. 2.).

Features	Total ($n = 250$)
Age (median, IQR)	56 [35.17–74.17]
Male (n, %)	120 (48%)
Hospitalized	86 (34.4%)
Hospitalized at neurology department	49 (19.6%)
Mortality	23 (9.2%)
Neurological category of consult based on ICD-9 code on release	
Stroke and Cerebrovascular disorders	35 (14%)
Seizure Disorders	28 (11.2%)
Headache disorders	32 (12.8%)
Neuromuscular Disorders	12 (4.8%)
Central demyelinating disorders	1 (0.4%)
Infections of the nervous system disorders	2 (0.8%)
Other disorder of CNS	8 (3.2%)
ICD-9 code without diseases of the nervous system and sense organs at release from ED	132 (52.8%)
Imaging Conducted at ED	
C.T	182 (72.8%)
MRI	9 (3.6%)
ECG	148 (59.2%)
Laboratory data	
Blood test	232 (92.8%)
Lumbar Puncture	36 (14.4%)
Urine test	12 (4.8%)

Table 1. Baseline demographic and clinical characteristics of patient neurological consultation reports.

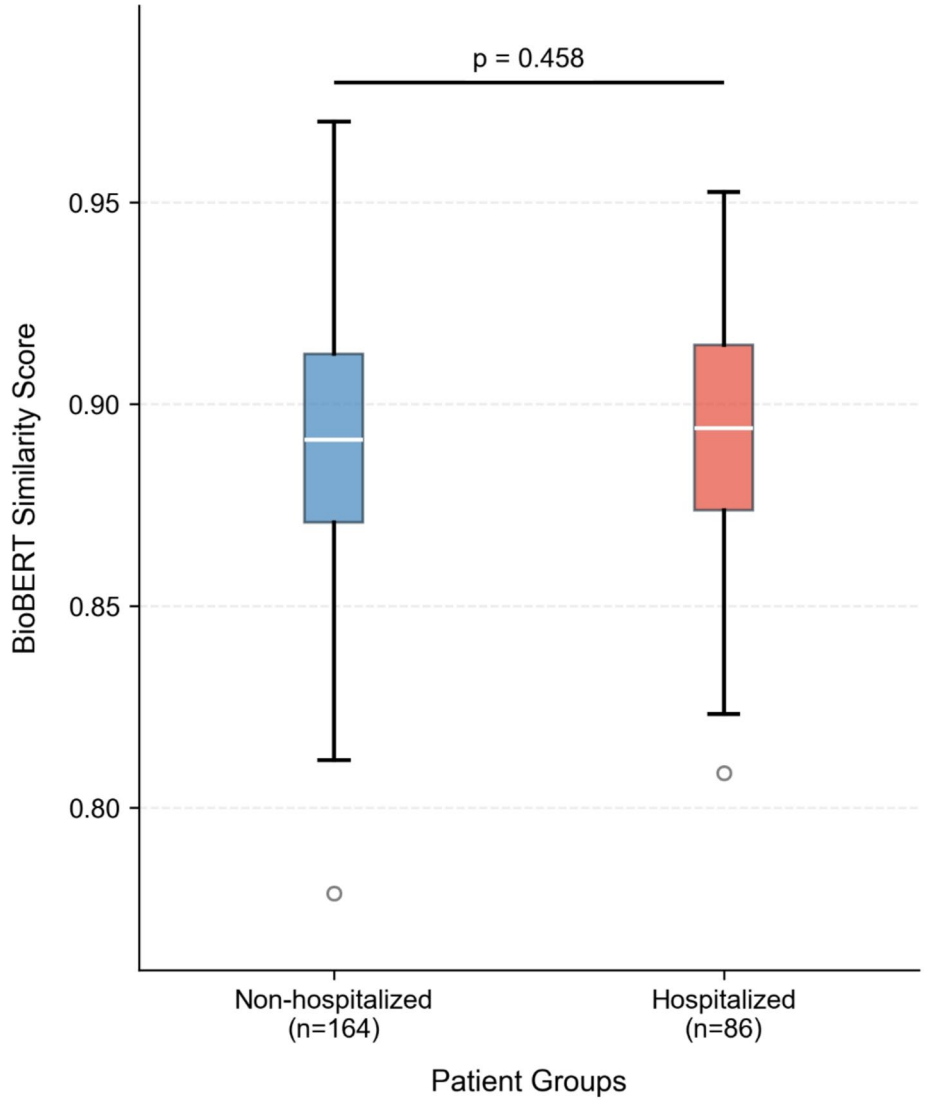

Fig. 2. Comparison of BioBERT-Derived Similarity Scores Between Non-Hospitalized and Hospitalized Patients: Box-and-whisker plots illustrating BioBERT similarity scores for non-hospitalized ($n = 164$, blue) and hospitalized ($n = 86$, red) patients. The central horizontal line within each box denotes the group median, with the box boundaries representing the interquartile range. Whiskers extend to the lowest and highest values excluding outliers, which are plotted individually. Statistical comparison revealed no significant difference in scores between the two groups ($p = 0.458$).

Similarity analysis

To benchmark the 0.88 AI-to-reference overlap, we randomly sampled 1 000 unique, unordered pairs within each cohort and computed BioClinicalBERT cosine similarity. Human-to-human pairs showed a median similarity of 0.98 (IQR 0.97–0.98), whereas AI-to-AI pairs were 0.97 (0.97–0.98); the difference, while statistically significant ($p < 0.001$), is numerically trivial ($\Delta = 0.01$). Further details can be seen on supplementary figure S1.

Hourly trends

Hourly analysis over a 24-hour period (Fig. 3) showed that both attending-summary (human authored) lengths and BioBERT-based similarity scores (between AI generated report to human authored report) fluctuated in ways that did not cleanly align throughout the day. Notably, both metrics rose to their highest levels in the mid-morning (around 09:00–10:00) before dipping sharply at around 11:00. Late-evening hours (e.g., 23:00) also showed relatively lower similarity scores alongside shorter summaries, suggesting that certain time blocks whether due to shift fatigue, varying patient loads, or other contextual factors may influence documentation patterns. These findings raise the possibility of temporal biases in clinical summary quality, underscoring the need for further investigation to clarify the roles of shift schedules, circadian rhythms, and systemic factors in shaping how neurology attendings generate their documentation.

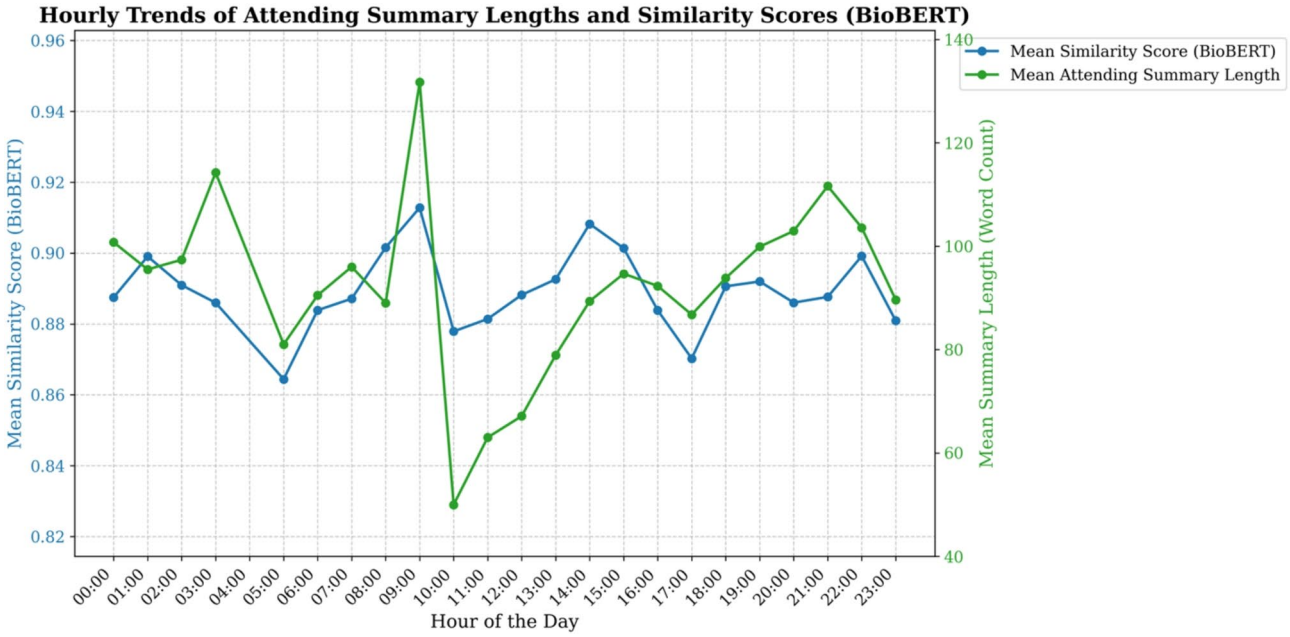

Fig. 3. Hourly Trends in Attending Summary Length and BioBERT-Derived Similarity Scores: A dual-axis line chart depicting the mean summary length of human authored reports (in word count; green line, right y-axis) and mean BioBERT similarity score (blue line, left y-axis) at each hour of the day (x-axis). Data range from 00:00 to 23:00, illustrating temporal fluctuations in both content length and language-based similarity scores of attending summaries over a 24-hour period.

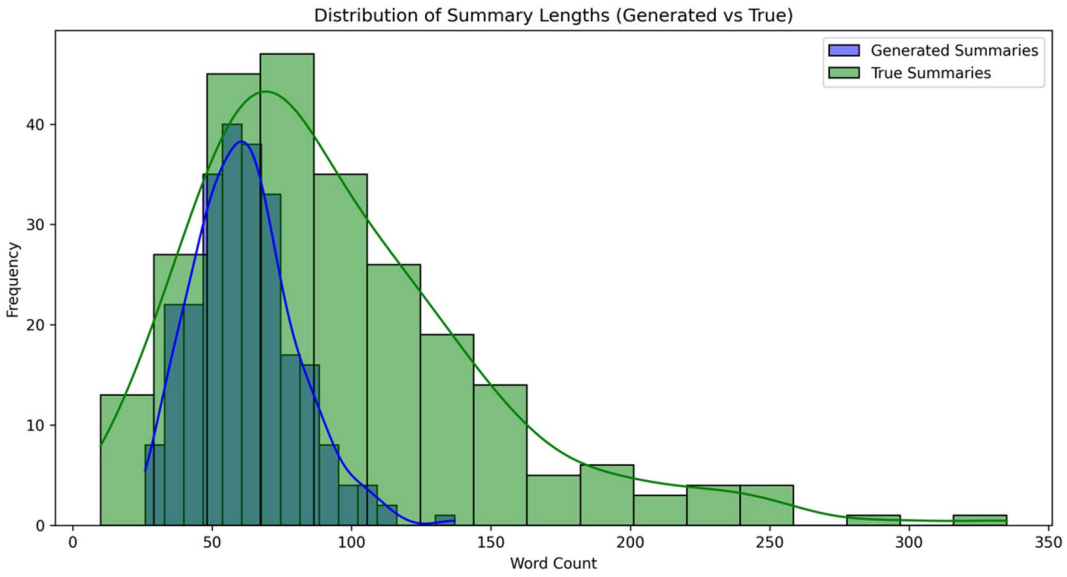

Fig. 4. Distribution of Summary Lengths: Generated vs. True Summaries: Overlaid histograms illustrating the word count distributions for AI-generated summaries (purple bars and curve) and actual (human-authored) summaries (green bars and curve). The x-axis denotes the number of words in each summary, while the y-axis represents the frequency of summaries in each bin. Overall, human-authored summaries extend to higher word counts, whereas AI-generated summaries tend to cluster at shorter lengths.

Summary length

In comparing the word counts of AI-generated summaries to their clinician-written counterparts, a clear difference in brevity emerged (Fig. 4). The AI-generated summaries displayed a pronounced left-shift in their distribution, with a mean of 61.57 words versus 94.75 words for the true summaries. This difference was significant ($p < 0.001$). Notably, the high similarity scores from BioBERT suggest that these concise summaries

effectively preserve the essential clinical information. This indicates that the model can maintain brevity without sacrificing critical content, underscoring its suitability for fast-paced clinical workflows.

Readability performance

The mean FKGL for the generated summaries was 11.30, compared to 12.22 for the true summaries. There was statistically significant difference between the report ($p < 0.001$), indicating that the generated summaries are written at a lower grade level and are therefore easier to comprehend. In contrast, the FRES analysis showed no significant difference between the generated and true summaries (p-value = 0.85). This suggests that both sets of summaries are comparable in terms of readability ease, but the generated summaries may require slightly less advanced literacy for comprehension. The balance of improved readability and retained clinical content underscores the potential usability of the AI-generated summaries in high-stress clinical environments.

Next step recommendations

To further provide for neurologists we evaluated the LLM regarding the next step after the consultation. With the model being correct 78.8%. Correctness was defined as agreement between the model's recommendation (admit vs. discharge) and the actual patient outcome. When the model erred, it was equally likely to recommend admission for patients who were ultimately discharged (18.9%) as to recommend discharge for patients who were ultimately admitted (34.1%, $p = 0.32$). In certain reports, the model suggested referrals to specialists, and these recommendations were consistent with actual outcomes, indicating the same specialist. Notably, these reports included patients with extensive medical histories specific to this specialization (e.g., oncology).

Discussion

We evaluated an LLM tailored for the ED, designed to function as AI-generated neurologic consultation reports to reduce documentation burnout for neurologist consultants. The LLM demonstrates strong performance in semantic similarity with mean 0.89 cosine similarity score. The model demonstrated strong performance in capturing clinically relevant information, achieving a high semantic similarity score (mean cosine similarity = 0.89). Notably, the accuracy of AI-generated reports remained consistent across different contexts, including night shifts and reports for hospitalized patients, suggesting robustness against contextual biases. Additionally, the LLM-generated reports were written in a more accessible style, potentially improving comprehension for both patients[17] and downstream care providers. The reduced length of the AI-generated report, when juxtaposed with its human-authored counterpart, maintains a comparable clinical relevance, as evidenced by a high clinical similarity score.We speculate that this has the potential to alleviate cognitive burden by shortening the time spent on EHRs, a frequent contributor to cognitive fatigue[18,19]. However, further research is necessary to evaluate the medicolegal implications and billing processes associated with AI-generated reports in comparison to those created by human providers. This is particularly crucial in the U.S., where physicians typically spend more time on EHRs compared to their counterparts in other countries[20]. While prior studies have demonstrated the ability of LLMs to generate accurate medical reports, it is important to note that only 33% of reports generated by GPT-4 were entirely free of errors, highlighting the need for continued validation and refinement of AI-assisted documentation tools[21]. Most reports contained hallucinations and omitted clinically relevant information[22]. This finding is pertinent to our study, where, under optimal conditions, the expected similarity score should exceed 0.89, approaching a score of 1. This indicates that either the LLMs are missing vital clinical information or that the physicians are neglecting it; however, the latter is less likely, given that only complete and comprehensive physician reports were included to the analysis.

This AI tool possess the potential of identifying, notifying and filling the missing crucial components of the human written consultation report which seems to be a prevalent need. It is important to recognize that, despite the extensive studies on generative reports using AI, such research is rarely conducted in the field of neurology[23]. Most studies tend to focus primarily on diagnostic applications. AI should not be limited to diagnostics; it should be integrated throughout the field to enhance patient care and improve the quality of life for neurologists as well. It is crucial to recognize that the principal objective of AI generative reports is to streamline documentation processes by providing physicians with a structured template. This approach alleviates the necessity for clinicians to draft notes from scratch, thereby mitigating cognitive load and work-related stress[24,25]. Despite initial promises of reducing the time physicians spend on documentation as shown in theoretical study[26], a subsequent quality improvement study assessing AI-generated draft replies to patient messages found no significant reduction in the time required to compose responses[27]. The study identified key challenges with AI-generated drafts, including a lack of empathy and personalization essential for patient-centered communication. Physicians also frequently criticized the drafts for being excessively long. While these issues raise concerns about the practicality of AI-generated reports, they highlight opportunities to refine LLMs for better alignment with clinical needs. By leveraging prompt engineering and RAG, we successfully optimized report length and relevance, making the AI-generated summaries more concise and clinically useful.

A secondary outcome was to evaluate the alignment of recommendations, revealing a 78.8% concordance with neurologist decision. This closely mirrors findings from previous research, which reported a 77.5% accuracy rate for ChatGPT-4[28]. These results suggest that LLMs exhibit difficulties in producing reliable prediction probabilities[29]. Traditional machine learning and deep learning architectures, as evidenced by various studies, may be more adept at prediction tasks[30,31]. This notion implies a potential for hybridizing these two architectures to leverage the predictive accuracy of machine learning techniques alongside the language processing capabilities of LLMs. Such a hybrid approach could enhance tasks such as consultation reports, which require both summarization and the provision of specific patient management pathways, including admission, discharge, or referrals to other specialists.

Our study presents several limitations worth noting. Firstly, we operated with a relatively small dataset, which can introduce significant variability, particularly affecting specific subgroups within the sample. This often results in skewed outcome estimations. Notably, our cohort exhibited a high prevalence of non-neurological cases, which could further complicate interpretations. Moreover, our methodological approach lacked a systematic manual review of the reports, as we did not implement a formal grading system. Our evaluation was solely based on automated text-overlap metrics (ROUGE-1/2/L), which, while providing a convenient and reproducible benchmark, fail to adequately assess the clinical utility of a consult note or confabulations of certain reccmondetions and diagnosis. That is, these metrics do not evaluate key factors such as clarity, accuracy, and the note's ability to guide the primary care team's management strategy. Incorporating expert evaluation into analysis is thus a critical next step, highlighting a significant limitation of our current study. The retrospective nature of the study raises additional concerns regarding the compliance of neurologists in utilizing the AI tool within high-pressure environments, such as the emergency room. There remains a gap in understanding how physicians interact with AI tools in hospital settings and the extent to which patients adhere to recommendations based on AI-generated assessments of medical information.

We encountered a key limitation stemming from our strict inclusion criteria for data completeness. Of the 1,368 emergency department encounters reviewed, only 246 (18%) had sufficient documentation to allow for automated summarization. This required the presence of triage demographics, neurological examinations, provisional ICD-9 codes, and at least one finalized laboratory or imaging report. The exclusion of the remaining cases primarily reflects challenges in retrieving structured data from the electronic health record. We chose to exclude incomplete cases to ensure a fair comparison between the model-generated summaries and comprehensive human-written notes, as missing inputs would inherently bias the evaluation in favor of the human reports. As a result, our model is currently applicable only when complete data are available—a limitation that highlights a broader issue in real-world settings, where incomplete documentation is unfortunately common.

In conclusion, augmented medical report generation can support ER neurologists by generating preliminary report drafts, reducing documentation time, and enabling clinicians to focus more on direct patient care and personalized communication. By streamlining documentation, these tools have the potential to enhance both physician efficiency and the overall patient experience. Future research should prioritize real-world implementation and evaluate how AI-driven reporting impacts clinical decision-making, workflow, and patient outcomes.

Ethics declarations

The authors declare no conflict of interest. All methods were carried out in accordance with relevant guidelines and regulations.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Received: 11 March 2025; Accepted: 30 September 2025
Published online: 06 November 2025

References

1. Burnout in Practicing Neurologists | Neurology Clinical Practice. (accessed 1 February 2025). https://www.neurology.org/doi/full/10.1212/CPJ.0000000000200422
2. Alzheimer's Association. 2015 Alzheimer's disease facts and figures. *Alzheimers Dement J Alzheimers Assoc* **11**(3), 332–384. https://doi.org/10.1016/j.jalz.2015.02.003 (2015).
3. Kowal, S. L., Dall, T. M., Chakrabarti, R., Storm, M. V. & Jain, A. The current and projected economic burden of parkinson's disease in the united States. *Mov. Disord Off J. Mov. Disord Soc.* **28** (3), 311–318. https://doi.org/10.1002/mds.25292 (2013).
4. Majersik, J. J. et al. A shortage of Neurologists – We must act now. *Neurology* **96** (24), 1122–1134. https://doi.org/10.1212/WNL.0000000000012111 (2021).
5. Patel, U. K. et al. Recommended strategies for physician Burnout, a Well-Recognized escalating global crisis among neurologists. *J. Clin. Neurol.* **16** (2), 191–201. https://doi.org/10.3988/jcn.2020.16.2.191 (2020).
6. Gaffney, A. et al. Medical Documentation burden among US Office-Based physicians in 2019: A National study. *JAMA Intern. Med.* **182** (5), 564–566. https://doi.org/10.1001/jamainternmed.2022.0372 (2022).
7. Moy, A. J. et al. Understanding the perceived role of electronic health records and workflow fragmentation on clinician Documentation burden in emergency departments. *J. Am. Med. Inf. Assoc. JAMIA*. **30** (5), 797–808. https://doi.org/10.1093/jamia/ocad038 (2023).
8. Moulin, T. et al. Impact of emergency room neurologists on patient management and outcome. *Eur. Neurol.* **50** (4), 207–214. https://doi.org/10.1159/000073861 (2003).
9. Caplan, L. R. Dizziness: how do patients describe Dizziness and how do emergency physicians use these descriptions for diagnosis? *Mayo Clin. Proc.* **82** (11), 1313–1315. https://doi.org/10.4065/82.11.1313 (2007).
10. Moy, A. J. et al. Measurement of clinical Documentation burden among physicians and nurses using electronic health records: a scoping review. *J. Am. Med. Inf. Assoc. JAMIA*. **28** (5), 998–1008. https://doi.org/10.1093/jamia/ocaa325 (2021).
11. Review article: Components of a good quality discharge summary: A systematic review - Wimsett – 2014 - Emergency Medicine Australasia - Wiley Online Library. (accessed 1 February 2025). https://doi.org/10.1111/1742-6723.12285
12. Schwarz, C. M. et al. A systematic literature review and narrative synthesis on the risks of medical discharge letters for patients' safety. *BMC Health Serv. Res.* **19**, 158. https://doi.org/10.1186/s12913-019-3989-1 (2019).
13. West, C. P., Tan, A. D., Habermann, T. M., Sloan, J. A. & Shanafelt, T. D. Association of resident fatigue and distress with perceived medical errors. *JAMA* **302** (12), 1294–1300. https://doi.org/10.1001/jama.2009.1389 (2009).
14. Tawfik, D. S. et al. Physician Burnout, Well-being, and work unit safety grades in relationship to reported medical errors. *Mayo Clin. Proc.* **93** (11), 1571–1580. https://doi.org/10.1016/j.mayocp.2018.05.014 (2018).
15. Busch, F. et al. Current applications and challenges in large Language models for patient care: a systematic review. *Commun. Med.* **5** (1), 1–13. https://doi.org/10.1038/s43856-024-00717-2 (2025).

16. Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large Language models in medicine: the potentials and pitfalls: A narrative review. *Ann. Intern. Med.* **177** (2), 210–220. https://doi.org/10.7326/M23-2772 (2024).
17. Safeer, R. S. & Keenan, J. Health literacy: the gap between physicians and patients. *Am. Fam Physician*. **72** (3), 463–468 (2005).
18. Khairat, S. et al. Association of electronic health record use with physician fatigue and efficiency. *JAMA Netw. Open.* **3** (6), e207385. https://doi.org/10.1001/jamanetworkopen.2020.7385 (2020).
19. Asgari, E. et al. Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review. *JMIR Med. Inf.* **12** (1), e55499. https://doi.org/10.2196/55499 (2024).
20. Holmgren, A. J. et al. Assessment of electronic health record use between US and Non-US health systems. *JAMA Intern. Med.* **181** (2), 251–259. https://doi.org/10.1001/jamainternmed.2020.7071 (2021).
21. Williams, C. Y. K. et al. Evaluating large Language models for drafting emergency department discharge Summaries. Published online April 4, 2024:2024.04.03.24305088. https://doi.org/10.1101/2024.04.03.24305088
22. Woolf, S. H., Kuzel, A. J., Dovey, S. M. & Phillips, R. L. A string of mistakes: the importance of cascade analysis in describing, counting, and preventing medical errors. *Ann. Fam Med.* **2** (4), 317–326. https://doi.org/10.1370/afm.126 (2004).
23. Gutman, B., Shmilovitch, A. H., Aran, D. & Shelly, S. Twenty-Five years of AI in neurology: the journey of predictive medicine and biological breakthroughs. *JMIR Neurotechnology*. **3** (1), e59556. https://doi.org/10.2196/59556 (2024).
24. Shah, S. J. et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and Documentation burden. *J. Am. Med. Inf. Assoc.* **32** (2), 375–380. https://doi.org/10.1093/jamia/ocae295 (2025).
25. Gandhi, T. K. et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? *JAMIA Open.* **6** (3), ooad079. https://doi.org/10.1093/jamiaopen/ooad079 (2023).
26. Menzies, D., Kirwan, S. & Albarqawi, A. AI managed emergency Documentation with a pretrained model. *Published Online August*. **17**. https://doi.org/10.48550/arXiv.2408.09193 (2024).
27. Tai-Seale, M. et al. AI-Generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open.* **7** (4), e246565. https://doi.org/10.1001/jamanetworkopen.2024.6565 (2024).
28. Glicksberg, B. S. et al. Evaluating the accuracy of a state-of-the-art large Language model for prediction of admissions from the emergency room. *J. Am. Med. Inf. Assoc.* **31** (9), 1921–1928. https://doi.org/10.1093/jamia/ocae103 (2024).
29. Gu, B., Desai, R. J., Lin, K. J. & Yang, J. Probabilistic medical predictions of large Language models. *Npj Digit. Med.* **7** (1), 1–9. https://doi.org/10.1038/s41746-024-01366-4 (2024).
30. Not the Models You Are Looking For: Traditional ML Outperforms LLMs in Clinical Prediction Tasks - PubMed. (accessed 5 February 2025). https://pubmed.ncbi.nlm.nih.gov/39677419/
31. Ghaffarzadeh-Esfahani, M. et al. Large Language models versus classical machine learning: performance in COVID-19 mortality prediction using High-Dimensional tabular data. *Published Online September*. **2**. https://doi.org/10.48550/arXiv.2409.02136 (2024).

Author contributions

AG- study concept and design, acquisition of data, analysis, and interpretation, drafting the manuscript, critical revision of the manuscript. SS- study concept and design, acquisition of data, analysis, and interpretation, drafting the manuscript, critical revision of the manuscript, study supervision. SP-drafting the manuscript, critical revision of the manuscript. YV-drafting the manuscript, critical revision of the manuscript. All authors have read and approved the final manuscript and consent to its publication.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-22769-7.

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.