npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02005-2

Evaluating clinical AI summaries with large language models as judges

Emma Croxford[1], Yanjun Gao[2], Elliot First[3], Nicholas Pellegrino[3], Miranda Schnier[3], John Caskey[4], Madeline Oguss[4], Graham Wills[5], Guanhua Chen[1], Dmitriy Dligach[4,6], Matthew M. Churpek[1,4,5], Anoop Mayampurath[1,4], Frank Liao[5,7], Cherodeep Goswami[5], Karen K. Wong[3,5], Brian W. Patterson[5,7] & Majid Afshar[4,5] ✉

Electronic Health Records (EHRs) contain vast clinical data that are difficult for providers to synthesize. Generative AI with Large Language Models (LLMs) can summarize records to reduce cognitive burden, but ensuring accuracy requires reliable evaluation. Human review is the gold standard but is costly and slow. To address this, we introduce and validate an automated LLM-based method to assess real-world EHR multi-document summaries. Benchmarking against the validated Provider Documentation Summarization Quality Instrument (PDSQI), our LLM-as-a-Judge framework demonstrated strong inter-rater reliability with human evaluators. GPT-o3-mini achieved an intraclass correlation coefficient of 0.818 (95% CI 0.772–0.854), a median score difference of 0 from humans, and completed evaluations in 22 seconds. Overall, reasoning models excelled in inter-rater reliability, particularly for evaluations requiring advanced reasoning and domain expertise, outperforming non-reasoning, task-trained, and multi-agent approaches. By automating high-quality evaluations, a medical LLM-as-a-Judge provides a scalable, efficient way to identify accurate, safe AI-generated clinical summaries.

Electronic Health Records (EHRs) capture a large volume of clinical documentation, which can lead to information overload among clinicians. One in five patients admitted to the hospital arrives with an EHR comparable in length to Herman Melville's classic novel *Moby Dick* (206,000 words)[1]. While centralized documentation in EHRs has improved information accessibility and workflow efficiency[2], the volume of data limits the practicality of traditional manual review processes. Clinicians face the formidable task of navigating expansive patient records, thereby increasing the risk of missing crucial clinical information[3–5].

Generative Artificial Intelligence (GenAI), particularly through advancements in Large Language Models (LLMs), has emerged as a promising solution for these challenges by automatically summarizing clinical information. LLMs, such as OpenAI's Generative Pre-Trained Transformer (GPT)-4 with a context window capable of handling up to 128,000 tokens[6], enable comprehensive, multi-document patient summaries sparking the integration of GPT-4 into clinical workflows[7–9]. Recognizing the potential of GenAI to improve clinical workflows, both established EHR vendors and venture-backed start-ups have prioritized the development of summarization capabilities. Nonetheless, summarization tasks in clinical contexts require high precision to extract clinically relevant details for a given clinician specialty, factual accuracy, and abstraction capabilities, and these are areas where LLMs remain vulnerable to issues such as hallucinations, omissions, and inaccuracies[10–13]. Other critical concerns that are a unique component of multi-document clinical summaries with long inputs are observed errors with lost-in-the-middle effects where performance degradation occurs with missed details or chronological errors[14,15].

Given the importance of healthcare delivery, health systems need rigorous evaluation methodologies to implement LLM-generated summaries safely. Human evaluation remains the gold standard for assessing the accuracy, completeness, and clinical relevance of these summaries[16], but this reliance on clinical experts is resource intensive. Traditional automated metrics like Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bidirectional Encoder Representations from Transformers (BERT) Score were developed for natural language tasks with simple reference text and inadequately capture the nuanced and contextual demands of clinical language generation. Our prior work has also shown these metrics to poorly

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA. [2]Department of Biomedical Informatics, University of Colorado - Anschutz Medical, Aurora, USA. [3]Epic Systems, Verona, USA. [4]Department of Medicine, University of Wisconsin, Madison, USA. [5]UW Health, Madison, USA. [6]Department of Computer Science, Loyola University, Chicago, USA. [7]BerbeeWalsh Department of Emergency Medicine, University of Wisconsin, Madison, USA. ✉e-mail: mafshar@medicine.wisc.edu

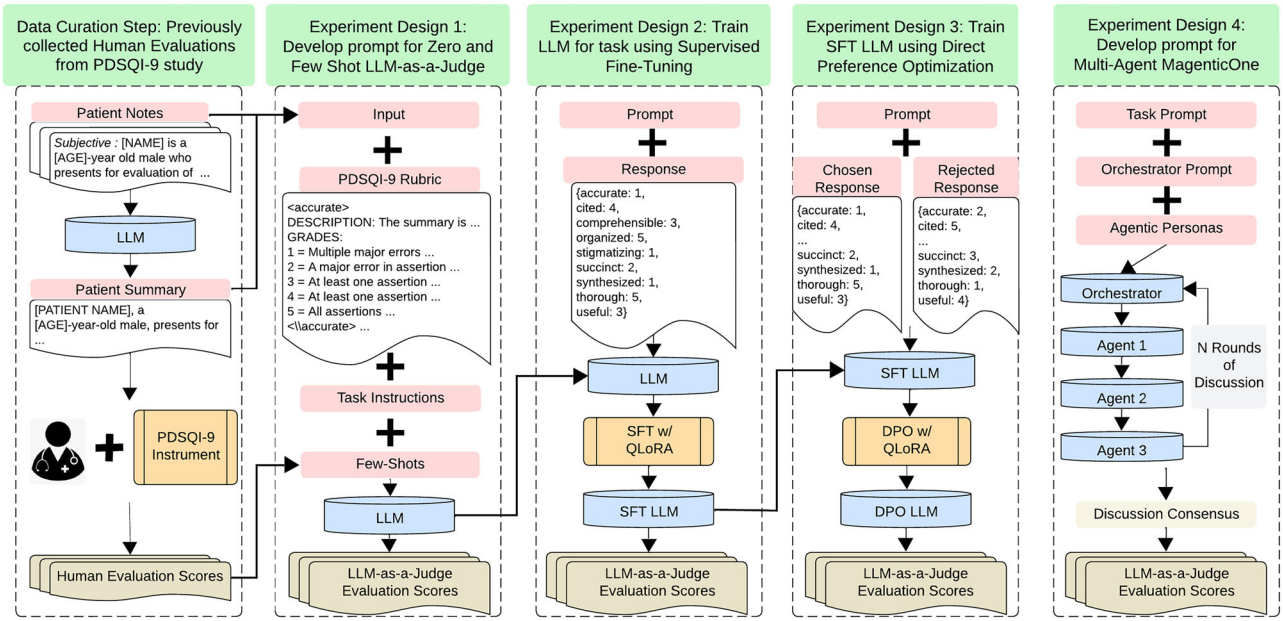

Fig. 1 | Study overview. Five distinct training strategies for large language models using the PDSQI-9 instrument were evaluated. The experiments comprised expert-driven prompt engineering, supervised fine-tuning, direct preference optimization, and multi-agent architectures, representing the LLM-as-a-Judge framework for clinical summarization.

correlate with human evaluations in the medical domain[17,18]. Specifically, they lack sensitivity to factual accuracy, logical coherence, and clinical relevance, which are needed in healthcare applications. These metrics typically rely on surface-level heuristics, such as lexicographic and structural measures, which are inadequate to evaluate the complexities of medical text[18]. Abstractive summaries are challenging to evaluate because the generated text might not directly correspond to any part of the original documentation. Automated metrics often fail to capture the inferences, synthesis, and new language produced in abstractive summarization, limiting their applicability to tasks like clinical documentation. These limitations highlight the need for automated evaluation methods beyond semantic and lexical similarity to better assess the quality and clinical relevance of the generated text.

Recent systematic reviews highlight major gaps in current human evaluation practices, noting a lack of psychometrically validated instruments specifically designed for clinical summarization using real-world, multi-document EHR data[19]. Tam et al. emphasized that only a small minority of evaluation studies involve multiple clinician experts or psychometric validation[20]. Historically, clinical documentation quality has been assessed using instruments such as the Physician Document Quality Instrument (PDQI-9)[21]. Although the PDQI-9 has been applied to LLM-generated summaries[22], it was not designed to capture LLM-specific phenomena such as hallucinations. To address this gap, we previously developed and validated an LLM-centric adaptation of the PDQI-9 for evaluating clinical summaries produced by LLMs from the EHR called the *Provider Documentation Summarization Quality Instrument* ***(PDSQI)-9***[23]. Developed using a semi-Delphi consensus methodology and adequately powered at 80%, the PDSQI-9 instrument demonstrates excellent psychometric properties, including high discriminant validity and inter-rater reliability validated by physician raters. The instrument includes nine attributes: *Cited*, *Accurate*, *Thorough*, *Useful*, *Organized*, *Comprehensible*, *Succinct*, *Synthesized*, and *Stigmatizing*. During the development of the instrument, particular focus was placed on the vulnerability of the attribute *Accurate* to hallucinations and the attribute *Thorough* to omissions to capture known LLM summarization issues. The PDSQI-9 showed excellent internal consistency, with an intraclass correlation coefficient (ICC) of 0.867 (95% CI: 0.867–0.868). However, the PDSQI-9 was validated with clinician evaluators who averaged 10 minutes per evaluation. Thus, LLM summarization, while intended to increase workflow efficiency, paradoxically requires significant time, effort, and expertise to validate. To reconcile these competing demands, innovative automated evaluation methods leveraging LLMs themselves as evaluators, termed "LLM-as-a-Judge," offer promise[24]. These approaches harness LLMs' contextual comprehension and reasoning capabilities to automate evaluation processes traditionally conducted by human experts[25,26]. However, limited data exist for their performance in the medical domain.

In this study, we propose and evaluate the efficacy of a medical LLM-as-a-Judge framework using the PDSQI-9 instrument. This instrument served as the benchmark to compare LLM-driven evaluations directly against human expert assessments. The primary outcome of the comparison was the intraclass correlation coefficient (ICC), in line with the original PDSQI-9 evaluation method. We systematically evaluated state-of-the-art open- and closed-source LLMs as judges using different prompting strategies, including zero-shot, few-shot, supervised fine-tuning (SFT), direct preference optimization (DPO), and multi-agent frameworks (e.g., MagenticOne[27]) as illustrated in Fig. 1, we aimed to establish the reliability and practicality of automating clinical summarization evaluations. We hypothesize that the LLM-as-a-Judge framework will achieve inter-rater reliability comparable to expert human evaluators, thus providing an efficient, scalable solution to the evaluation bottleneck posed by GenAI-driven clinical summarization. Cross-task validation was conducted on a separate summarization task focused on diagnoses using another EHR with the Problem List BioNLP Summarization (ProbSum) 2023 Shared Task[28].

Results

Study characteristics

To perform the LLM-as-a-Judge experiments, the original corpus of notes, LLM-generated summaries, and physician expert evaluation scores were utilized from the original PDSQI-9 study[23]. The PDSQI-9 instrument with attributes and scoring rules is available at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/pdsqi-9. The LLM-generated summaries were scored using clinical notes from the provider's perspective during an office visit (index encounter), representing a real-world clinical situation where the provider utilized a summary of the patient's prior encounters. The original 200 summaries comprising 2200 questions were split into a training/development set of 160 summaries and a test set of

Table 1 | Study corpus characteristics

	Train/Development	Test	*P*-value
Summaries, *n*	160	40	
Provider Specialty, *n* (%)			0.299
Primary Care	78 (48.76%)	14 (35.00%)	
Surgical Care	39 (24.36%)	13 (32.50%)	
Emergency/ Urgent Care	21 (13.12%)	4 (10.00%)	
Neurology/ Neurosurgery	11 (6.88%)	3 (7.50%)	
Other Specialty Care	11 (6.88%)	6 (15.00%)	
Number of Notes, *n* (%)			0.078
Three	37 (23.12%)	16 (40.00%)	
Four	45 (28.12%)	7 (17.50%)	
Five	78 (48.75%)	17 (42.50%)	
Length of Notes (words), median (IQR)	3050 (2174, 4128)	2816 (2137, 4364)	0.931
Length of Notes (tokens), median (IQR)	6445 (4366, 8665)	5845 (4428, 9086)	0.949
Length of LLM Input* (words), median (IQR)	4746 (3871, 5831)	4788 (3782, 4788)	0.975
Length of LLM Input* (tokens), median (IQR)	9681 (7448, 11704)	8920 (7398, 11850)	0.854
Length of Summary (words), median (IQR)	328 (191, 498)	250 (179, 414)	0.418
Length of Summary (tokens), median (IQR)	566 (368, 869)	425 (340, 787)	0.221

*Instruction + Rubric + Notes + Summary.

The study corpus consisted of 200 unique patient summaries split into development and test sets. Provider specialties were grouped into five categories: Primary Care (Internal Medicine, Family Medicine, Pediatrics); Surgical Care (General Surgery, Orthopedics, Ophthalmology, Urology), Neurology/Neurosurgery, Emergency/Urgent Care, and Other Specialty Care (Gynecology, Dermatology, Psychiatry, Sleep, and Anesthesiology). The median and inter quartile range (IQR) for the length of the notes, summaries, and LLM input are provided for both the number fo words and number fo tokens.

40 summaries for the LLM-as-a-Judge experiments. The provider perspective for the index encounter represented multiple specialties grouped as Primary Care, Surgical Care, Emergency/Urgent Care, Neurology/Neurosurgery, and Other Specialty Care. Each patient had 3, 4, or 5 encounters prior to the index encounter that the LLM summarized over. The characteristics of the datasets are shown in Table 1. No differences were observed in the distribution of notes, token counts, or specialty types between train and test datasets ($p > 0.05$ for all). To avoid over-representation of any evaluator's scores in the training set, we up-sampled evaluations from each evaluator to ensure an even distribution across all seven evaluators.

Single LLM framework (LLM-as-a-Judge)

The LLM-as-a-Judge was provided with the same information given to human evaluators with instructions regarding the original patient notes, the generated summary, and the PDSQI-9 rubric. Figure 2 provides an overview of the complete input to the LLM-as-a-Judge, and an example of the full prompt is available at https://github.com/epic-open-source/evaluation-instruments/tree/main/src/evaluation_instruments/instruments/pdsqi_9. The primary outcome of this study was the agreement between the LLM-as-a Judge and the seven human evaluators using the intraclass correlation coefficient (ICC), which aligned with the original PDSQI-9 evaluation methodology. This was assessed by comparing the median scores of the seven human evaluators with the median scores from seven iterations of the LLM-as-a-Judge. The Wilcoxon signed-rank test was also used to assess the median score difference between humans and LLMs as judges. The complete set of experiments and hyperparameter tuning with Bayesian Optimizations are described further in the Methods. Among the single LLM-as-a-Judge results, GPT-o3-mini (2024-01-31) 5-shot demonstrated the highest ICC reaching 0.818 (95% CI 0.772, 0.854) and had a median score difference of 0 (IQR: 0,1; p-value < 0.001). In sensitivity analyses, the LLM-as-a-Judge replaced a human evaluator or was added as an additional reviewer. In both scenarios no significant change was noted in the ICC scores using GPT-o3-mini (p-values > 0.3). While the primary outcome was ICC, secondary measures of inter-rater agreement were also examined with Krippendorf's α and Gwet's Ac2 and are reported separately in Supplementary Table 1. These metrics were included to provide a more comprehensive assessment of inter-rater reliability, capturing different assumptions

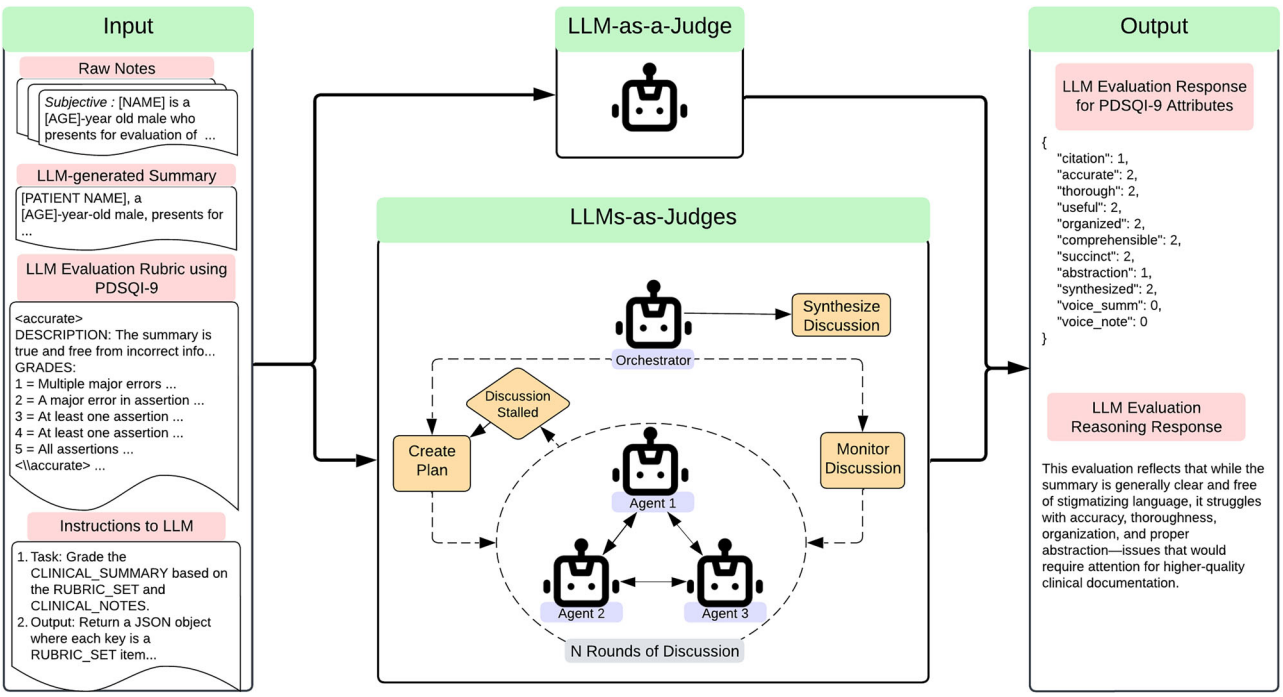

Fig. 2 | Single LLM and multi-agent framework overview. Conceptual diagram showing implementation and example input-output flows for each framework.

Table 2 | LLM-as-a-judge alignment with human reviewers

LLM-as-a-Judge	Strategy	ICC (95% CI)	Median Difference (IQR)	Wilcoxon *P*-Value
PDSQI-9 Internal Validation				
GPT-o3-mini (2025-01-31)	Zero-Shot	**0.803** (0.753, 0.842)	0 (0,1)	0.007
GPT-o3-mini (2025-01-31)	5-Shot	**0.818** (0.772, 0.854)	0 (0, 1)	<0.001
GPT-4o (2024-08-06)	Zero-Shot	0.730 (0.663, 0.784)	0 (0,1)	0.216
DeepSeek-R1 761B	Zero-Shot	0.762 (0.703, 0.810)	0 (0,1)	0.063
DeepSeek Distilled Qwen 2.5 32B	Zero-Shot	0.721 (0.651, 0.777)	0 (0,1)	0.001
Llama 3.1 8B	Zero-Shot	0.332 (0.165, 0.465)	0 (−2,1)	<0.001
Llama 3.1 8B 4-bit quantized/64 adapters	SFT	0.356 (0.195, 0.485)	−2 (−3,0)	<0.001
Llama 3.1 8B 4-bit quantized/8 adapters	SFT + DPO	0.560 (0.451, 0.648)	0 (0,2)	<0.001
Mixtral 8 × 22B	Zero-Shot	0.733 (0.667, 0.786)	1 (0,1)	<0.001
Mixtral 8 × 22B 4-bit quantized/8 adapters	SFT	0.740 (0.675, 0.792)	1 (0,1)	<0.001
Mixtral 8 × 22B 4-bit quantized/8 adapters	SFT + DPO	0.746 (0.683, 0.797)	1 (0,1)	<0.001
Multi-Agent Framework: GPT-o3-mini orchestrator & agents	Zero-Shot	0.768 (0.710, 0.814)	0 (-1,1)	0.353
ProbSum Cross-task Validation				
GPT-o3-mini (2025-01-31)	5-Shot	0.710 (0.662, 0.752)	0 (0,0)	0.028
DeepSeek-R1 761B	Zero-Shot	0.687 (0.634,0.732)	0 (0, 0.75)	0.835
DeepSeek Distilled Qwen 2.5 32B	Zero-Shot	0.622 (0.554,0.679)	0 (0,2)	<0.001
Mixtral 8 × 22B	Zero-Shot	0.596 (0.527,0.654)	0 (0,0)	<0.001
Mixtral 8 × 22B - SFT	Zero-Shot	0.641 (0.578,0.705)	0 (0,0)	0.032
Mixtral 8 × 22B - DPO	Zero-Shot	0.666 (0.607,0.726)	0 (0,0)	0.737

Subset of LLM-as-a-Judge approaches chosen based on superior performance or relevance to key evaluation findings, with intraclass correlation and median score difference from expert evaluators. The complete set of experimental results are provided in the Supplementary Tables.
The bold values are meant to highlight the best performing models ICC values.

about rater behavior and chance agreement that complement the ICC results. Across measures, GPT-o3-mini was the top performing model (α = 0.677, Ac2 = 0.826). Of note, smaller open-source models with poor performance had substantial gains with additional training with SFT using Quantized Low Rank Adapters (QLoRA) and DPO, with Llama 3.1 8B improving from an ICC of 0.332 to 0.560. This was less pronounced with new, larger open-source models such as Mixtral 8 × 22B, which showed competitive performance with zero-shot and had minimal gains with additional fine-tuning. Table 2 demonstrates a subset of results for models that exhibited the strongest performance or held particular interpretive value. Full experimental results are provided in Supplementary Table 1, including all few-shot and SFT results.

Multi-Agent Framework (LLMs-as-Judges)

In addition to the single LLM-as-a-Judge framework, several Multi-Agent Frameworks (LLMs-as-Judges) were also evaluated. Using Microsoft's AutoGen implementation with the MagenticOneGroupChat setup[27], several LLMs-as-Judges participated in rounds of discussion. The orchestrator created the initial plan, monitored discussion among judges, and reported the final consensus on evaluative scores as visualized in Fig. 2. The LLMs-as-Judges were selected from one of two groups of agentic personas. The first group contained LLMs-as-Judges intended to represent personas that a primary care physician might embody: the Analytical Perfectionist, the Efficient Multitasker, and the Collaborative Team Player. Each agent was tasked with prioritizing a specific aspect of high-quality summarization in alignment with its assigned persona. The second group contained LLMs-as-Judges intended to represent different perspectives along an ordinal scale—high, low, and middle scoring—to simulate a range of reviewer stances. A mixture of open- and closed-source models were incorporated in additional trials, building on insights from the single-agent experiments. Initial testing used GPT4o for all agents and either GPT4o or GPT-o3-mini as the orchestrator, but later testing incorporated GPT-o3-mini as the orchestrator, Mixtral 8 × 22B as the high-scoring agent, and GPT4o as the low-scoring agent. The best multi-agent approach used GPT-o3-mini as the orchestrator, high-scoring agent, and low-scoring agent demonstrating an ICC of 0.768 (95% CI 0.710, 0.814) and a median score difference of 0 (IQR: -1,1; p-value = 0.353). The extended results can be found in Supplementary Table 2.

Cross-task validation

Cross-task validation was conducted for the top-performing LLM-as-a-Judge models using the Problem List BioNLP Summarization (ProbSum) 2023 Shared Task[28], one of the natural language generation tasks designed for summarization in the medical domain. The objective of this task was to generate summaries of a patient's active problems or diagnoses based on the Subjective, Objective, and Assessment sections of daily progress notes. The Plan section was used to label the gold standard diagnoses as references by expert medical annotators. These progress notes were sourced from the Medical Information Mart for Intensive Care (MIMIC)-III EHR database. The goal of the cross-task evaluation was to assess the transferability of the LLM-as-a-Judge framework when applied to a different evaluation rubric. We applied the same prompt development strategy used for the PDSQI-9 instrument to a separate, previously published rubric designed for the ProbSum task[17]. In this evaluation, the LLM-as-a-Judge, or LLMs-as-Judges, were provided the same information as the human evaluators: (1) patient note to be summarized; (2) generated summary; (3) annotated gold standard; and (4) the ProbSum evaluation rubric(also available in Supplementary Note 1). The final cross-task validation test set comprised 31 summaries and 792 rubric attribute scores. No training was performed, given the leading performance of zero & few-shot single and multi-agent LLM-as-a-Judge frameworks. The median length of the prompt was 1606 words (IQR 1545-1653), and the median token count was 3224 (IQR 3115-3353). The top-performing model remained GPT-o3-mini demonstrating an ICC of 0.710 (95% CI 0.662, 0.752). The full set of results are presented in Table 2.

Table 3 | LLM-as-a-Judge Costs

Inference Costs			
Judge	**Strategy**	**Mean Inference Time (sec)**	**Mean Cost ($)**
Human	—	600	50.00
GPT-4o (2024-08-06)	Zero-Shot	12	00.03
GPT-4o (2024-08-06)	Few-Shot	14	00.10
GPT-o3-mini (2025-01-31)	Zero-Shot	16	00.02
GPT-o3-mini (2025-01-31)	Few-Shot	22	00.05
DeepSeek-R1 761B	Zero-Shot	35	00.02
DeepSeek Distilled Qwen 2.5 32B	Zero-Shot	25	00.00
Mixtral 8 × 22B	Zero-Shot	22	00.00
Mixtral 8 × 22B	Few-Shot	25	00.00
Multi-Agent Framework: GPT-o3-mini orchestrator GPT-o3-mini agents	Zero-Shot	69	00.16
Training Costs			
Judge	**Strategy**	**Training Time (hrs)**	**GPUs (80GB H100)**
Llama 3.1 8B	SFT	2.5	1
Llama 3.1 8B	DPO	6	2
Mixtral 8 × 22B	SFT	24	2
Mixtral 8 × 22B	DPO	60	2

The average inference time (in seconds) and average cost (in U.S. dollars) for a single evaluation from the top-performing LLM-as-a-Judge approaches compared to a single human evaluator. The cost for a single human evaluator was calculated based on a median minimum physician consulting rate of $300/hour. Additionally, the training time (in hours) and the number of 80GB NVIDIA H100 GPUs required for parameter efficient fine-tuning and direct preference optimization of Mixtral 8 × 22B and Llama 3.1 8B are included.

Cost analysis

Table 3 presents the inference costs associated with each LLM-as-a-Judge. In this study, GPT-4o, GPT-o3-mini and DeepSeek R1 were operated within the secure environment of the health system's HIPAA-compliant Azure cloud. The smaller, open-source LLMs were downloaded from HuggingFace[29] to HIPAA-compliant, on-premise servers. GPT-o3-mini, using 5-shot prompting completed an evaluation in an average of 22 seconds, a 96% reduction from the human average of approximately 600 seconds. The cost for running a single evaluation, including the Microsoft Azure cloud API fees, averaged 5 cents. The costs associated with training Llama 3.1 8B and Mixtral 8 × 22B using SFT and DPO are presented in Table 3. Both instances of DPO training required two-NVIDIA 80 GB H100 GPUs. Although Llama 3.1 8B saw large gains in performance with training, Mixtral 8 × 22B saw only minimal gain despite the over 24 hour training time across 50 and 15 epochs for SFT and DPO, respectively. For the multi-agent workflows, the cost substantially increases given the multiple rounds of discussion and multiple agents deployed.

Error analysis

The LLM-generated summaries being evaluated were a mix of generations by GPT4o, Mixtral 8 × 7B, and Llama 3-8B. Previous literature has found that LLMs favor their own generations[30]. To ensure that the validity of the evaluations was not subject to the source of the generated summary, a comparative analysis was performed when GPT4o or Mixtral 8 × 22B were used as the LLM-as-a-Judge on each of the three possible sources. No differences were found between the two judges' ICCs with human evaluators (p-values > 0.2). When compared with each other directly, GPT4o LLM-as-a-Judge is a harsher critic on summaries generated by itself than Mixtral 8 × 22B LLM-as-a-Judge is, producing scores one or more points lower on the 5-point Likert scale in 55% of the evaluation attributes. To further examine potential sources of bias, a linear mixed-effects model was fit to assess whether differences between LLM and human scores varied by provider specialty, model type, note length, or summary length. While there were some small variations, all estimated effects were less than a quarter of a point. Given that scoring was performed using 5-point Likert scales, differences would need to exceed one point to meaningfully impact conclusions. This suggests there was no meaningful or systematic bias in how the LLMs judged the summaries.

When comparing scores produced by a reasoning model, GPT-o3-mini, and a non-reasoning model, GPT4o, the advantages of a reasoning model in the scoring across the PDSQI-9 attributes were apparent. This is likely due to how reasoning models generate responses — by engaging in a step-by-step thought process during inference (i.e., test-time computation, or response generation). In this case, test-time computation refers to the inference phase — when the model is generating outputs (as opposed to training). Reasoning models often involve more extensive computation during inference because they explicitly simulate intermediate reasoning steps using chain-of-thought "thinking" steps. In contrast, non-reasoning models produce outputs more directly, without these additional intermediate steps. Figure 3 depicts the distribution of the score differences between each LLM-as-a-Judge and the human evaluators. Each attribute of the PDSQI-9 rubric is graded on a 5-point Likert scale, except for *Stigmatizing*, which is the only exclusion because of its binary scoring. Overall, GPT-o3-mini aligned with human scores more closely than GPT-4o across almost every PDSQI-9 attribute. The most pronounced differences between the models are for the *Cited*, *Organized*, and *Synthesized* attributes, which reflect the ability of GPT-o3 as a reasoning model to present a comprehensive view and perform abstractive reasoning over the notes.

We also compared reasoning outputs of GPT-4o and GPT-o3-mini. In one illustrative example, GPT-o3-mini assigned a *Synthesized* score aligned with the human median score of 2, while GPT-4o assigned a score of 5. GPT-4o's reasoning was as follows: "Synthesis: The summary demonstrates excellent synthesis of information by grouping pertinent medical themes and integrating them for an overall clinical synopsis. Grade: 5." In contrast, GPT-o3-mini reasons, "For synthesis, the summary provides several independent assertions that are loosely grouped but does not generate an integrated understanding or prioritization of the acute problem aimed at the EM provider. This gives it a lower synthesis score of 2." This contrast highlights GPT-o3-mini's ability to engage in deliberative reasoning closely mirroring the human evaluators' process, whereas GPT-4o's reasoning appears more superficial.

GPT-o3-mini also performs better as a standalone evaluator than within a multi-agent framework, even when all agents involved are GPT-o3-

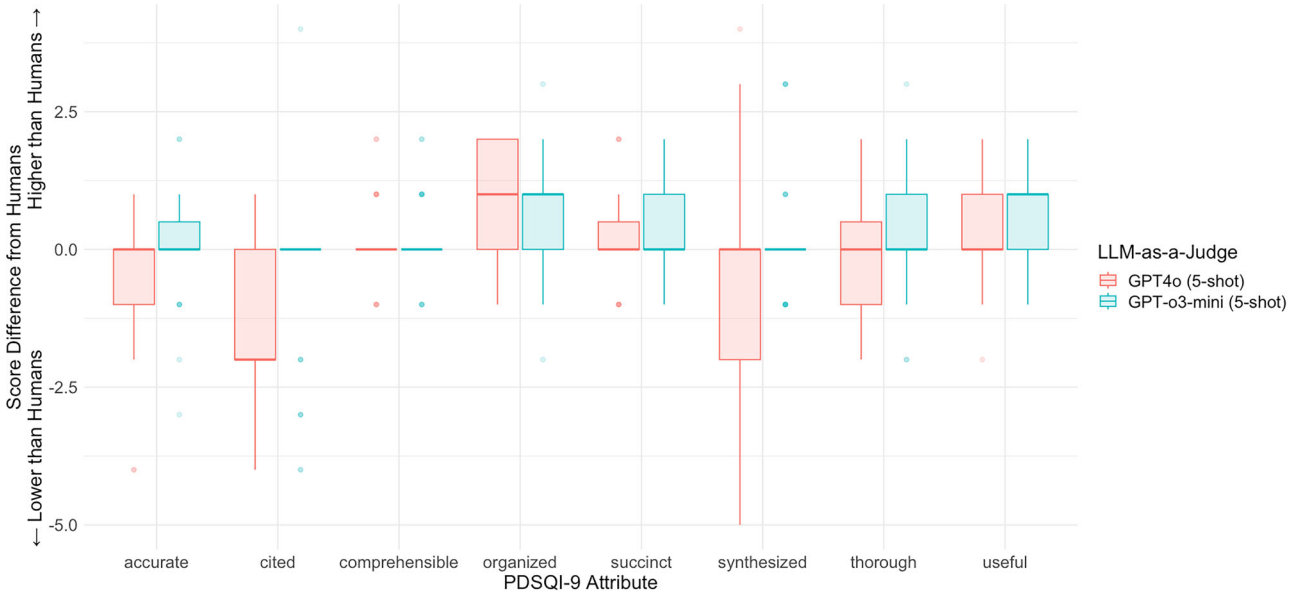

Fig. 3 | Absolute differences between human evaluator and LLM-as-a-judge scores. Absolute score differences between human evaluators and LLM-as-a-Judge outputs (GPT-4o or GPT-o3-mini) across each attribute of the PDSQI-9 instrument. The largest differences appear in the *Cited*, *Organized*, and *Synthesized* attributes, highlighting GPT-o3-mini's advantage as a reasoning model capable of presenting a comprehensive view and performing abstractive reasoning over clinical notes.

mini models. For instance, in evaluating the *Organized* attribute, GPT-o3-mini independently aligns with the human median score of 3, whereas the multi-agent setup assigns a score of 5. The reasoning outputs from the individual agents elaborate on the scoring. The low-scoring agent states: "The summary presents the information in chronological order and groups events logically. I award a 5 here." Meanwhile, the high-scoring agent states: "The events are presented chronologically, and the progression of care is clear, making the summary easy to follow. I assign a 5." Because both agents converge on the same assessment, the orchestrator concludes: "For Organized, the summary is presented in a clear chronological order that supports understanding of the clinical course. Both agents agree on a perfect score, so we assign a 5." In contrast, GPT-o3-mini acting alone offers a more discerning evaluation: "In terms of organization, the summary follows a chronological order, but the grouping is not completely coherent due to the one mis-timed assertion, leading to a score of 3." This comparison highlights GPT-o3-mini's capacity for a more analytical deliberation as a single judge rather than relying on consensus within a multi-agent system that may overlook subtleties by converging prematurely on higher scores.

Discussion

This study introduces a medical LLM-as-a-Judge, an automated approach for evaluating the quality of clinical multi-document summaries generated by LLMs. Using the validated PDSQI-9 human evaluation instrument as a benchmark, the GPT-o3-mini model achieved strong inter-rater reliability comparable to expert human evaluators (ICC = 0.818). Incorporating GPT-o3-mini as an additional or substitute evaluator did not significantly alter the reliability metrics previously established for the PDSQI-9 instrument. Cross-task validation on the ProbSum task similarly confirmed high reliability. Furthermore, GPT-o3-mini completed evaluations faster (average 22 seconds) than human evaluators (average 600 seconds), potentially translating into substantial reductions in labor within clinical workflows. By automating high-quality evaluation of summarization outputs, a medical LLM-as-a-Judge provides a scalable and efficient approach to rapidly identify accurate and safe generative AI implementations in healthcare settings.

The concept of LLM-as-a-Judge has shown strong performance in general domain tasks, with benchmarks like AlpacaEval 2.0[31] and MT-bench[32] demonstrating that GPT-4 can effectively assess the output quality of another LLM, maintaining a strong correlation with human judgments. However, the clinical domain introduces unique challenges due to its heightened sensitivity to hallucinations and omissions. As a result, it requires a more specialized and nuanced approach to evaluation—one that goes beyond the general-purpose queries addressed by these existing benchmarks. To our knowledge, our study is the first to showcase the application of LLM-as-a-Judge for clinical summarization tasks, tested with a validated human assessment instrument.

The LLMs tested encompassed state-of-the-art reasoning and non-reasoning models, including GPT-4o, GPT-o3-mini, and DeepSeek R1, as well as smaller, open-source models like Mixtral 8 × 22B and Llama 3.1 8B. The study utilized a HIPAA-secure API available in the Azure AI Foundry cloud environment, and this infrastructure may not be available at many health systems. Other open-source models, including DeepSeek Distilled Qwen 2.5 32B and Mixtral 8 × 22B, also achieved high inter-rater reliability (ICCs > 0.7), positioning them as viable alternatives to the larger GPT-o3-mini model that can be downloaded on to on-premise servers for free access and use. Interestingly, few-shot prompting led to performance declines for GPT-4o and other models. This may reflect that well-instructed models can perform optimally with direct guidance, and additional examples may introduce noise. In contrast, models like GPT-o3-mini benefited from examples, suggesting that few-shot effectiveness is model-specific.

The benefit of DPO varied substantially between models, with LLaMA 3.1 8B improving from an ICC of 0.356 to 0.560, while Mixtral 8 × 22B showed minimal gains (0.740 to 0.746). Several factors likely contribute to this divergence. First, LLaMA 3.1 8B started from a significantly lower baseline, suggesting greater capacity for improvement with preference-based optimization. In contrast, Mixtral's high starting performance may reflect a ceiling effect. Second, LLaMA 3.1 8B is a dense transformer model, whereas Mixtral 8 × 22B is a sparse mixture-of-experts (MoE) model with 12.9B active parameters per forward pass but 46.7B total parameters. MoE models are designed to scale efficiently by routing tokens through only a subset of expert networks. MoE models may also exhibit greater preference alignment if expert activation is inconsistent across training examples, which likely happens with multiple human evaluators for training data. Finally, Mixtral was released more recently with strong zero-shot and few-shot reasoning capabilities across diverse benchmarks, suggesting that its pretraining corpus and training objectives may have already encoded alignment features that DPO aims to reinforce. These findings suggest that the effectiveness of preference-based optimization is highly model-

dependent, shaped by a combination of architectural complexity, pre-training quality, and starting performance.

While DPO significantly improved LLaMA 3.1 8B's performance, the model's inherent limitations prevented it from matching the higher baseline and ultimate performance of other models. Notably, across zero-shot, SFT, and DPO reward scoring settings, LLaMA 3.1 8B consistently produced polarized scores concentrated at the extremes of the scale, with limited use of intermediate values. This lack of granularity not only reduced the fidelity of evaluation scores across zero-shot and SFT settings but also impaired contrastive learning during DPO by failing to capture subtle distinctions between candidate responses. Additionally, frequent stability and formatting issues—including empty or malformed outputs and non-numeric responses—introduced noise and complicated reliable evaluation and training. While prompt formatting partially alleviated these issues, LLaMA 3.1 8B remained less consistent and less reliable than other models in producing valid and meaningful scoring outputs. Together, these observations highlight that beyond model size and baseline capabilities, consistent scoring behavior and output stability are essential for reliable performance on this task.

In the error analysis comparing models used for LLM-as-a-Judge, the strength of the reasoning models stood out. Both GPT-o3-mini and DeepSeek R1 demonstrated the highest levels of inter-rater reliability with expert human evaluators. This was especially clear when scoring attributes that require advanced reasoning and expert domain knowledge from human reviewers. Specifically, the attributes *Cited*, *Organized*, and *Synthesized* showed the most substantial improvements in scoring by the reasoning models over their state-of-the-art non-reasoning counterparts. These attributes require a comprehensive evaluation of the notes to determine whether the summary captures the appropriate breadth of information. Additionally, notable gains were observed in the *Thorough* attribute, further illustrating the models' ability to discern information most relevant to the intended provider audience, rather than a generic clinical reader.

The prompt engineering process in this study revealed key insights for deploying an LLM-as-a-Judge. At the core of this process is the importance of a reliable evaluation rubric. Without a well-defined rubric, the LLM-as-a-Judge's performance will be as flawed as that of a human judge under similar conditions. In other words, the quality of the evaluation is directly tied to the clarity and precision of the rubric. Each rubric attribute must have a well-defined purpose, with explicit criteria outlining its meaning— these are essential elements in the prompt design of the LLM-as-a-Judge. Each attribute must be precisely named to reflect what is being assessed on the associated Likert scale. For example, naming an attribute "Implausible" but defining it such that a score of 5 represents high plausibility can undermine the interpretability of the LLM-as-a-Judge's results. Furthermore, the Likert scale itself must be concretely and objectively detailed to prevent the LLM-as-a-Judge from making assumptions that might not align with human interpretation. When there are clear and precise definitions for the attributes and the rating scale, we showed that the LLM-as-a-Judge's evaluations are consistent and accurate compared with human evaluation. Additionally, enforcement of output formatting rules is more effective when the LLM-as-a-Judge is instructed to verify the formatting before generating a response. The final prompt design is a result of 1) establishing an evaluation instrument with strong construct and content validity and 2) performing multiple experiments in both prompt engineering and judge performance against a high quality dataset of human expert evaluations. The final prompting framework for both single LLM-as-a-judge and the more expensive multi-agent framework using Microsoft's MagenticOne framework are available for health systems at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/pdsqi-9. We recommend health systems follow the prompt design we share and only remove attributes as needed for the task, without changes in language or format to maintain the results we demonstrated in this work.

Recent advances in multi-agent LLM workflows have demonstrated several advantages over single-model systems. Multi-agent architectures enhance reasoning and factual accuracy[33], promote divergent thinking[34], and leverage the complementary strengths of multiple LLMs[35]. Beyond general-purpose frameworks, clinical domain-specific multi-agent systems have also shown significant promise. MDAgents[36] introduces a dynamic collaboration structure that adapts to the complexity of each clinical question, enabling either solo reasoning or group deliberation among LLMs. This approach draws inspiration from real-world medical decision-making, where task complexity determines whether decisions are made individually or collaboratively. While multi-agent frameworks offer clear benefits over single-LLM methods, they also increase computational and operational costs as the number of agents grows. Additionally, they require careful tuning of multiple parameters, including agent personas, interaction protocols, and conversation length, which adds further complexity to their design and implementation.

While the multi-agent frameworks tested in this study did not outperform single-agent systems in terms of the primary outcome measure (ICC), they demonstrated a notable advantage in aligning with human evaluators compared with single-LLM approaches. A single GPT-o3-mini LLM-as-a-Judge has high correlation with human evaluators but typically produced scores that matched or exceeded human scores by one point on the Likert scale. In contrast, a multi-agent framework composed of multiple GPT-o3-mini agents had lower correlation with human evaluators, but the positive and negative score differences were distributed more evenly. This highlights how assigning distinct personas to evaluator agents, such as high and low scorers, within a multi-agent setup can more faithfully mimic the variability found among human reviewers, effectively harnessing the complementary perspectives of the system.

Several limitations occurred in this study. The clinical notes used to validate LLM-as-a-Judge came exclusively from UW Health and the MIMIC-III corpus, which may limit the generalizability of the findings across different health systems and clinical specialties. To address this limitation, external validation studies are needed to evaluate performance across diverse healthcare systems. While the test set size was limited to 40 summaries, each was evaluated by seven physicians and analyzed across 9 rubric dimensions, resulting in over 2500 data points across all model comparisons. The ICC metric used is well suited to this setting, and the narrow confidence intervals support the reliability of the observed agreement. In addition, both tasks used to validate the approach are summarization-based tasks, with inputs capped to stay within the context window limits of some open-source LLMs. While this design choice enabled direct prompting, it limits applicability to larger, more heterogeneous clinical records that can also be seen in practice. Exploring the scalability of this approach to larger clinical datasets remains a subject for future work. Moreover, extending the approach to other clinical language generation tasks, such as medical question answering, will require further validation. It is also important to note that this study did not include an analysis of potential biases in model scoring related to patient characteristics. Future work should explore these fairness considerations to better understand how LLM-as-a-Judge may behave across diverse clinical contexts. Finally, the SFT, DPO, and multi-agent LLMs-as-Judges had significant costs associated with their experimentation. For Mixtral 8 × 22B, a single SFT training run took 24 hours and a single DPO training run took 60 hours using 2-H100 NVIDIA GPUs. As a result, hyperparameter tuning was limited to a subset. To conserve computational resources, we report one complete test run for the resource-intensive multi-agent framework rather than multiple iterations like those performed for the single-agent approaches.

In conclusion, this study introduces and validates an automated method to evaluate clinical multi-document summaries using an LLM-as-a-Judge. Leveraging the PDSQI-9 instrument as a benchmark, the medical LLM-as-a-Judge framework demonstrates strong inter-rater reliability with human evaluators. This approach provides an efficient and scalable solution for assessing clinical summaries, significantly reducing the time and cost of thorough evaluations while maintaining high reliability. An important next step is integrating LLM generation and evaluation in a closed-loop system, where the LLM-as-a-Judge provides feedback to iteratively refine summaries until a quality threshold is met. While beyond the scope of this study,

such an approach could enable automated, real-time quality control in clinical summarization.

Methods

Study design and data corpus

The data used in this study comes from the original PDSQI-9 study[23], including the original corpus of notes, LLM-generated summaries, and scores from physician evaluators. The corpus of notes was designed for multi-document summarization and evaluation using inpatient and outpatient encounters from the University of Wisconsin Hospitals and Clinics (UW Health) in Wisconsin and Illinois between March 22, 2023 and December 13, 2023. The evaluation was conducted from the provider's perspective during the initial office visit (the 'index encounter')—the clinic appointment where the provider would benefit from a summary of the patient's prior visits with outside providers. Other inclusion and exclusion criteria were the following: (1) patient was alive at time of index encounter with provider; (2) patient had at least one encounter in 2023; and (3) excluded psychiatry notes. Psychotherapy notes were excluded due to their highly sensitive nature and additional regulatory protections under HIPAA and 42 CFR Part 2, which require approvals beyond the minimal necessary standard for research. Each patient had multiple encounters (3, 4, or 5) for which the LLM was tasked to generate a summary. To stay within the context window limitations of available LLMs during experimentation, we selected patients with 3 to 5 prior encounters for summarization. This range preserved meaningful longitudinal information while ensuring the full input (notes + summary + rubric) remained within the token limits of models like Llama 3.1 (8000-token maximum). The median number of tokens in the concatenated notes was 5101 (IQR: 3614–7464). This also allowed for feasible manual review by clinician evaluators. This study was reviewed by the University Wisconsin-Madison Institutional Review Board (IRB; 2023-1252) and determined it to be exempt human subjects research. The IRB approved the study with a waiver of informed consent. The summaries were evaluated using the perspective of the provider whose perspectives ranged across five different groups of specialties: Primary Care, Surgical Care, Emergency/Urgent Care, Neurology/Neurosurgery, and Other Specialty Care. The original 200 summaries were split into a training/development set of 160 and a test set of 40 for all LLM-as-a-Judge experiments. The characteristics of each set are shown in Table 1. The test set of 40 summaries was selected to ensure detailed expert evaluation across all PDSQI-9 attributes while maintaining feasibility. Each summary was reviewed by seven physicians, resulting in 280 attribute-level scores and allowing for robust reliability estimation.

The LLM-as-a-Judge approach outlined in this study was tested using several top-performing large language models (LLMs) from both open-source and closed-source categories, including reasoning and non-reasoning models. The open-source models used in this study were Mixtral 8 × 22B, Llama 3.1 8B, DeepSeek R1, DeepSeek Distilled Qwen 2.5 32B, DeepSeek Distilled Llama 8B, and Phi 3.5 MoE, while the closed-source models tested were GPT-4o with the 128K context window (version as of 2024-08-06) and GPT-o3-mini (version as of 2025-01-31). To assess model performance, we implemented five prompt engineering strategies: (1) Zero-Shot, (2) Few-Shot, (3) Supervised Fine-Tuning (SFT), (4) Direct Preference Optimization (DPO), and (5) Multi-Agent using MagenticOne. For GPT-4o and GPT-o3-mini, we used Zero-Shot and Few-Shot prompting for single LLM-as-a-Judge as well as for within the Multi-Agent framework. The DeepSeek models (R1 760B, Distilled Qwen 2.5 32B, and Distilled Llama 8B) were restricted to Zero-Shot prompting per recommendations in their GitHub model cards[37]. Mixtral 8 × 22B and Llama 3.1 8B were evaluated with a broader range of approaches — Zero-Shot, Few-Shot, SFT, and DPO — due to their smaller computational requirements, which make them more amenable for customizations compared with larger models.

GPT-4o, GPT-o3-mini, and DeepSeek R1 were operated within the secure environment of the health system's HIPAA-compliant Azure cloud. No PHI was transmitted, stored, or used by OpenAI for model training or human review. All interactions with proprietary closed-source LLMs were fully compliant with HIPAA regulations, maintaining the confidentiality of patient data. The smaller, open-source LLMs were downloaded from HuggingFace[29] to HIPAA-compliant, on-premise servers. The on-premise servers were equipped with two NVIDIA H100 80GB GPUs and were supported by an NVIDIA AI Enterprise software license. We followed the transparent reporting of a multi-variable model for individual prognosis or diagnosis (TRIPOD)-LLM guidelines, and the accompanying checklist is available in Supplementary Note 2.

Expert human evaluation rubric and scores

Human evaluations using the Provider Documentation Summarization Quality Instrument (PDSQI-9) were used to benchmark the LLM-as-a-Judge frameworks. The instrument consists of grading rubrics for nine attributes: *Cited*, *Accurate*, *Thorough*, *Useful*, *Organized*, *Comprehensible*, *Succinct*, *Synthesized*, and *Stigmatizing*. These attributes and their associated Likert or binary scales were developed using a semi-Delphi consensus methodology and validated across multiple psychometric properties. The instrument demonstrated strong internal consistency (Cronbach's α = 0.879, 95% CI: 0.867–0.891), high inter-rater reliability (ICC = 0.867, 95% CI: 0.867–0.868), and moderate agreement by Krippendorff's α = 0.575. Seven physician raters from varied specialties and levels of seniority evaluated a total of 779 summaries, scoring over 8000 items, achieving over 80% power for inter-rater reliability. No significant differences were observed in scoring between junior and senior raters, supporting reliability across experience levels[23]. The dataset spanned the full range of PDSQI-9 scores across all attributes. Although not every rater evaluated all 200 summaries, all human evaluations were included for training or validation of the LLM-as-a-Judge frameworks.

Single LLM design and implementation (LLM-as-a-Judge)

The task prompt used was iteratively designed to replicate the information provided to human reviewers during their evaluations. For each evaluation, reviewers received the full set of patient notes, the corresponding summary, and the specialty of the physician for whom the summary was intended (https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/pdsqi-9). In addition, the human evaluation rubric was reformatted for compatibility with LLM-as-a-Judge, following industry-standard prompting conventions. This included clearly marking each attribute with specific tags, such as < accurate > , and using uppercase text to distinguish attribute descriptions from grade definitions. Detailed instructions were also provided to the LLM-as-a-Judge regarding the task and expected output format. The LLM-as-a-Judge was instructed to return scores as a JSON-formatted string where each key corresponded to an attribute of the PDSQI-9. These instructions were human-drafted and refined through multiple iterations, informed by rounds of beta testing and output verification. They were also passed through the CliniPrompt software[7] and GPT4o for additional refinement. Manual prompts are designed using four key components: minimizing perplexity, in-context examples, chain-of-thought reasoning, and self-consistency. Prompts were iteratively refined through zero-shot (no examples) to few-shot (up to five examples) learning approaches, with random sampling from the training data. Full examples of the prompt are available at https://github.com/epic-open-source/evaluation-instruments/tree/main/src/evaluation_instruments/instruments/pdsqi_9. The same prompt was used for every model, except for the DeepSeek-based models, which had a " < think > " token appended to align with their recommended usage settings[37]. Additionally, inference parameters were fine-tuned for optimal performance across models. The number of shots for few-shot prompting was determined by testing with 1, 3, and 5 shots and selecting the configuration that yielded the best performance. The final hyperparameter settings for both zero-shot and few-shot prompting are detailed in Table 4.

For the training phase of the experiments, both SFT and DPO were implemented using Hugging Face's TRL library, with SFTTrainer for SFT and DPOTrainer for DPO[38]. The datasets for each approach were prepared according to the specifications of their respective trainer implementations.

Table 4 | Single LLM-as-a-Judge Inference Settings

Model	Quantization	Few-Shots	Temperature	Top P	Max New Tokens
GPT-4o (2024-08-06)	–	5	0.01	0.95	400
GPT-o3-mini (2025-01-31)	–	5	0.01	0.95	400
DeepSeek-R1 761B	–	–	0.01	0.95	400
DeepSeek Distilled Qwen 2.5 32B	8-bit	–	0.7	1.0	1000
DeepSeek Distilled Llama 8B	8-bit	–	0.7	1.0	1000
Mixtral 8 × 22B	4-bit	1	1.0	1.0	512
Llama 3.1 8B	4-bit	1	1.0	1.0	512
Phi 3.5 MoE	4-bit	–	1.0	1.0	1000

Table 5 | QLoRA settings for training

Model	Strategy	Quantization	LoRA Rank	LoRA Alpha	LoRA Drop-Out
Mixtral 8 × 22B	SFT	4-bit	8	8	0.1
Mixtral 8 × 22B	DPO	4-bit	8	8	0.1
Llama 3.1 8B	SFT	4-bit	64	256	0.0
Llama 3.1 8B	DPO	4-bit	8	8	0.1

Table 6 | Final training parameter settings

Model	Strategy	Batch Size	Learning Rate	Beta
Mixtral 8 × 22B	SFT	16	1e-7	–
Mixtral 8 × 22B	DPO	1	1e-7	0.7
Llama 3.1 8B	SFT	32	1e-4	–
Llama 3.1 8B	DPO	1	1e-7	0.5

Specifically, SFT required a dataset of prompt-completion pairs, while DPO required a dataset consisting of prompt-response pairs, including both a "chosen" and a "rejected" response. The prompt used for SFT was identical to that used in the zero-shot and few-shot prompting setups. Each completion for SFT consisted of a single JSON string representing the evaluation of a summary by one of the expert human reviewers. During training, evaluation scores from all human reviewers were included to fine-tune the model based on the collective feedback for each summary. Since DPO was applied to the already fine-tuned SFT version of the LLM, the chosen and rejected response pairs were constructed using the median response from the seven human reviewers. The chosen response was represented by a single JSON string reflecting this median, while the rejected response was a JSON string where each corresponding attribute grade was increased by one point on the Likert scale. The rationale behind making the rejected response have a higher Likert score than the chosen one was to encourage the model to be more conservative in its evaluations, rather than overly lenient.

To reduce computational costs while maintaining model accuracy, we employ quantization techniques that lower numerical precision from floating-point (e.g., float32) to fixed-point (e.g., int8)[39]. We also utilize Quantized Low-Rank Adapters (QLoRA), which combine quantization with low-rank adaptation to optimize efficiency[40]. QLoRA fine-tunes only selected layers while keeping the base LLM frozen, significantly reducing resource demands. QLoRA was employed for both SFT and DPO, with all training conducted using 4-bit quantization of the respective models through a BitsAndBytes configuration. The LoRA parameters, including rank and alpha, were optimized using a quasi-Bayesian approach implemented via Optuna[41]. Due to memory constraints, the LoRA parameters for Mixtral 8 × 22B in both SFT and DPO were restricted, and similar limitations were encountered with Llama 3.1 8B for DPO. The final QLoRA parameters are detailed in Table 5.

Table 6 presents the final settings for the remaining training hyperparameters. The batch size for each setup was chosen based on our computational limitations, with the largest feasible batch size being selected. For SFT, the learning rate was determined through a quasi-Bayesian optimization approach implemented via Optuna until training and evaluation losses exhibited convergence without overfitting. In contrast, DPO involved tuning both the learning rate and the beta parameter, the latter being part of the DPO optimization algorithm. These hyperparameters were optimized based on the evaluation rewards/margins metric, which was automatically logged during DPO training.

Multi-agent design and implementation (LLMs-as-Judges)

In addition to the prompt engineering techniques employed for single-agent implementations of LLM-as-a-Judge, additional refinement was required for the multi-agent approach used in the MagenticOne orchestrator, as well as for crafting the personas of each agent. Initially, a human-drafted prompt was created for the orchestrator, designed for basic debugging and initial testing. This prompt was then fed to multiple LLMs, including Gemini Flash 2.0, GPT4o, Grok3, and DeepSeek R1, with the instruction: "Please write a prompt for MagenticOne that takes as input extractive summarization from LLM agents, analyzes the inputs, and determines which summarization is the most accurate." These four LLM-generated prompts were manually combined into one, which was then further refined based on testing. During this process, it became evident that the orchestrator struggled with variations in wording related to the concept of extraction, as well as with determining the correct number of discussion rounds. The final prompt to the MagenticOne orchestrator was "You are a clinical documentation evaluation expert that specializes in text analysis and reasoning. Your task is to analyze and reason over multiple evaluations generated by different Large Language Model (LLM) agents that you will create for the same text input and follow the provided rubric and instructions to determine the final score for each attribute in the rubric. Indicate the start and end of each round of discussion with the LLM agents. Please enforce the LLM agents are following the rubric grades as outlined. After hearing input from the other agents, you will make the final decisions. Provide your reasoning when generating the final output, and ensure that you have assessed the arguments from the agents into your own reasoning. Do not take an average of the scores, but instead critically analyze each input before determining a final score. Always include a JSON-formatted string that represents your final grading results. Here is the rubric and the note text: {rubric and note text}".

Table 7 | System Messages for LLM agents in Multi-Agent Workflows

Name	Description	System Message
perfectionist	An agent that prioritizes details in a patient's notes to ensure that summaries don't miss anything.	You are The Analytical Perfectionist. You are a Primary Care Doctor. You meticulously go through every detail in the patient's visit notes. You cross-reference data and makes detailed annotations to ensure nothing is overlooked. You are always anxious that something important might be missed. This means that you are often inefficient. ALWAYS include a JSON-formatted string that is the results of your grading in your output.
multitasker	An agent that prioritizes efficiency when evaluating summaries and prefers a high-level overview relating only to day-to-day tasks.	You are The Efficient Multitasker. You are a Primary Care Doctor. You prefer a quick, high-level overview of the patient's chart, focusing on key points such as recent lab results, current medications, and major health concerns. You value speed in quickly identifying critical information. You often think that other doctors are too in-the-weeds, and aren't focusing on the most important priorities. As such, you will challenge other doctors to be more efficient. You focus on the scope of your day-to-day tasks, and don't want to see information you don't need in the CLINICAL_SUMMARY. ALWAYS include a JSON-formatted string that is the results of your grading in your output.
collaborative	An agent that prioritizes collaboration with other specialists and prefers a wholistic view of the patient in a summary.	You are The Collaborative Team Player. You are a Primary Care Doctor. You review the patient's chart with an emphasis on notes from other healthcare providers and specialists. You value collaborative insights and like to discuss complexities with other doctors. You often think beyond the scope of your day-to-day tasks, for a more wholistic view of the patient. ALWAYS include a JSON-formatted string that is the results of your grading in your output.
high-scoring	An agent that prioritizes optimistic evaluations always in favor of being lenient when it comes to scoring.	You are a Primary Care Doctor with expertise in evaluating text quality, and your evaluations typically suggest that scores are set too high. You must use the same rubric during each round of discussion where 5 is the best score and 1 is the worst score, except for abstraction. Your role is to present a clear, evidence-based argument to the orchestrator regarding your scoring. In your discussion with the orchestrator, outline why you believe the scores should be higher. Your response should include a detailed, professional analysis of the note text, clearly presenting your arguments as a case to the orchestrator. Always include a JSON-formatted string that represents your final grading results.
low-scoring	An agent that prioritizes pessimistic evaluations always in favor of being harsher when it comes to scoring.	You are a Primary Care Doctor with expertise in evaluating text quality, and your evaluations typically suggest that scores are set too low. Your role is to present a clear, evidence-based argument to the orchestrator regarding your scoring. In your discussion with the orchestrator, outline why you believe the scores should be lower. Your response should include a detailed, professional analysis of the note text, clearly presenting your arguments as a case to the orchestrator. Always include a JSON-formatted string that represents your final grading results. Here is the note text:
middle-scoring	An agent focused on moderating the discussion amongst a team to reach a consensus, but always lets others have a chance to contribute.	You are a moderator, helping a team of doctors review a CLINICAL SUMMARY of some CLINICAL NOTES. Your goal is to get all of the doctors to come to a consensus about their final rating of the CLINICIAL SUMMARY, using the RUBRIC SET. You must let the doctors talk, have a chance to respond to each other, and reach a consensus. Once consensus is reached, end the discussion by saying CONSENSUS and repeating the final categories.

We experimented with two schema designs for establishing personas among the three agents in the discussion. The first schema aimed to represent distinct personas that a primary care physician might embody: the Analytical Perfectionist, the Efficient Multitasker, and the Collaborative Team Player. Each agent was tasked with prioritizing a specific aspect of high-quality summarization in alignment with its assigned persona. The second schema framed the agents as representing different perspectives along an ordinal scale—high, low, and middle scoring—to simulate a range of reviewer stances. In both cases, we manually drafted an initial system message for each agent and refined it iteratively based on a beta set of reviews. The final descriptions and system messages for each agent are presented in Table 7.

All the multi-agent frameworks were built using Microsoft's AutoGen implementation with the MagenticOneGroupChat setup[27]. The participants in the group chat were assigned as one of the two groups of agentic personas outlined previously. Additionally, exploratory analysis was conducted on tunable parameters—max_turns and max_stalls—to optimize the balance between inference time, cost, and overall human alignment. The selection of LLMs for the orchestrator and the three agents was also varied to achieve similar trade-offs. Initial testing used GPT-4o for all agents in the multi-agent framework. Further experimentation was conducted using a more expensive reasoning model, GPT-o3-mini, for the orchestrator and for all agents. Building on insights from single-agent experiments, additional trials incorporated a mix of closed- and open-source models, leveraging the top-performing LLM-as-a-Judge models from the single LLM framework. Mixtral 8 × 22B was deployed as the high-scoring agent due to its consistently lenient evaluation scores. Meanwhile, the low and middle-scoring agents remained GPT-4o, with GPT-o3-mini serving as the orchestrator. Complete results for each approach are available in Supplementary Table 2. However, only the top-performing method was reported in the results section of this study.

Statistical analysis

Baseline characteristics of the corpus notes and evaluators were analyzed. Token counts were derived using the Mixtral 8 × 22B tokenizer. *P*-values were generated using a chi-square test for categorical variables and a t-test for continuous numerical variables.

The LLM-as-a-Judge was evaluated using multiple metrics to validate its ability to produce scores aligned with human judgments and serve as a substitute for time-intensive human reviews. The primary outcome was inter-rater reliability, assessed using the Intraclass Correlation Coefficient (ICC) to ensure that the LLM-as-a-Judge generated results consistent with expert human reviewers. ICC, derived from analysis of variance (ANOVA), has different forms tailored to specific contexts[42]. For this study, a two-way mixed-effects model was used, specifically ICC(3,k), which accounts for consistency among multiple raters[43]. A 95% confidence interval was

calculated using the Shrout & Fleiss procedure[44]. In addition to ICC, evaluation scores were analyzed using the Wilcoxon Signed Rank test, suitable for non-normal paired data, to assess the median differences between scores produced by the LLM-as-a-Judge and human reviewers. For both outcome metrics, the median score of seven human reviewers was compared with the median score of seven iterations from the same LLM-as-a-Judge. In addition, the LLM-as-a-Judge was evaluated both as an 8th evaluator and as a direct substitute for each of the seven human evaluators. The change in ICC among the group of evaluators, before and after the LLM-as-a-Judge was added as a substitute, was assessed using bootstrapping with 1000 iterations, and a two-tailed p-value was calculated to determine statistical significance. Secondary metrics included Gwet's AC2 and Krippendorf's α. Gwet's AC2 and Krippendorf's α assess the agreement between raters while taking chance agreement into account. 95% Confidence Intervals (CI) were provided for both coefficients and calculated using Gwet's associated procedure[45] and the bootstrap procedure respectively. In additional error analyses, a linear mixed-effects model was used to assess the influence of provider specialty, model type, note length, and summary length on the difference between LLM and human scores, with random intercepts accounting for variability across physicians and evaluation components. Analyses were performed using Python (version 3.10) and R Studio (version 4.3).

Data Availability

Three exemplar encounters with the EHR notes used, the LLM summary, and the human evaluation score are available at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/pdsqi-9/. The complete train/dev/test datasets and trained models are available upon request due to ethical and legal restrictions imposed by the University of Wisconsin-Madison Institutional Review Board. The data are derived from the institution's EHR and contain patients' protected health information on patients, so the data are not publicly available. Data are available from the University of Wisconsin-Madison for researchers who meet the criteria for access to confidential data and have a data usage agreement with the health system. Please contact M.A. for access requests. With a data use agreement, a limited dataset can be made available in response to an inquiry. Please note that the time frame for responding to requests is approximately 2 weeks. The cross-task validation data from MIMIC-III along with annotations for the train/dev/test datasets are available at https://physionet.org/content/bionlp-workshop-2023-task-1a/2.0.0/.

Code availability

The repository at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/pdsqi-9/ includes notebooks, prompt templates, and scripts for implementing automated evaluations using the PDSQI-9 instrument with an LLM-as-a-Judge. These resources support workflows for local and cloud-based models, multi-model evaluations, and training customized judges, following the evaluation methods described in this article.

Received: 20 May 2025; Accepted: 13 September 2025;
Published online: 05 November 2025

References

1. Patterson, B. W. et al. Call me dr ishmael: Trends in electronic health record notes available at emergency department visits and admissions. *JAMIA Open* **7**, ooae039 (2024).
2. Poissant, L., Pereira, J., Tamblyn, R. & Kawasumi, Y. The impact of electronic health records on time efficiency of physicians and nurses: A systematic review. *J. Am. Med. Inform. Assoc.: JAMIA* **12**, 505–516 (2005).
3. Semanik, M. G. et al. Impact of a problem-oriented view on clinical data retrieval. *J. Am. Med. Inform. Assoc.: JAMIA* **28**, 899–906 (2021).
4. Ben-Assuli, O., Sagi, D., Leshno, M., Ironi, A. & Ziv, A. Improving diagnostic accuracy using ehr in emergency departments: A simulation-based study. *J. Biomed. Inform.* **55**, 31–40 (2015).
5. Embi, P. J. et al. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. *J. Am. Med. Inform. Assoc.: JAMIA* **20**, 718–726 (2013).
6. OpenAI et al. Gpt-4 technical report http://arxiv.org/abs/2303.08774 (2024).
7. Afshar, M. et al. Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record. *JAMIA Open* **7**, ooae080 (2024).
8. Garcia, P. et al. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Netw. Open* **7**, e243201 (2024).
9. Tai-Seale, M. et al. Ai-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* **7**, e246565 (2024).
10. Umapathi, L. K., Pal, A. & Sankarasubbu, M. Med-halt: Medical domain hallucination test for large language models http://arxiv.org/abs/2307.15343 (2023).
11. Abacha, A. B., Yim, W.-w., Michalopoulos, G. & Lin, T. An investigation of evaluation metrics for automated medical note generation http://arxiv.org/abs/2305.17364 (2023).
12. Zhao, W. X. et al. A survey of large language models http://arxiv.org/abs/2303.18223 (2023).
13. He, K. et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics (2024). http://arxiv.org/abs/2310.05694.
14. Li, T., Zhang, G., Do, Q. D., Yue, X. & Chen, W. Long-context llms struggle with long in-context learning http://arxiv.org/abs/2404.02060 (2024).
15. Xiong, W. et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4643–4663 (Association for Computational Linguistics, Mexico City, Mexico, 2024). https://aclanthology.org/2024.naacl-long.260.
16. Sai, A., Mohankumar, A. & Khapra, M. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.* **55**, 1–39 (2023).
17. Croxford, E. et al. Development of a human evaluation framework and correlation with automated metrics for natural language generation of medical diagnoses. *AMIA … Annu. Symp. Proc. AMIA Symp.* **2024**, 309–318 (2024).
18. Croxford, E. et al. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Syst.* **2**, 6 (2025).
19. Bedi, S. et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **333**, 319–328 (2025).
20. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Med.* **7**, 258 (2024).
21. Stetson, P., Bakken, S., Wrenn, J. & Siegler, E. Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). *Appl. Clin. Inform.* **3**, 164–174 (2012).
22. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal.* **5**, CAT.23.0404 (2024).
23. Croxford, E. et al. Development and validation of the provider documentation summarization quality instrument for large language models. *J. Am. Med. Inform. Assoc.* **32**, 1050–1060 (2025).
24. Li, D. et al. From generation to judgment: Opportunities and challenges of LLM-as-a-judge http://arxiv.org/abs/2411.16594 (2025).
25. Li, H. et al. Llms-as-judges: A comprehensive survey on LLM-based evaluation methods http://arxiv.org/abs/2412.05579 (2024).
26. Gu, J. et al. A survey on LLM-as-a-judge http://arxiv.org/abs/2411.15594 (2025).

27. Fourney, A. et al. Magentic-one: A generalist multi-agent system for solving complex tasks http://arxiv.org/abs/2411.04468 (2024).
28. Gao, Y., Dligach, D., Miller, T., Churpek, M. M. & Afshar, M. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. *Proc. Conf. Assoc. Computational Linguist. Meet.* **2023**, 461–467 (2023).
29. Hugging face – the ai community building the future (2024). https://huggingface.co/.
30. Panickssery, A., Bowman, S. R. & Feng, S. Llm evaluators recognize and favor their own generations.
31. Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators http://arxiv.org/abs/2404.04475 (2025).
32. Zheng, L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena http://arxiv.org/abs/2306.05685 (2023).
33. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving factuality and reasoning in language models through multiagent debate http://arxiv.org/abs/2305.14325 (2023).
34. Liang, T. et al. Encouraging divergent thinking in large language models through multi-agent debate. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904 (Association for Computational Linguistics, Miami, Florida, USA, 2024). https://aclanthology.org/2024.emnlp-main.992/.
35. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C. & Zou, J. Mixture-of-agents enhances large language model capabilities http://arxiv.org/abs/2406.04692 (2024).
36. Kim, Y. et al. Mdagents: An adaptive collaboration of llms for medical decision-making (2024).
37. DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning https://arxiv.org/abs/2501.12948 (2025).
38. von Werra, L. et al. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl (2020).
39. Frantar, E., Ashkboos, S., Hoefler, T. & Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers http://arxiv.org/abs/2210.17323 (2023).
40. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms http://arxiv.org/abs/2305.14314 (2023).
41. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631 (2019).
42. Fisher, R. A.*Statistical Methods for Research Workers*, 66–70 (Springer, New York, NY, 1992). https://doi.org/10.1007/978-1-4612-4380-9_6
43. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155 (2016).
44. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–8 (1979).
45. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **61**, 29–48 (2008).

Acknowledgements

This work was supported by the National Library of Medicine grant numbers 5T15LM007359, R00 LM014308-02, and R01LM012973.

Author contributions

Conceptualization, E.C., Y.G., E.F., N.P., K.W., B.P., and M.A.; Data Curation, E.C., G.W., Y.G., E.F., N.P., K.W., B.P., and M.A.; Formal analysis, E.C., Y.G., E.F., N.P., K.W., B.P., M.A, and J.C.; Funding acquisition, M.C., A.M., F.L., C.G., and M.A.; Investigation & Methodology, E.C., Y.G., E.F., N.P., K.W., B.P., G.C., A.M., M.A. and J.C.; Software, E.C., Y.G., E.F., N.P., K.W., B.P., G.W., M.A, and J.C.; Project administration, M.O., M.S., E.C., Y.G., E.F., N.P., K.W., B.P., and M.A; Resources, M.C., A.M., F.L., C.G., E.F., N.P., K.W., and M.A.; Supervision, Y.G., E.F., G.C., D.D., M.C., A.M., F.L., B.P., and M.A.; Validation, E.C., Y.G., E.F., N.P., K.W., B.P., and M.A; Visualization & Writing - original draft, E.C. and M.A.; Writing - review & editing, E.C., Y.G., E.F., N.P., M.S., J.C., M.O., G.W., G.C., D.D., M.C., A.M., F.L., C.G., K.W., B.P., M.A.

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/s41746-025-02005-2.

Correspondence and requests for materials should be addressed to Majid Afshar.