scientific reports

OPEN

Analyzing patient perspectives with large language models: a cross-sectional study of sentiment and thematic classification on exception from informed consent

Aaron E. Kornblith[1], Chandan Singh[1,2], Johanna C. Innes[3], Todd P. Chang[4], Kathleen M. Adelgais[5], Maija Holsti[6], Joy Kim[7], Bradford McClain[8], Daniel K. Nishijima[9], Steffanie Rodgers[10], Manish I. Shah[11], Harold K. Simon[12], John M. VanBuren[13], Caleb E. Ward[14] & Catherine R. Counts[15]

Large language models (LLMs) can improve text analysis efficiency in healthcare. This study explores the application of LLMs to analyze patient perspectives within the exception from informed consent (EFIC) process, which waives consent in emergency research. Our objective is to assess whether LLMs can analyze patient perspectives in EFIC interviews with performance comparable to human reviewers. We analyzed 102 EFIC community interviews from 9 sites, each with 46 questions, as part of the Pediatric Dose Optimization for Seizures in Emergency Medical Services study. We evaluated 5 LLMs, including GPT-4, to assess sentiment polarity on a 5-point scale and classify responses into predefined thematic classes. Three human reviewers conducted parallel analyses, with agreement measured by Cohen's Kappa and classification accuracy. Polarity scores between LLM and human reviewers showed substantial agreement (Cohen's kappa: 0.69, 95% CI 0.61–0.76), with major discrepancies in only 4.7% of responses. LLM achieved high thematic classification accuracy (0.868, 95% CI 0.853–0.881), comparable to inter-rater agreement among human reviewers (0.867, 95% CI 0.836–0.901). LLMs enabled large-scale visual analysis, comparing response statistics across sites, questions, and classes. LLMs efficiently analyzed patient perspectives in EFIC interviews, showing substantial sentiment assessment and thematic classification performance. However, occasional underperformance suggests LLMs should complement, not replace, human judgment. Future work should evaluate LLM integration in EFIC to enhance efficiency, reduce subjectivity, and support accurate patient perspective analysis.

Keywords Large language models, Sentiment analysis, Emergency medical services, Pediatrics, Research ethics

Abbreviations

API	Application Programming Interface
EFIC	Exception From Informed Consent
EMS	Emergency Medical Services

[1]University of California San Francisco, San Francisco, CA, USA. [2]Microsoft Research, Seattle, WA, USA. [3]Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA. [4]Keck School of Medicine of University of Southern California & Children's Hospital Los Angeles, Los Angeles, CA, USA. [5]University of Colorado School of Medicine, Aurora, CO, USA. [6]Primary Children's Medical Center, University of Utah, Salt Lake City, UT, USA. [7]Oregon Health & Science University, Portland, OR, USA. [8]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [9]UC Davis School of Medicine, Sacramento, CA, USA. [10]Nationwide Children's Hospital, Columbus, OH, USA. [11]Stanford University School of Medicine, Palo Alto, CA, USA. [12]Emory University School of Medicine & Children's Healthcare of Atlanta, Atlanta, GA, USA. [13]University of Utah, Salt Lake City, UT, USA. [14]Children's National Hospital, The George Washington University, Washington, D.C, USA. [15]University of Washington, Seattle, WA, USA. ✉email: Aaron.Kornblith@ucsf.edu

| https://doi.org/10.1038/s41598-025-89996-w

FDA	Food and Drug Administration
IRB	Institutional Review Board
LLMs	Large Language Models
PediDOSE	Pediatric Dose Optimization for Seizures in Emergency Medical Services
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology

Background & significance

Large language models (LLMs) have emerged as powerful tools for generating, summarizing, and analyzing complex text with natural language understanding[1–3]. Recently, LLMs have been applied in healthcare, enhancing tools such as diagnostics, decision-making support, and predictive analytics[4]. While these applications have garnered considerable interest, the challenge of using LLMs to summarize and understand patient perspectives has received less attention despite its critical importance[5,6]. This involves analyzing and aggregating large volumes of patient narratives, a task for which LLMs are well-suited.

We explore how LLMs can be applied to understand patient perspectives in healthcare, using the exception from informed consent (EFIC) process as a use case. The EFIC process allows the enrollment of subjects in clinical trials when consent isn't feasible, e.g., for administering an automated external defibrillator to unconscious patients[7]. Instead of obtaining consent directly from the patient, EFIC requires community-wide interviews before starting a trial[9]. Analyzing these interviews presents a major challenge, requiring extensive manual labor and subjective judgment with ambiguous standardization[8–11].

Objective

This challenge presents an opportunity for LLMs to efficiently analyze and interpret patient perspectives, addressing issues of scale and consistency. In this study, we hypothesized that LLMs would match human reviewers in determining patient perspectives, including sentiment (e.g., very positive to very negative), and in classifying responses from EFIC interviews into themes. We tested this hypothesis using LLMs to analyze sentiment polarity and classify interviews from the Pediatric Dose Optimization for Seizures in Emergency Medical Services (PediDOSE) study, comparing their performance with human reviewers.

Results

LLM polarity score analysis

We first used GPT-4 to assign polarity scores to interview responses across the study questions (Fig. 1**).** Of 3,692 responses available for analysis, 1,000 were coded "no response." GPT-4 classified 2.8% ($n = 104$) of all site responses as very negative, 13.1% ($n = 482$) as negative, 32.7% ($n = 1207$) as neutral, 32.3% ($n = 1191$) as positive, and 19.2% ($n = 708$) as very positive. Additionally, visualizing sentiment polarity across sites generally revealed consistent trends in responses, while also highlighting specific questions that differed between sites. For example, Question 32 on seizure awareness elicited more negative responses from Site A compared to the other two sites.

To compare LLM-generated judgments to human judgments, 5 LLMs and 3 human reviewers assigned polarity scores to 123 responses to 9 questions from 3 PediDOSE sites (Fig. 2**).** The mean human reviewer polarity score substantially agreed with GPT-4, the highest-performing LLM of the five LLMs evaluated, yielding a Cohen's kappa k = 0.69 (95% CI 0.61–0.76), Fig. 2. In comparison, individual human reviewer polarity scores yielded a slightly higher agreement among the 3 reviewers, ranging from k = 0.78–0.92. Mistral (7B) k = 0.63, GPT-3.5 Turbo k = 0.65, and GPT-4 k = 0.69 substantially agreed with mean human reviewer polarity score. The highest performing LLMs (GPT-4, GPT 3.5 turbo, Mistral) also had the most substantial agreement with human reviewer 2. LLAMA 2 (7B), k = 0.31 and LLAMA (70B) k = 0.44 had the lowest agreement with the mean human reviewer polarity score. The lowest performing LLMs (LLAMA (7B) and (70B)) had very poor agreement with each other, k = 0.19.

Major discrepancies, in which the GPT-4 and the human reviewer assigned opposite polarity scores (e.g. positive vs. negative), were seen in 4.7% of all scores. Human reviewers were 62% less likely to assign extreme values (very positive and very negative) than GPT-4. Most questions that generated a positive polarity score from LLM were also scored as positive by the human reviewers. For example, LLM and human reviewers scored this question as positive, Question 1, "How important do you think it is to do this study in your community?" **(Table A1**, **Supplementary File 1)**. Similarly, questions that yielded a negative polarity score from LLM we also scored as negative by the human reviewers. Most negative polarity score questions were about the background or are phrased in a manner that they expect negative responses, e.g., Question 46 "Do you have any remaining questions about research or informed consent?" A common answer is "nothing else".

Thematic classification

Up to 15 responses for each question were randomly selected from three study sites. In our text classification analysis of GPT-4, we collected 188 responses from 22 free-text questions sorted into classes by each human reviewer; an example is shown in Fig. 3. On average, GPT-4 generated 3.24 classes per question. Human reviewers categorized responses into the same classes as GPT-4 86.8% (95% CI 86.3–87.3%) of the time, suggesting that thematic classification by GPT-4 is similar to that by human reviewers (Table 1). Inter-reviewer thematic classification accuracy was 86.7% (95% CI 85.3–88.1%).

Thematic classification accuracy of response classes by human review compared to GPT-4, inter-reviewer agreement (% accuracy ± 95% confidence interval).

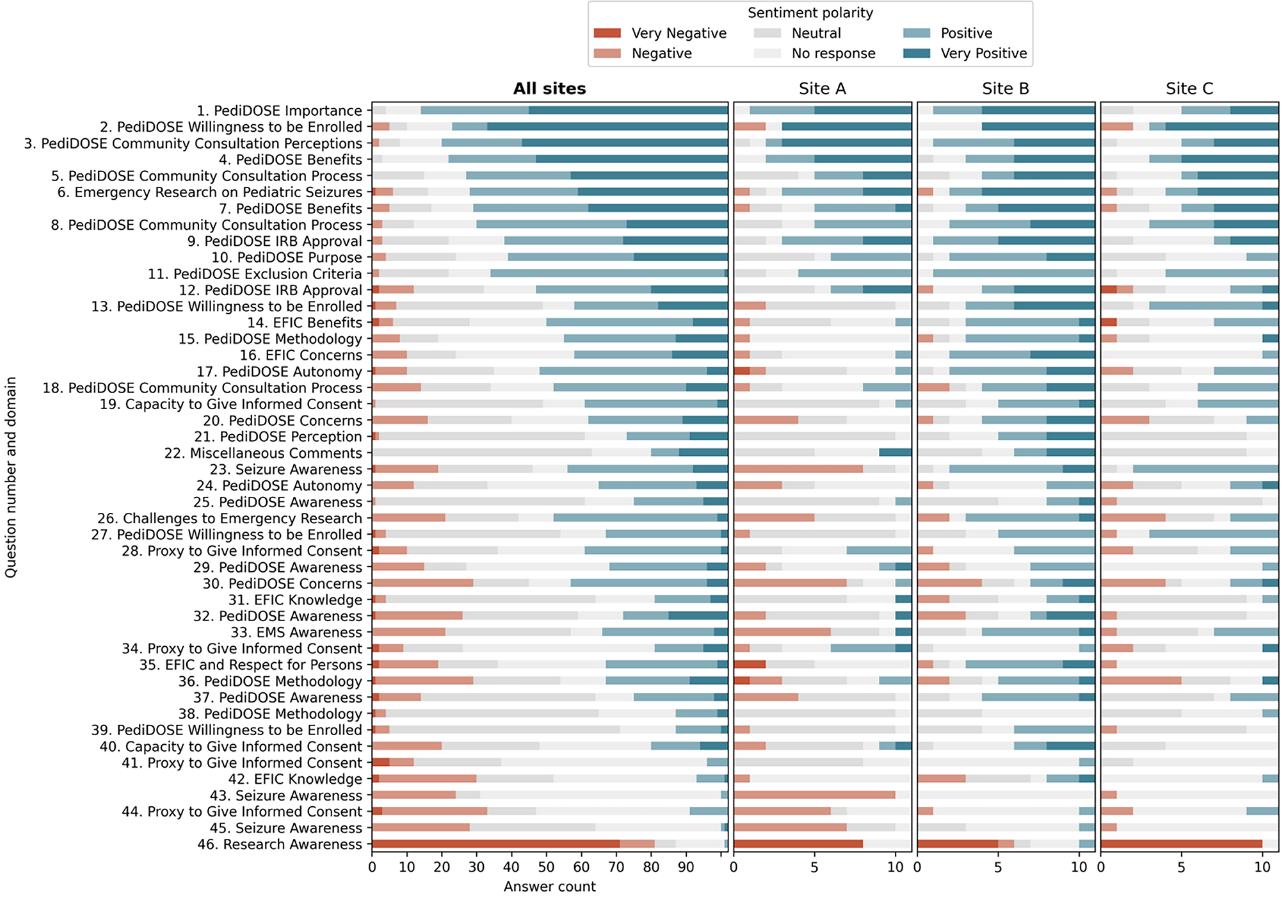


Fig. 1. Large language models (LLM) polarity score across participating sites ($n = 3{,}692$ responses). Each row represents the mean LLM response polarity score to each question, color indicates assigned polarity. Each question is represented by question number and representative domain. *All sites* represent the mean LLM response polarity score from all nine sites, while the subsequent three columns provide a breakdown of the response polarity score at three sites, selected for detailed comparison to simplify review.

Discussion

Our study demonstrates that LLMs assigned sentiment polarity scores and classified responses from EFIC community interviews for PediDOSE nearly as well as human reviewers. However, LLMs had less agreement in assessing polarity scores compared to human reviewers, highlighting the need for LLMs to complement, rather than replace, human oversight. By combining human and LLM evaluations, we achieve a more comprehensive quantitative analysis, which enhances our understanding of patient sentiment. LLMs can also provide clear, concise class-based summaries, enabling investigators, Institutional Review Board (IRB), and ethics boards to rapidly assess overall responses. Sentiment polarity scores offer a quantitative assessment of community sentiment, helping stakeholders understand not only overall trends but also variations in participant responses to individual questions, which can inform targeted improvements in communication and study design. Quantitative plots generated from LLM data, such as those in Fig. 1, facilitate the rapid analysis of large datasets, enabling visual interpretation of patterns, including variations and outliers. Our study showed that responses to the same question across 3 sites were generally consistent, with few exceptions. These plots helped identify trends and outliers between sites. For example, responses to Question 23 (personal experiences with seizures) at Site A showed significantly more negative sentiment compared to the other two sites. Such insights help understand patient sentiment nuances, enhancing overall analysis.

LLMs were more likely than human reviewers to assign extreme polarity values. One explanation is that LLMs may struggle to recognize the subtleties of human language. Another possibility is that the model's performance metric, designed to reward higher confidence, could bias polarity scoring toward extremes. Additionally, LLMs may struggle to interpret emotional context, which can result in more extreme classifications based on the text's surface content rather than its deeper emotional nuances[12]. This underscores the need for sentiment polarity assessments to be considered alongside thematic classification for more accurate interpretation.

Our thematic classification process improved our understanding of community feedback and may assist Food and Drug Administration (FDA) EFIC procedures related to study protocol approvals, adjustments, or disapprovals. Classifying responses into distinct classes provides researchers and regulatory bodies with deeper insights into community sentiments, facilitating more comprehensive evaluations of clinical studies. Thematic

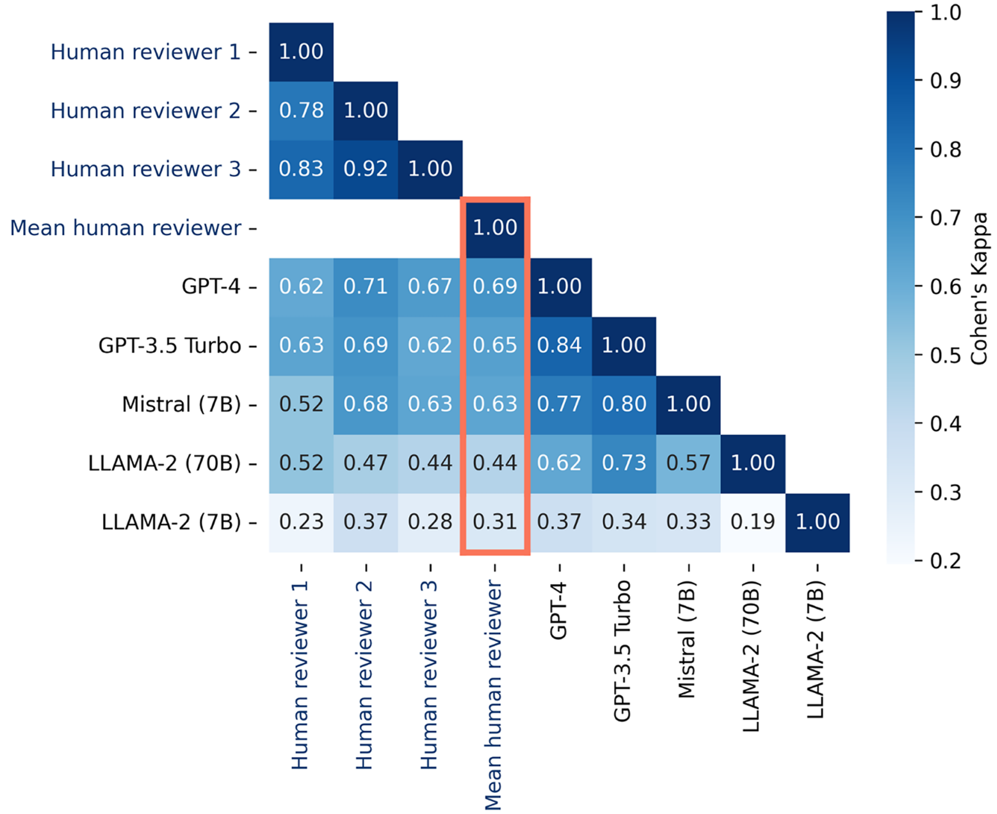
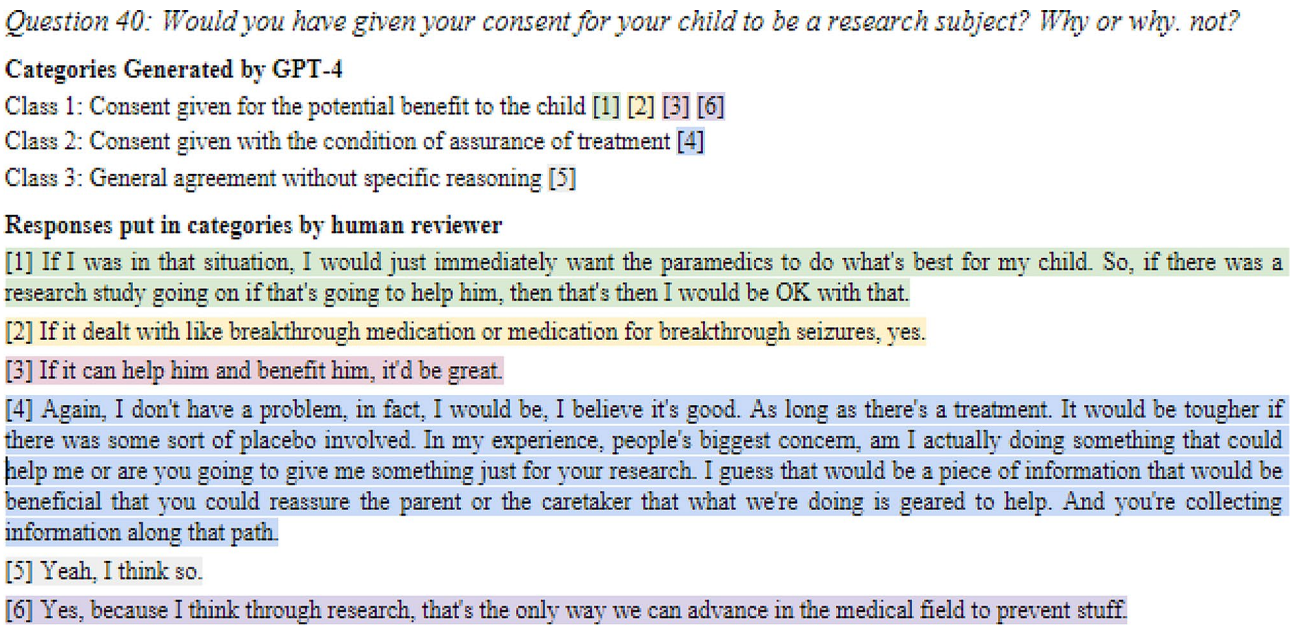


Fig. 2. Polarity score agreement between mean, individual human reviewers, and large language models (LLM) across human reviewed sample ($n = 123$ responses). Using Cohen's Kappa, this heatmap shows agreement of response polarity scores between human reviewers and LLM classes.



Question 40: Would you have given your consent for your child to be a research subject? Why or why. not?

Categories Generated by GPT-4

Class 1: Consent given for the potential benefit to the child [1] [2] [3] [6]

Class 2: Consent given with the condition of assurance of treatment [4]

Class 3: General agreement without specific reasoning [5]

Responses put in categories by human reviewer

[1] If I was in that situation, I would just immediately want the paramedics to do what's best for my child. So, if there was a research study going on if that's going to help him, then that's then I would be OK with that.

[2] If it dealt with like breakthrough medication or medication for breakthrough seizures, yes.

[3] If it can help him and benefit him, it'd be great.

[4] Again, I don't have a problem, in fact, I would be, I believe it's good. As long as there's a treatment. It would be tougher if there was some sort of placebo involved. In my experience, people's biggest concern, am I actually doing something that could help me or are you going to give me something just for your research. I guess that would be a piece of information that would be beneficial that you could reassure the parent or the caretaker that what we're doing is geared to help. And you're collecting information along that path.

[5] Yeah, I think so.

[6] Yes, because I think through research, that's the only way we can advance in the medical field to prevent stuff.

Fig. 3. Thematic classification by human reviewers. Human reviewers classified the responses to Question 40 into GPT-4 generated classes.

Human Reviewer-GPT-4 Accuracy **n = 188 responses**			**Human Reviewer-Human Reviewer Accuracy**
Reviewer 1	Reviewer 2	Reviewer 3	Inter-reviewer agreement
86.1 ± 4.9%	86.7 ± 4.9%	87.7 ± 4.7%	86.7% ±2.7%

Table 1. Text classification accuracy.

classification has been applied across various fields, including healthcare. For instance, non-generative LLMs have classified patients into health categories from discharge summaries[13]. In our study, Question 40 highlighted a range of community opinions about children's participation in research, emphasizing the requisite for clear and transparent communication about the EFIC process to future participants.

Classifying large volumes of lengthy responses is a complex, time- and resource-intensive process. In our study, LLMs performed similarly to human reviewers, offering valuable support in enhancing the efficiency of thematic classification, particularly with large datasets. Our findings demonstrate that LLMs can assist in assigning responses to predefined thematic classes, a key step in the EFIC process that traditionally requires significant human effort. This capability enables investigators and ethics boards to rapidly summarize and interpret feedback, potentially streamlining future community consultation workflows and improving the consistency of data interpretation. Through prompt engineering, LLMs generate classes and organize responses under these classes with corresponding citations. This method helps indicate which responses align with each class. LLMs greatly facilitated the work of human reviewers, enabling quick assessment of class validity, while reviewers can easily verify the accuracy of classifications by referencing source responses.

LLMs did not match humans in analyzing patient perspectives, but the field is advancing rapidly. Improvements in modeling[14,15], prompting[16,17], and interpretation[18]will enhance sentiment analysis. "Hallucinations," or fabricated information generated by LLMs, remain a challenge[19,20]. Investigators are exploring innovative methods for LLMs to detect these issues, such as self-verification[19]. Similarly, as LLMs' context windows grow, they'll handle larger datasets, improving classification of lengthy interviews.

Further studies are needed to assess whether human reviewers can adjust and refine LLM-generated classifications to align with nuanced human judgments. This approach leverages the strengths of both LLMs and human oversight for more reliable and accurate text classification. One study noted that analyzing survey data for the FDA is challenging due to a lack of standardized formats and raw data[21]. Our findings suggest that using LLMs can improve the efficiency of analyzing community consultation data for EFIC trials and help standardize reporting, building trust for future studies. Although this study did not measure workflow metrics, such as the time required to create summary reports or IRB review turnaround times, the visual representation of quantified sentiment and thematic classifications illustrates how LLM-generated outputs can streamline regulatory processes. For example, LLM-generated sentiment polarity plots could help IRBs and ethics committees quickly identify trends, such as a notable percentage of participants raising concerns about consent processes or reporting confusion about study procedures. This approach reduces the cognitive burden of manually reviewing extensive unstructured data and enables more efficient, data-driven decision-making. Future studies should consider quantifying time savings and evaluating user feedback to further demonstrate the practical impact of LLMs in regulatory workflows.

Limitations

Our study has several limitations. First, we limited our analysis to a manageable number of human reviews. Manually reviewing and annotating interviews is labor-intensive, underscoring the challenge of the EFIC process and the potential utility of LLMs in facilitating those tasks. Regardless, we conducted ample human annotations to reach reliability and meaningful correlation. Second, we limited our human reviewers to investigators. Future studies could include other EFIC team members or other stakeholders, such as IRB members, to review and modify the data output to suit their needs. Third, although LLMs do not need to analyze yes/no responses, we included all question types for sentiment polarity assessment to demonstrate the full range of LLM capabilities and to avoid introducing bias through selective exclusion. While some questions may seem binary, participant responses often included elaborations that provided valuable sentiment insights. By including all questions, we ensured that the analysis captured these nuances and demonstrated the LLM's ability to handle a wide range of input formats. This approach allows future users to tailor their analyses to specific research needs while maintaining transparency in how LLMs can be applied. Fourth, our study design limited our ability to investigate how demographics such as race, ethnicity, language, and other patient-level experiences influence the responses to the interview questions. Fifth, presenting the survey questions in a set order may have introduced a bias based on sequence. However, this approach reflects how surveys are done in the real world, offering an authentic portrayal of EFIC activities. Similarly, to mitigate bias and ensure that each response from the LLM was generated solely based on the text of that response, without undue influence from prior responses, the interface was systematically reset after each query. Sixth, another limitation of our study is the absence of demographic data linked to individual interview responses, which prevented a detailed analysis of LLM performance across different population subgroups. Demographic factors such as language proficiency, cultural norms, and socio-economic background may influence response patterns and LLM classifications. Future studies should consider integrating demographic information to explore whether LLMs perform differently across subgroups and to detect potential biases in sentiment classification. Addressing these biases is essential to ensuring that LLM-based approaches in healthcare research remain equitable and reflective of diverse community perspectives.

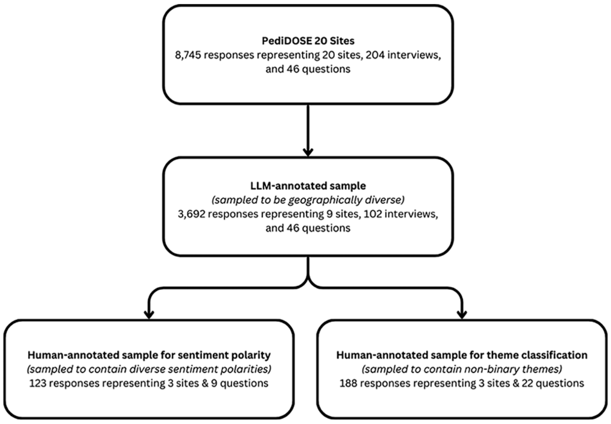

PediDOSE 20 Sites
8,745 responses representing 20 sites, 204 interviews, and 46 questions
LLM-annotated sample
(sampled to be geographically diverse)
3,692 responses representing 9 sites, 102 interviews, and 46 questions
Human-annotated sample for sentiment polarity
(sampled to contain diverse sentiment polarities)
123 responses representing 3 sites & 9 questions
Human-annotated sample for theme classification
(sampled to contain non-binary themes)
188 responses representing 3 sites & 22 questions

Fig. 4. Sampling Data Flow of Interview Responses. A visual overview of the data flow, showing the original dataset of Exception From Informed Consent (EFIC) community interviews collected across 20 Pediatric Dose Optimization for Seizures in Emergency Medical Services (PediDOSE) study sites, the geographically diverse subset annotated by large language models (9 sites), and the final human-annotated samples used for sentiment polarity and thematic classification comparisons.

Conclusion

Overall, LLMs demonstrated substantial agreement with human reviewers in analyzing patient perspectives, including sentiment polarity and thematic classification, using EFIC community interviews. While LLM polarity scoring showed lower reliability compared to human reviewers, thematic classification performance closely matched human assessments. Our study emphasizes the importance of using LLMs to supplement, rather than replace, human oversight. LLMs provide an effective method for rapidly summarizing and visualizing large datasets. Future efforts should explore how stakeholders can leverage LLMs to deepen insights into patient perspectives across various healthcare settings.

Materials and methods

Setting

PediDOSE is a multicenter EFIC trial (NCT05121324) that seeks to decrease the number of children who continue to have seizures upon arrival in the emergency department. Specifically, it evaluates the effectiveness of a standardized emergency medical services (EMS) protocol with age-based, paramedic-administered midazolam dosing[22].

Following the EFIC process, PediDOSE sites conducted EFIC community interviews to obtain IRB approval. Specifically, 10–14 community consultation interviews were conducted across 20 centers using an interview guide composed of 46 questions **(Supplement Table A1**, **Supplementary File 1).**[23] These questions were grouped into 25 domains preselected by EFIC PediDOSE investigators (Fig. 1). Interviewers were allowed to skip questions that seemed irrelevant based on previous responses from the interviewee and to ask unscripted follow-up questions. Interviews were conducted in English or Spanish, recorded, translated into English by certified interpreters, and de-identified for local IRB and investigator approval. Each site archived and transcribed interviews, which were submitted to the central IRB for review. All PediDOSE EFIC interview data were collated for manual review and final EFIC approval.

Study design

We retrospectively reviewed interview transcripts from the EFIC community interviews in the PediDOSE study. This study was approved by the central IRB, the University of Utah, and all participating site IRBs. All methods were performed in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. Written informed consent was obtained from all interview participants, as applicable. Identifiable information was removed to ensure participant anonymity. All actions were part of the PediDOSE EFIC activities, except for the post-hoc analysis of the transcripts performed as part of this study. For our analysis, we selected nine geographically diverse PediDOSE study sites that were representative of the various regions and patient demographics under study (Fig. 4). The University of Utah approved a waiver of written informed consent to collect and analyze the interview recordings; no identifiable information was requested during the recorded and transcribed interview. The source code used for performing the analyses is available at https://github.com/csinva/pedidose-efic-analysis. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines and the JAMA Network Guidance for evaluating and reporting LLM research[24–26].

Analysis methods

We evaluated two closed-source LLMs: GPT-4 (gpt-4-0613, Open AI)[17] and GPT-3.5 (gpt-3.5-turbo-0613, Open AI)[27], which were both accessed securely through the Azure OpenAI Application Programming Interface (API). We additionally assessed three open-source models: Mistral 7B (Mistral AI), LLaMA-2-7B (Meta), and LLaMA-2-70B (Meta), which were run locally for secure HIPAA-compliant computing. To minimize randomness, each LLM query was run in an independent session with a sampling temperature of 0. Python version 3.11 (Python Software Foundation) was used for all statistical analyses.

LLM-based sentiment polarity analysis

We evaluated each LLM's ability to assess the sentiment polarity for all responses in the studied interviews. Specifically, an LLM was prompted to categorize the polarity of a response on a 5-point Likert scale (very positive, positive, neutral, negative, or very negative*)*. For example, in response to a question such as *Is this study important to you?*, answers can vary from *very positive* (indicating it is important) to *very negative* (suggesting it is not important). If a question was not asked or if the respondent did not respond directly, it was coded as *no response*.

LLM-based thematic classification

In our theme classification analysis, we used two-step prompts to categorize participant responses. The first prompt listed the 25 predefined classes relevant to the EFIC process (e.g., "concerns about consent," "clarity of study purpose") and asked GPT-4 to sort responses into one of these classes. The second prompt aggregated responses within each class to generate a comprehensive summary of participant feedback. The detailed prompt structure can be found in **Table A2**, **Supplementary File 1**.

Human reviewer assessments

We evaluated the LLM-based sentiment polarity analysis and the LLM-based thematic classification by comparing LLM assessments to those of 3 blinded human reviewers. Human reviewers were made to do the same two analyses as the LLM. To allocate human reviewer time effectively, we subsampled each analysis's responses. For sentiment polarity analysis, we selected 15 responses from 9 questions across 3 random sites, stratified to yield a uniform range of negative, neutral, and positive answers. For thematic classification, we excluded simple 'Yes' or 'No' responses, as they do not require LLM analysis and would not add complexity to the problem. In cases of disagreement among reviewers, the majority opinion was used for analysis.

Evaluation metrics

We hypothesized that there would be a substantial association between the LLMs' outputs and human reviewers. For sentiment polarity scores, we measure the association using Cohen's Kappa k. For thematic classification, we measure the association using classification accuracy. Then, we use the LLMs scores to generate quantitative plots that enable understanding sentiment polarity distributions across sites and questions.

Data availability

Partial datasets and data dictionaries for the parent investigation, Pediatric Dose Optimization for Seizures in Emergency Medical Services Study (PediDOSE), will be available from the Data Coordinating Center (DCC) of the Pediatric Emergency Care Applied Research Network (PECARN) in a de-identified format 3 years after the last participant enrollment (anticipated July 2029). However, the qualitative transcripts of this secondary analysis cannot be made available because they cannot be de-identified. Please contact Dr. Manish Shah, mshah5@stanford.edu.

Received: 27 November 2024; Accepted: 10 February 2025
Published online: 20 February 2025

References

1. Zhang, W. et al. Sentiment Analysis in the Era of Large Language Models: A Reality Check. Published online May 24, 2023. Accessed December 7, (2023). http://arxiv.org/abs/2305.15005
2. Zhong, Q. et al. Can ChatGPT Understand too? A comparative study on ChatGPT and Fine-tuned BERT. *Published Online March*. **2** https://doi.org/10.48550/arXiv.2302.10198 (2023).
3. Wang, Z. et al. Is ChatGPT a good sentiment Analyzer? A preliminary study. *Published Online April*. **9** https://doi.org/10.48550/arXiv.2304.04339 (2023).
4. Peng, C. et al. A study of generative large language model for medical research and healthcare. *Npj Digit. Med.* **6** (1), 210. https://doi.org/10.1038/s41746-023-00958-w (2023).
5. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large Language models in Medicine. *JAMA* **330** (9), 866–869. https://doi.org/10.1001/jama.2023.14217 (2023).
6. Christou, P. The Use of Artificial Intelligence (AI) in qualitative research for Theory Development. *Qual. Rep.* **28** (9), 2739–2755. https://doi.org/10.46743/2160-3715/2023.6536 (2023).
7. *Exception from Informed Consent Requirements for Emergency Research*. FDA; Accessed February 10, 2024. (2020). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/exception-informed-consent-requirements-emergency-research
8. Denecke, K. & Deng, Y. Sentiment analysis in medical settings: new opportunities and challenges. *Artif. Intell. Med.* **64** (1), 17–27. https://doi.org/10.1016/j.artmed.2015.03.006 (2015).
9. Chisolm-Straker, M. et al. Exception from informed consent: how IRB reviewers Assess Community Consultation and Public Disclosure. *AJOB Empir. Bioeth.* **12** (1), 24–32. https://doi.org/10.1080/23294515.2020.1818878 (2021).
10. Lanspa, M. J., Fan, E. & Morris, A. H. How should we apply the Wisdom of the crowd to clinical trials with exception from informed consent? *JAMA Netw. Open.* **2** (7), e197569. https://doi.org/10.1001/jamanetworkopen.2019.7569 (2019).
11. Johnson, A. R. et al. An Approach to reviewing local context for exception from informed consent trials using a single IRB. *Ethics Hum. Res.* **43** (6), 42–48. https://doi.org/10.1002/eahr.500109 (2021).
12. Shah, M. *Pediatric Dose Optimization for Seizures in EMS (PediDOSE)*. clinicaltrials.gov; Accessed December 31, 2023. (2023). https://clinicaltrials.gov/study/NCT05121324

13. Ward, C. E. et al. Public support for and concerns regarding pediatric dose optimization for seizures in emergency medical services: an exception from informed consent (EFIC) trial. *Acad. Emerg. Med. Off J. Soc. Acad. Emerg. Med. Published Online March*. **7** https://doi.org/10.1111/acem.14884 (2024).
14. von Elm, E. et al. The strengthening the reporting of Observational studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann. Intern. Med.* **147** (8), 573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010 (2007).
15. Perlis, R. H. & Fihn, S. D. Evaluating the application of large Language models in Clinical Research contexts. *JAMA Netw. Open*. **6** (10), e2335924. https://doi.org/10.1001/jamanetworkopen.2023.35924 (2023).
16. Flanagin, A. et al. Reporting Use of AI in Research and Scholarly Publication—JAMA Network Guidance. *JAMA Published Online March*. **7** https://doi.org/10.1001/jama.2024.3471 (2024).
17. OpenAI, Achiam, J. et al. GPT-4 Technical Report. Published online December 18, (2023). https://doi.org/10.48550/arXiv.2303.08774
18. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Published Online March*. **4** https://doi.org/10.48550/arXiv.2203.02155 (2022).
19. Gehrmann, S. et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS One*. **13** (2), e0192360. https://doi.org/10.1371/journal.pone.0192360 (2018).
20. Tu, T. et al. Towards conversational diagnostic AI. *Published Online January*. **10** https://doi.org/10.48550/arXiv.2401.05654 (2024).
21. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620** (7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2 (2023).
22. Morris, J. X. et al. Tree Prompting: efficient Task Adaptation without Fine-tuning. *Published Online Oct*. **21** https://doi.org/10.48550/arXiv.2310.14034 (2023).
23. Nori, H. et al. Can Generalist Foundation models outcompete special-purpose tuning? Case Study in Medicine. *Published Online November*. **27** https://doi.org/10.48550/arXiv.2311.16452 (2023).
24. Singh, C. et al. Rethinking interpretability in the era of large Language models. *Published Online January*. **30** https://doi.org/10.48550/arXiv.2402.01761 (2024).
25. Gero, Z. et al. Self-Verification improves few-shot clinical information extraction. *Published Online May*. **30** https://doi.org/10.48550/arXiv.2306.00024 (2023).
26. Pan, L. et al. Automatically correcting large Language models: surveying the landscape of diverse self-correction strategies. *Published Online August*. **29** https://doi.org/10.48550/arXiv.2308.03188 (2023).
27. Feldman, W. B. et al. Public Approval of Exception from Informed Consent in emergency clinical trials: a systematic review of Community Consultation surveys. *JAMA Netw. Open*. **2** (7), e197591. https://doi.org/10.1001/jamanetworkopen.2019.7591 (2019).

Acknowledgements

Pediatric Emergency Care Applied Research Network (PECARN) Pediatric Dose Optimization for Seizures in EMS (PediDOSE) Study Investigators (See Online Resource 1 for list of investigators)This work was supported in part by the Pediatric Emergency Care Applied Research Network (PECARN). PECARN is supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS), in the Maternal and Child Health Bureau (MCHB), under the Emergency Medical Services for Children (EMSC) program through the following cooperative agreements: EMSC Data Center (EDC)-University of Utah (UJ5MC30824), GLACiER-Nationwide Children's Hospital (U03MC28844), HOMERUN-Cincinnati Children's Hospital Medical Center (U03MC22684), PEMNEWS-Columbia University Medical Center (U03MC00007), PRIME-University of California at Davis Medical Center (U03MC00001), CHaMP node- State University of New York at Buffalo (U03MC33154), STELAR - Seattle Children's Hospital (U03MC33156), and SPARC node - Emory University School of Medicine (U03MC49671);

Author contributions

Concept and design: AK, CS, MS; Acquisition, analysis or interpretation of data: AK, CS, JI, TC, KA, MH, JK, BM, DN, SR, MS, HS, JV, CW, CC; Draft manuscript: AK, CS, CC; Statistical Analysis: CS, JMV; Supervision: AK, CS, MS, CC. All authors contributed to the analysis of results, reviewed this manuscript, and revised it critically for important intellectual content and approve of the version as submitted.

Funding

This work was supported in part by the Pediatric Emergency Care Applied Research Network (PECARN). PECARN is supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS), in the Maternal and Child Health Bureau (MCHB), under the Emergency Medical Services for Children (EMSC) program through the following cooperative agreements: EMSC Data Center (EDC)-University of Utah (UJ5MC30824), GLACiER-Nationwide Children's Hospital (U03MC28844), HOMERUN-Cincinnati Children's Hospital Medical Center (U03MC22684), PEMNEWS-Columbia University Medical Center (U03MC00007), PRIME-University of California at Davis Medical Center (U03MC00001), CHaMP node- State University of New York at Buffalo (U03MC33154), STELAR - Seattle Children's Hospital (U03MC33156), and SPARC node - Emory University School of Medicine (U03MC49671). This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by HRSA, HHS or the U.S. Government. Dr. Aaron Kornblith is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number K23HD110716 (AK).

Declarations

Competing interests

Dr. Aaron Kornblith is a consultant and co-founder for the University of California San Francisco spinout Capture Dx, Inc. and is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number K23HD110716 (AK). This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by HRSA, HHS, or the U.S. Government. All the remaining authors declare no conflict of interest.

Ethics approval and consent to participate

This study was approved under a Single IRB by the University of Utah Institutional Review Board, under the PECARN SIRB PediDOSE Protocol 052, IRB Approval number 001397941.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89996-w.

Correspondence and requests for materials should be addressed to A.E.K.

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.