JAMA Surgery | Original Investigation

Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication

Philip Chung, MD, MS; Christine T. Fong, MS; Andrew M. Walters, MD; Nima Aghaeepour, PhD; Meliha Yetisgen, PhD; Vikas N. O'Reilly-Shah, MD, PhD

Invited Commentary page 937

Supplemental content

IMPORTANCE General-domain large language models may be able to perform risk stratification and predict postoperative outcome measures using a description of the procedure and a patient's electronic health record notes.

OBJECTIVE To examine predictive performance on 8 different tasks: prediction of American Society of Anesthesiologists Physical Status (ASA-PS), hospital admission, intensive care unit (ICU) admission, unplanned admission, hospital mortality, postanesthesia care unit (PACU) phase 1 duration, hospital duration, and ICU duration.

DESIGN, SETTING, AND PARTICIPANTS This prognostic study included task-specific datasets constructed from 2 years of retrospective electronic health records data collected during routine clinical care. Case and note data were formatted into prompts and given to the large language model GPT-4 Turbo (OpenAI) to generate a prediction and explanation. The setting included a quaternary care center comprising 3 academic hospitals and affiliated clinics in a single metropolitan area. Patients who had a surgery or procedure with anesthesia and at least 1 clinician-written note filed in the electronic health record before surgery were included in the study. Data were analyzed from November to December 2023.

EXPOSURES Compared original notes, note summaries, few-shot prompting, and chain-of-thought prompting strategies.

MAIN OUTCOMES AND MEASURES F1 score for binary and categorical outcomes. Mean absolute error for numerical duration outcomes.

RESULTS Study results were measured on task-specific datasets, each with 1000 cases with the exception of unplanned admission, which had 949 cases, and hospital mortality, which had 576 cases. The best results for each task included an F1 score of 0.50 (95% CI, 0.47-0.53) for ASA-PS, 0.64 (95% CI, 0.61-0.67) for hospital admission, 0.81 (95% CI, 0.78-0.83) for ICU admission, 0.61 (95% CI, 0.58-0.64) for unplanned admission, and 0.86 (95% CI, 0.83-0.89) for hospital mortality prediction. Performance on duration prediction tasks was universally poor across all prompt strategies for which the large language model achieved a mean absolute error of 49 minutes (95% CI, 46-51 minutes) for PACU phase 1 duration, 4.5 days (95% CI, 4.2-5.0 days) for hospital duration, and 1.1 days (95% CI, 0.9-1.3 days) for ICU duration prediction.

CONCLUSIONS AND RELEVANCE Current general-domain large language models may assist clinicians in perioperative risk stratification on classification tasks but are inadequate for numerical duration predictions. Their ability to produce high-quality natural language explanations for the predictions may make them useful tools in clinical workflows and may be complementary to traditional risk prediction models.

Author Affiliations: Department of Anesthesiology, Perioperative & Pain Medicine, Stanford University, Stanford, California (Chung, Aghaeepour); Department of Anesthesiology & Pain Medicine, University of Washington, Seattle (Fong, Walters, O'Reilly-Shah); Department of Biomedical & Health Informatics, University of Washington, Seattle (Yetisgen); Department of Linguistics, University of Washington, Seattle (Yetisgen).

Corresponding Author: Philip Chung, MD, MS, Department of Anesthesiology, Perioperative & Pain Medicine, Stanford University, 300 Pasteur Dr, Grant Building S238, Stanford, CA 94305 (chungp@stanford.edu).

JAMA Surg. 2024;159(8):928-937. doi:10.1001/jamasurg.2024.1621
Published online June 5, 2024.

Instruction-tuned large language models (LLMs) have been successful at knowledge retrieval,[1-4] text extraction,[5-9] summarization,[10-12] and reasoning[13-17] tasks without requiring domain-specific fine-tuning. Prompting LLMs with instruction and data contexts described in natural language has emerged as a means for task and domain specification as well as controllability of model behaviors.[18] This investigation assesses the capability of general-domain LLMs in performing preoperative risk stratification and prognostication. This involves assigning a risk score or predicting a postoperative outcome metric based on patient information and details of a surgery or procedure. Such assessments are valuable for proceduralists, surgeons, and anesthesiologists, aiding them in evaluating the risks and benefits associated with proceeding or considering alternatives including canceling or delaying the procedure for medical optimization.

General-domain LLMs have been shown to excel at medical question-and-answer (Q&A) tasks such as US Medical Licensing Exam questions[19-21] or summarization of electronic health record (EHR) text.[22] However, these examinations are not reflective of the real-world clinical setting.[19-21] Multiple-choice test questions present a preselected list of possible answers and often ask questions with clear answers that exist within a well-defined knowledge source such as a medical textbook.[23,24] Real-world EHRs contain patient contexts with uncertain, incomplete, or erroneous information, and a clear answer may be elusive. This investigation was performed using a dataset from this real-world context to benchmark the capabilities of LLMs in perioperative risk prediction and prognostication.

Because there is no single postoperative outcome measure of risk, LLM capabilities were surveyed on 8 different tasks: (1) assignment of the American Society of Anesthesiologists Physical Status (ASA-PS) classification,[25-27] (2) prediction of postanesthesia care unit (PACU) phase 1 duration, (3) hospital admission, (4) hospital duration, (5) intensive care unit (ICU) admission, (6) ICU duration, (7) whether the patient will have an unanticipated hospital admission, and (8) whether the patient will die in the hospital. The LLM-generated responses were compared against ground-truth labels extracted from patients' EHR, and performance metrics were reported based on this comparison (**Figure 1**). Prompting strategies explored include zero-shot prediction, few-shot prediction (also known as in-context learning), and chain-of-thought (CoT) reasoning. Few-shot prompting involves adding representative task and solution examples into the prompt before the actual query task to demonstrate the desired pattern of task and response.[1,28,29] In zero-shot prompting, only the query task is given to the LLM. CoT instructs language models to respond with step-by-step reasoning before providing a final answer.[14,15] Both few-shot and CoT techniques are commonly used to improve task performance. It was hypothesized that LLMs can perform preoperative risk stratification and prognostication using real-world EHR data, and the prediction would be more accurate with few-shot and CoT prompting.

Key Points

Question Can a large language model perform risk stratification and predict postoperative outcome measures using a description of the procedure and a patient's preoperative clinical notes from the electronic health record?

Findings In this prognostic study of task-specific datasets, each with 1000 cases, large language model GPT-4 Turbo (OpenAI) achieved an F1 score of 0.50 for American Society of Anesthesiologists Physical Status, 0.64 for hospital admission, 0.81 for intensive care unit (ICU) admission, 0.61 for unplanned admission, and 0.86 for hospital mortality prediction but was unable to accurately predict duration outcomes such as postanesthesia care unit phase 1 duration, hospital duration, and ICU duration.

Meaning Current generation large language models may be able to assist clinicians with perioperative risk stratification in classification tasks but not numerical prediction tasks.

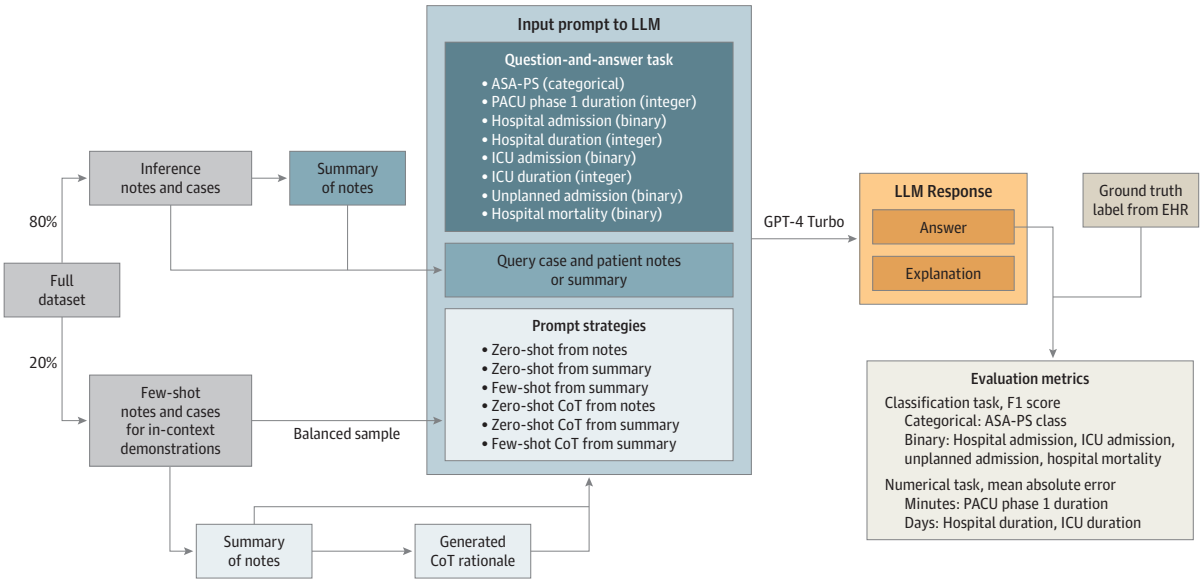
Methods

This was a retrospective prognostic study of routinely collected health records data, approved by the University of Washington (UW) institutional review board with a waiver of consent owing to practicality of carrying out a large-scale retrospective study and minimal risk to participants. The computational environment (eFigure 1 in Supplement 1) for use of protected health information and personally identifiable information was reviewed and approved by UW Medicine Information Technology. The study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines.[30]

Dataset Creation and Experimental Approach

Inclusion criteria were patients who had a surgery or procedure with anesthesia at 3 hospitals (UW Medical Center-Montlake, UW Medical Center-Northwest, Harborview Medical Center) in Seattle, Washington, from April 1, 2021, to May 5, 2023, where the patient had an anesthesia preoperative evaluation note for the case and at least 1 other clinician note filed in the EHR before the case. Up to the last 10 clinician-written notes filed in the EHR before the surgery, excluding the anesthesia preoperative evaluation note, were used. Short notes less than 100 token lengths were excluded. Patient cases, notes, and information from associated hospital admission were then used to extract ground-truth labels and create task-specific datasets targeting 1000 cases per task (eMethods 1 in Supplement 1). Six prompt strategies were composed for each case and used as input to the LLM GPT-4 Turbo (gpt-4-1106-preview) via Microsoft Azure OpenAI service using the OpenAI python client (Figure 1). The prompting strategies were (1) zero-shot Q&A using original notes, (2) zero-shot Q&A using note summaries, (3) few-shot Q&A using note summaries, (4) zero-shot CoT Q&A using original notes, (5) zero-shot CoT Q&A using note summaries, and (6) few-shot CoT Q&A using note summaries. **Figure 2** and eFigure 2 in Supplement 1 depict representative prompts.

Figure 1. Overview of Experimental Apparatus



Overview of the experimental apparatus. Each task-specific dataset is divided into an inference dataset of query cases and a few-shot dataset used to construct few-shot prompts in an 80%-20% split. GPT-4 Turbo (OpenAI) is used as the large language model (LLM) in all steps. Each prompt to the LLM is unique based on the task, prompt strategy, and query case for which an answer and explanation are generated. Unplanned admission refers to patients who were planned for outpatient surgery but were actually admitted postoperatively. Hospital mortality refers to postoperative in-hospital mortality and not 30-day mortality. Zero-shot prompt strategy is conducted with both original clinical notes and a summary of the clinical notes. Few-shot prompts include example demonstrations from the few-shot dataset. Each few-shot demonstration is a question, procedure description, summary of patient notes, and answer. Summaries are generated using the LLM. The few-shot chain-of-thought (CoT) prompt strategy requires a CoT rationale for each few-shot demonstration that links the question to the answer, which is also generated using the LLM. Few-shot demonstrations are dynamically selected for each query case with inverse frequency sampling to balance the distribution of answers of few-shot demonstrations. Answers provided by the LLM are compared against the ground-truth label derived from electronic health record (EHR) data, and either an F1 score or mean absolute error is computed, depending on whether the outcome variable for the task is categorical/binary or numerical. ASA-PS indicates American Society of Anesthesiologists Physical Status; ICU, intensive care unit; PACU, postanesthesia care unit.

Statistical Analysis

The primary measure of performance was as follows: (1) F1 score for binary tasks, (2) F1 microaverage (F1-micro) score for categorical ASA-PS prediction, and (3) mean absolute error (MAE) for duration tasks. 95% CIs were estimated using 2500 bootstrap iterations. Statistical significance was tested using Wilcoxon signed rank test for pairwise comparison of prompt strategies using 2-sided *P* values, controlling the false discovery rate using Benjamini-Hochberg procedure ($\alpha = .05$). Analysis were conducted using statsmodels, version 0.14.0, and SciPy, version 1.1.4. Token length of notes was evenly stratified into 3 categories—short, medium, and long—and performance of each stratum was also reported. Prompt strategies for each task were compared against the following baselines: F1 or F1-micro score of a random classifier if the task is a binary or categorical outcome, respectively, and a dummy regressor that always predicts the mean duration in the dataset for numerical outcomes.

Results

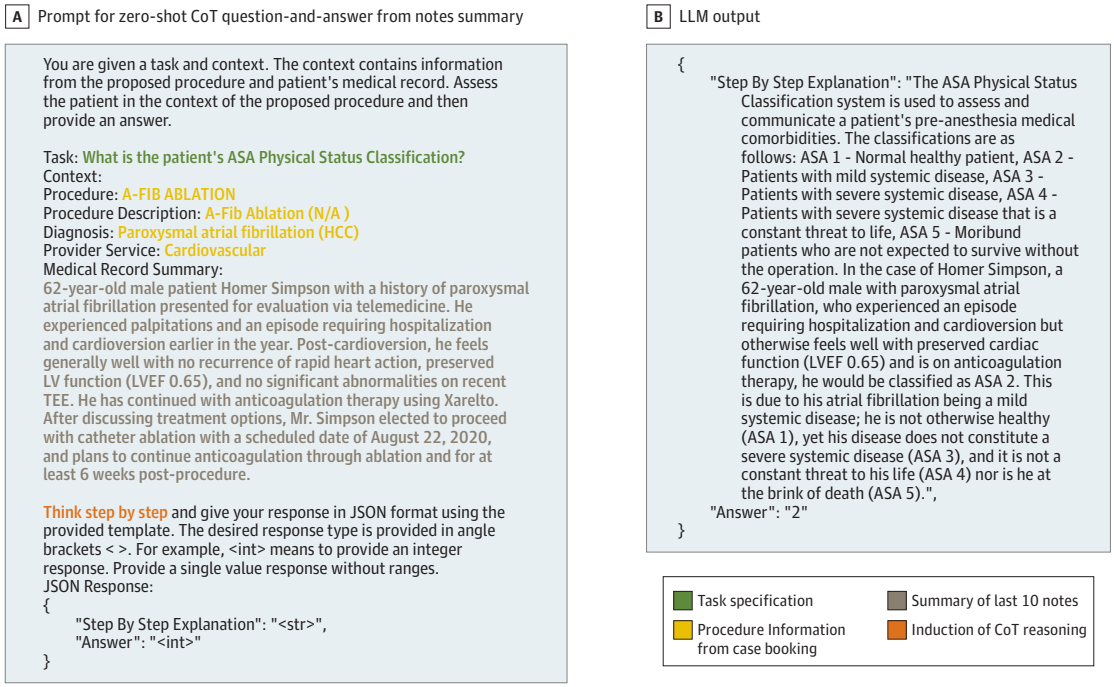
Datasets

Task-specific datasets are described in eTable 1 in Supplement 2, note type and author type in eTable 2 in Supplement 2, and experiment costs in eTable 3 in Supplement 2. Task-specific datasets created for this study include the following: (1) the ASA-PS dataset with 1000 patients, mean (SD) age of 51.9 (19.9) years, and 55.3% male; (2) the hospital admission dataset with 1000 patients, mean (SD) age of 50.5 (19.3) years, and 56.0% male; (3) the ICU admission and duration dataset with 1000 patients, mean (SD) age of 53.5(18.8) years, and 50.7% male; (4) the unplanned admission dataset with 949 patients, mean (SD) age of 54.8 (18.6) years, and 50.4% male; (5) the hospital mortality dataset with 576 patients, mean (SD) age of 58.2 (18.9) years, and 58.3% male; (6) the PACU phase 1 duration dataset with 1000 patients, mean (SD) age, 50.0 (19.5) years, and 50.6% male; and (7) the hospital duration dataset with 1000 patients, mean (SD) age, 56.1 (19.1) years, and 54.4% male. Race and ethnicity data were not collected. The dataset creation flow diagram is shown in eFigure 3 in Supplement 2, and overlap of the datasets is shown in eFigure 4 in Supplement 2. Details are available in the eResults in Supplement 1.

Association of Prompt Strategy With Perioperative Risk Prediction Tasks

Figure 2 depicts an example prompt and LLM response to illustrate LLM inputs and outputs. Performance of each prompt strategy and risk prediction task is summarized in **Figure 3** and **Figure 4** and is reported in detail with CIs in eTables 4 to 24 in

Figure 2. Annotated Example Prompt and Large Language Model (LLM) Response



A prompt and LLM output for zero-shot chain-of-thought (CoT) question and answer (Q&A) from notes summary prompt strategy. A, Prompt with highlights to show how data from cases and notes are inserted into the prompt template. Black text is the template for the specific prompt strategy, which is the same for all tasks with the exception of variable type being substituted in the response specification (eg, "<int>" may be "<bool>" for binary prediction tasks). Procedure information is inserted into the prompt template without modification. Note summaries are generated from clinical notes in a separate step using the LLM and then the summary is inserted into the prompt template. B, LLM output explanation shows that the LLM understands the definition for American Society of Anesthesiologists Physical Status (ASA-PS) classification and provides a valid rationale for classifying the patient's ASA-PS. Because LLMs are left-to-right causal language models, CoT prompt strategies always request generation of the step-by-step explanation before the final answer to ensure the LLM considers the explanation when generating the final answer. Although the content of this example is derived from a real patient and case from the electronic health record, all protected health information and personally identifiable information are removed with names obfuscated and dates and times shifted. More detail on all prompt strategies used in experiments including prompts used to generate summaries and CoT rationales are depicted in eFigure 2 in Supplement 1. A-Fib indicates atrial fibrillation; HCC, Hierarchical Condition Category; JSON, JavaScript Object Notation; LVEF, left ventricular ejection fraction.

Supplement 2. Statistical significance comparing each pair of prompt strategies for each task is shown in eFigures 5 to 12 in Supplement 1.

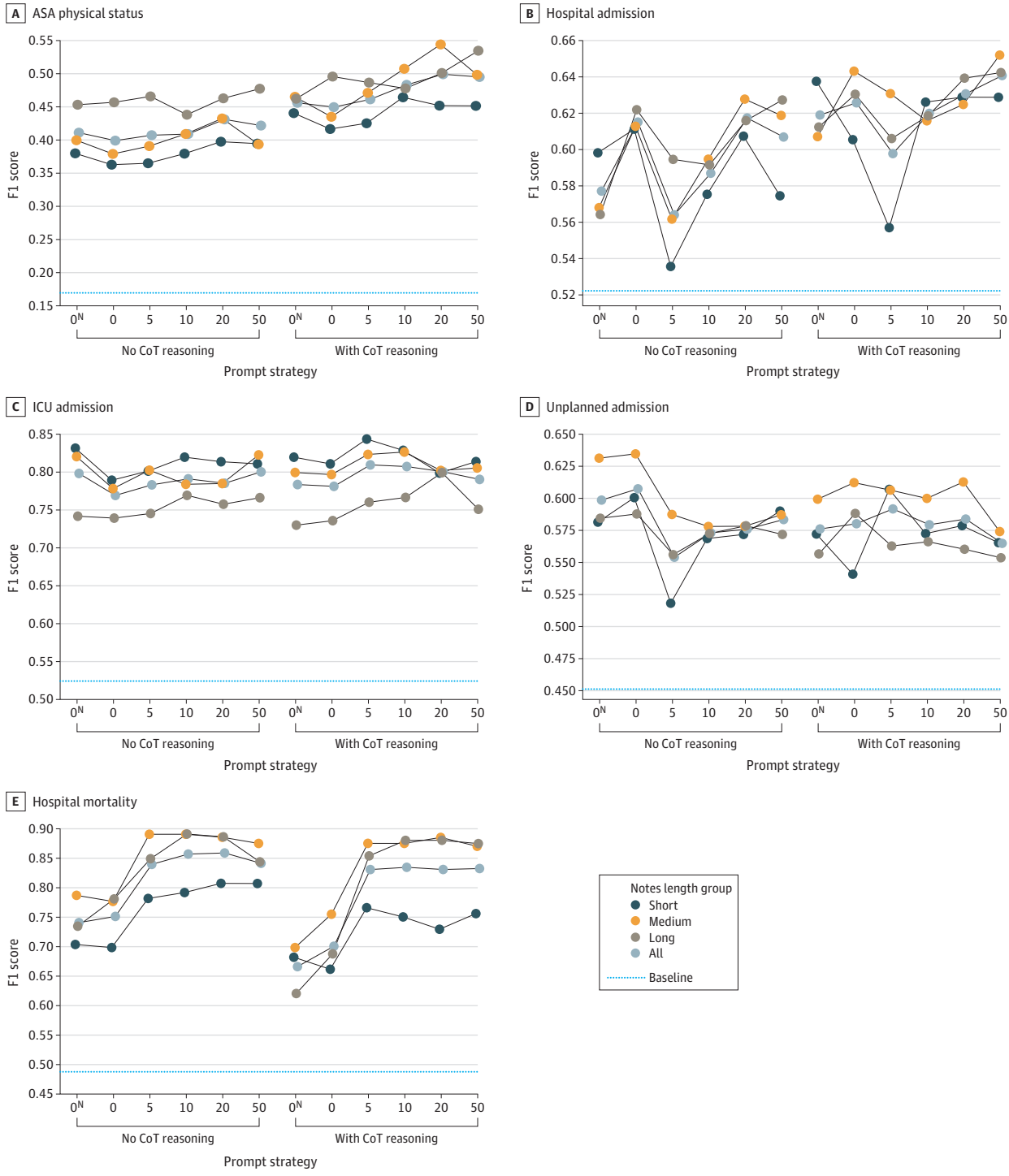
Binary and Categorical Outcomes

Figure 3 shows that all prompt strategies outperform the random baseline for binary and categorical prediction tasks, which include ASA-PS, hospital admission, ICU admission, unplanned admission, and hospital mortality. Compared against no CoT, the CoT prompt strategy improved F1 score for ASA-PS by 0.4 to 0.8; similarly, F1 score for hospital admission improved by 0.3 to 0.4 (Figure 3). However, CoT did not show improvements in F1 score with ICU admission, unplanned admission, and hospital mortality. Increasing the number of few-shot examples to 20 or 50 generally resulted in better performance than zero-shot prompting. Few-shot prompting increased F1 score for ASA-PS (from 0.45 to 0.5 with CoT), hospital admission (from 0.58 to 0.62 without CoT), and hospital mortality (from 0.74 to 0.86 without CoT) (Figure 3). Combining few-shot prompting with CoT yielded synergistic gains in F1 score for ASA-PS and hospital admission, but this was not observed for hospital mortality (Figure 3). ICU admission prediction performance was high across all prompt strategies with F1 score ranging from 0.77 to 0.81, suggesting that the LLM is easily able to perform this task regardless of prompt strategy (Figure 3). Zero-shot without CoT using note summaries was the best for unplanned admission with F1 score of 0.61 and demonstrated that CoT rationales do not help with all prediction tasks (Figure 3). In addition to F1, eTables 4 to 21 in Supplement 2 show sensitivity, specificity, positive predictive value, negative predictive value, and Matthew correlation coefficient for the binary and categorical tasks, and confusion matrices are shown in eFigures 13 to 17 in Supplement 1.

Numerical Outcomes

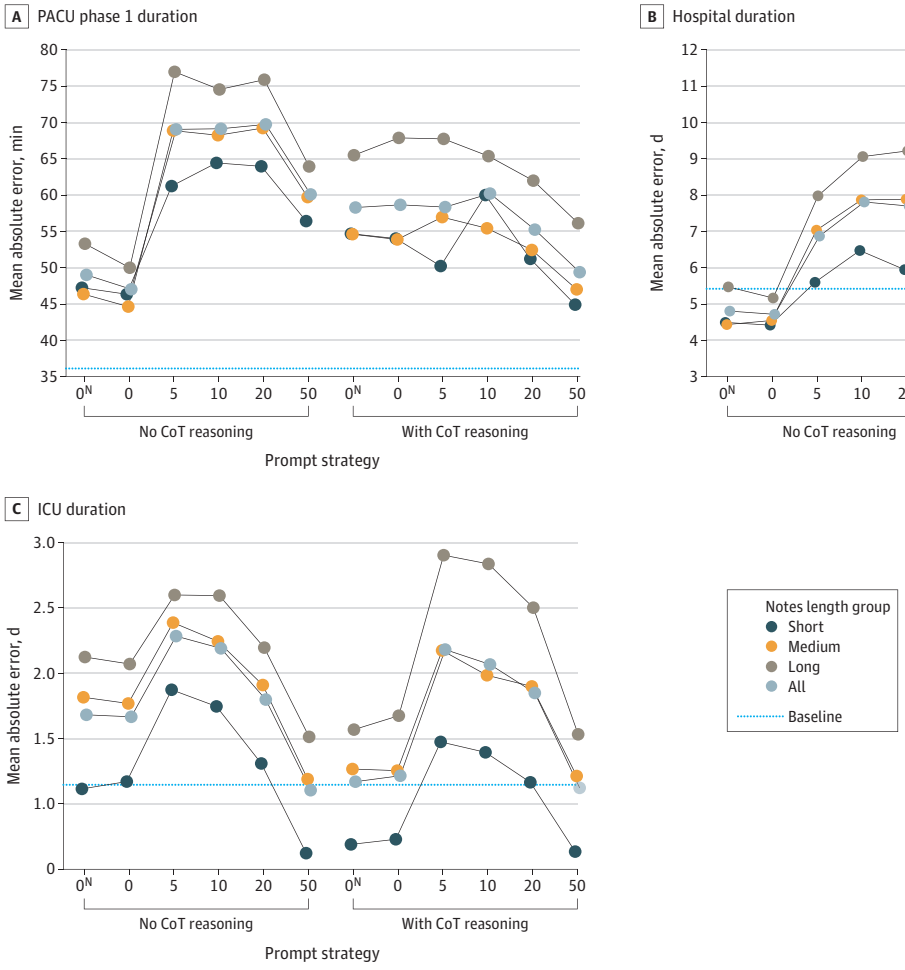
Figure 4 shows that GPT-4 Turbo (OpenAI) fails to perform numerical predictions better than the dummy regressor baseline. For PACU phase 1 duration prediction, all prompt strategies performed worse than baseline. For hospital duration prediction, zero-shot CoT using notes summaries achieved an MAE of 4.55 days (95% CI, 4.18-4.98 days) vs baseline MAE of 5.4 days (95% CI, 5.04-5.85) (Figure 4 and eTable 23 in Supplement 2). Few-shot prompting worsened hospital duration prediction, despite few-shot and CoT prompting improving

Figure 3. Association of Prompt Strategy With Binary and Categorical Prediction Tasks



The large language model (LLM) GPT-4 Turbo (OpenAI) prediction performance on the 5 binary and categorical prediction tasks. The x-axis shows the different prompt strategies with the first 6 prompt strategies without chain-of-thought (CoT) reasoning and the second 6 with CoT reasoning. "N" indicates that original clinical notes were inserted into the prompt. The remaining prompt strategy numbers are where clinical notes were first summarized using the LLM and then the summary was inserted into the prompt. "0" Corresponds to "0-shot," which indicates that zero-shot prompt strategy was used. "5" Corresponds to "5-shot," "10" with "10-shot," "20" with "20-shot," and "50" with "50-shot"; all refer to few-shot prompts with the respective number of few-shot demonstrations. All few-shot prompts used note summaries for both few-shot demonstrations and query case. y-axis is F1 score for classification tasks where higher score is better. Baseline for each task is different and represents the score achieved by random guessing. The clinical notes are stratified into short, medium, and long length groups, which represent the 1/3 shortest, 1/3 middle, and 1/3 longest notes by token count (word subunits used by LLMs, 1 token approximately equals 3/4 word) and performance is shown for each stratification. CIs are omitted for legibility but are available in eTables 4 to 21 in Supplement 2. ASA-PS indicates American Society of Anesthesiologists Physical Status; ICU, intensive care unit; PACU, postanesthesia care unit.

Figure 4. Effect of Prompt Strategy on Numerical Prediction Tasks



The large language model (LLM) GPT-4 Turbo (Open AI) prediction performance on the 3 numerical prediction tasks. The x-axis shows the different prompt strategies with the first 6 prompt strategies without chain-of-thought (CoT) reasoning and the second 6 with CoT reasoning. "N" indicates that original clinical notes were inserted into the prompt. The remaining prompt strategy numbers are where clinical notes were first summarized using the LLM and then the summary was inserted into the prompt. "0" corresponds to "0-Shot," which indicates that zero-shot prompt strategy was used. "5" Corresponds with "5-shot," "10" with "10-shot," "20" with "20-shot," "50" with "50-shot"; all refer to few-shot prompts with the respective number of few-shot demonstrations.

All few-shot prompts used note summaries for both few-shot demonstrations and query case. y-axis is mean absolute error (MAE) for numerical prediction tasks where lower error is better. The baseline for numerical prediction tasks represents the MAE achieved by a dummy regressor that always predicts the mean outcome value in the dataset. The clinical notes are stratified into short, medium, and long length groups, which represent the 1/3 shortest, 1/3 middle, and 1/3 longest notes by token count (word subunits used by LLMs, 1 token approximately equals 3/4 word) and performance is shown for each stratification. CIs are omitted for legibility but are available in eTables 22 to 24 in Supplement 2. ICU indicates intensive care unit; PACU, postanesthesia care unit.

the analogous hospital admission binary prediction task (Figure 4). ICU duration prediction only achieved parity with baseline when few-shot and CoT was used, despite the analogous ICU admission binary prediction task achieving significantly better performance (Figure 3 and Figure 4). To understand why numerical outcome prediction was poor, we visualized the distribution of the predictions for PACU phase 1 duration (eFigure 18 in Supplement 1), hospital duration (eFigure 19 in Supplement 1), and ICU duration (eFigure 20 in Supplement 1), revealing that LLMs tend to predict quantized outputs, often with a ceiling effect. The application of few-shot and CoT prompting helps remove the quantization effects but does not improve prediction accuracy.

Overall

The best-performing prompt strategies for each prediction task included an F1 score of 0.50 (95% CI, 0.47-0.53) for ASA-PS, 0.64 (95% CI, 0.61-0.67) for hospital admission, 0.81 (95% CI, 0.78-0.83) for ICU admission, 0.61 (95% CI, 0.58-0.64) for unplanned admission, and 0.86 (95% CI, 0.83-0.89) for hospital mortality prediction. The specific details of each task are described subsequently.

ASA-PS | Both 20-shot CoT and 50-shot CoT prompting achieved an F1-micro score of 0.50, which is 2.94 times (194% greater than) the random classifier baseline with an F1-micro score of 0.17 (eTable 4 in Supplement 2).

Hospital Admission | 50-shot CoT had an F1 score of 0.64, which is 1.23 times (23% greater than) the random classifier baseline with an F1 score of 0.52 (eTable 10 in Supplement 2).

ICU Admission | 5-shot CoT prompt strategy had an F1 score of 0.81, which is 1.56 times (56% greater than) the random classifier baseline with an F1 score of 0.52 (eTable 13 in Supplement 2).

Unplanned Admission | Zero-shot using note summaries had an F1 score of 0.61, which is 1.36 times (36% greater than) the random classifier baseline with an F1 score of 0.45 (eTable 16 in Supplement 2).

Hospital Mortality | 10-shot and 20-shot had an F1 score of 0.86, which is 1.76 times (76% greater than) the random classifier baseline with an F1 score of 0.49 (eTable 19 in Supplement 2)

PACU Phase 1 Duration | Zero-shot using original notes had an MAE of 49 minutes (95% CI, 46-51 minutes), which is 1.36 times (36% greater error than) the dummy regressor baseline MAE of 36 minutes (95% CI, 34-38 minutes) (eTable 22 in Supplement 2).

Hospital Duration | Zero-shot CoT using notes summary had an MAE of 4.5 days (95% CI, 4.2-5.0 days), which is 0.83 times (20% lower error than) the dummy regressor baseline MAE of 5.4 days (95% CI, 5.0-5.8 days) (eTable 23 in Supplement 2).

ICU Duration | Both 50-shot and 50-shot CoT had an MAE of 1.1 days (95% CI, 0.9-1.3 days), which is roughly the same error as the dummy regressor baseline MAE of 1.1 days (95% CI, 1.0-1.3 days) (eTable 24 in Supplement 2).

Outcome of Summary Representation of Notes

Prior work has shown that LLM-generated summaries in the clinical domain may be preferable to human-written summaries.[22] Comparison of zero-shot prompts using original notes vs zero-shot prompts using LLM-generated summaries resulted in slight degradation of performance as seen in ASA-PS, ICU admission, PACU phase 1 duration, and hospital duration but also boosted performance for hospital admission, unplanned admission, and hospital mortality prediction. The magnitude of these outcomes was small.

Association of Note Length With Perioperative Risk Prediction Tasks

Note length had a differential association with several tasks, including better performance for ASA-PS prediction and hospital mortality prediction. Because up to the last 10 clinical notes were used, increased note length was due to either longer notes or more notes being written about the patient. However, longer input note lengths performed worse for ICU admission prediction, PACU phase 1 duration prediction, and hospital duration prediction.

Discussion

Results of this prognostic study suggest that general-domain LLMs such as GPT-4 Turbo (OpenAI) have the capability to perform some aspects of perioperative risk assessment and prognostication, especially with categorical and binary prediction tasks. Strong performance for prediction of ASA-PS, postoperative ICU admission, and hospital mortality across all prompt strategies was observed. ASA-PS assignment is known to be subjective and has only moderate interrater agreement among human anesthesiologists[31,32]; therefore, it is unlikely that any prediction system can achieve a perfect score. In this context, an F1-micro score of 0.5 or 2.94 times greater than an F1 score for random guessing has meaningful clinical utility. The multimodal LLM tends not to make large errors in ASA-PS prediction, and confusion matrices show that ASA-PS misclassifications made by the LLM with few-shot and CoT prompting are almost always an adjacent ASA class (eFigure 13 in Supplement 1). However, an F1-micro score does not give any credit for adjacent score predictions, which is a harsh penalty given the poor interrater agreement among humans. The F1 score also does not score the clinical utility from the LLM's natural language explanation that explains the predicted ASA-PS. In practice, both ASA-PS score and text explanation have clinical utility, and the F1 score likely underestimates the true value of LLM prediction for the ASA-PS task.

Hospital admission and unplanned admission prediction performance is better than random guessing but not as impressive as ICU admission and hospital mortality where the illness severity of a patient is likely more apparent and makes for an easier prediction scenario. Still, the multimodal LLM achieves remarkable predictive performance from only procedure description and clinical notes with no specialized clinical training and no fine-tuning for perioperative risk prediction tasks.

Few-shot and CoT prompting reveal significant gains in categorical prediction tasks where synthesizing prior clinical knowledge is important, such as determination of ASA-PS, hospital admission prediction, and hospital mortality. These outcomes are additive and synergistic, but the benefits of these prompting techniques do not apply to all outcomes. These prediction tasks likely benefit from the prompting strategies because they are heavily dependent on preoperative illness severity, which would be reflected in a patient's clinical notes. Few-shot examples help the LLM compare and contrast among similar cases, whereas CoT rationales help expand on the concepts mentioned in clinical notes, both of which guide the LLM toward more accurate predictions. In contrast, it is possible that these gains are not seen in outcomes such as unplanned admission because factors leading to unexpected admission are predominantly due to intraoperative events—information not available in preoperative clinical notes and procedure booking data presented to the LLM—and no amount of deliberation or rationalization would affect the outcome.

LLMs struggle with numerical prediction tasks such as PACU phase 1 duration, hospital duration, and ICU duration. The LLM predicts quantized values, which we suspect is due to the LLM memorizing length of stay estimations from hospital websites,

textbooks, and journal articles. Few-shot demonstrations and prompting the multimodal LLM to rationalize about the patient's procedure and medical history help overcome this quantization phenomenon, but the continued poor results may be attributed to the architectural design of LLMs. Namely, LLMs enforce a discrete tokenized output where each token's representation is derived from text contexts. For continuous-valued outcomes, it is meaningful for humans to interpolate between numerical values. An LLM's training data and training process do not provide a robust way for the model to interpolate numerical values. Potential strategies to overcome this limitation include multimodal enhancements to LLMs to treat numbers as distinct data modalities and directly mapping of continuous values to and from the embedding space of neural network layers.[33-36] Although many visual-language[37-39] and multimodal models adopt these strategies to combine text and other data modalities such as pixel intensity in the same model, no widely available LLM has yet used these solutions for general numerical predictions. Future foundational models for health care or EHR data should consider model architectures and pretraining routines that enable better performance for these kinds of numerical prediction tasks. Another alternative is equipping LLMs with tool use,[40-43] but this relegates the LLM as a natural language information extractor and outsources the actual prediction task to an external model rather than taking full advantage of the LLM's capabilities in information synthesis.

Note length stratification indicated that longer input contexts do not necessarily result in better performance. Contrary to the intuition that providing the LLM more clinical context would enable a more accurate prediction, increased note length may also correlate with greater presence of tangential, outdated, or conflicting information that detracts from accurate predictions. Similarly, although transforming notes to summaries could result in loss of information useful for the predictive task, summaries may also help focus relevant information. The advantage of using summaries in our experiments was the ability to scale to 50-shot examples while staying within the LLM's input context window, resulting in significantly better performance for some tasks. The attention mechanism used in LLMs is biased toward the beginning and end of prompts, which may be an artifact of failure to train models on long-context data.[44] This phenomenon could also explain why shorter note length or summaries sometimes outperform longer inputs.

Overall, these results suggest that currently available general-domain LLMs may be useful for perioperative risk stratification workflows in hospitals and may assist in stratifying the preoperative patient population for these outcomes. The LLM in this study exhibited very good performance at ASA-PS classification prediction, ICU admission prediction, and hospital mortality prediction. When strictly comparing metrics such as F1 score, LLMs still underperform dedicated classification models utilizing tabular features.[45-58] Traditional machine learning models are rarely utilized in the clinical setting because of difficulty in interpreting a model's predictions. In contrast, LLMs present natural language explanations understandable to human clinicians, and can develop a rationale for each outcome variable of interest. These explanations can provide a valuable starting point for clinicians in comprehensive perioperative risk assessment and may be more useful than standalone risk predictions. Further work is necessary to assess the clinical accuracy and utility of these explanations.

Limitations

There are several limitations to this study. There is no ground-truth label to evaluate the clinical utility of an LLM's text explanations; study evaluations mainly quantified the binary, categorical, or numerical answer to each prediction task. The dataset did not contain 30-day hospital mortality labels; therefore, only in-hospital mortality prediction was studied. Hospital readmission was not studied; unanticipated admission was defined as a surgery booked as outpatient, but the patient was admitted. Furthermore, the incidence of some outcomes was rare. Of the 137 535 cases considered in the 2-year span from which the datasets were derived, only 0.49% of cases had postoperative ICU admissions, 0.43% of cases had unanticipated hospital admission, and 0.3% of cases had postoperative in-hospital mortality (eFigure 3 in Supplement 1). Outcome-balanced task-specific datasets were created to measure the LLM's performance, but true performance of the model against the real-world prevalence of outcomes requires further investigation. It is also costly to use models like the LLM used in this study on long-context clinical notes for large number of patient cases, which imposed practical constraints on our choice of data set size (eTable 3 in Supplement 2). Future research is needed to explore better methods for evaluating LLM outputs in terms of their clinical utility, assessing the performance of specialized clinical domain-specific language models,[6,7,19,59] and investigating whether advanced prompting strategies such as dynamic k-Nearest-Neighbor few-shot or retrieval-augmentation enhance performance.[21,23,60-62] Future large-scale prospective clinical validation is necessary to verify the observed performance, study whether LLM-based clinical decision support systems significantly bias clinician judgment, and compare against existing perioperative prediction algorithms.[57,58]

Conclusions

Results of this prognostic study suggest that although general-domain text-only LLMs are capable of perioperative risk prediction and prognostication when framed as classification or binary prediction tasks, they were unable to predict continuous-valued outcomes such as PACU, hospital, and ICU length of stay. Few-shot prompting and CoT reasoning improved prediction performance for perioperative prediction tasks. Future prospective studies are needed to verify the effectiveness of large language models as tools to assist perioperative risk stratification.

ARTICLE INFORMATION

Accepted for Publication: March 8, 2024.

Published Online: June 5, 2024. doi:10.1001/jamasurg.2024.1621

Author Contributions: Dr Chung had full access to all of the data in the study and takes responsibility

for the integrity of the data and the accuracy of the data analysis.
Concept and design: Chung, Walters, Aghaeepour, Yetisgen, O'Reilly-Shah.
Acquisition, analysis, or interpretation of data: Chung, Fong, Walters, Aghaeepour, O'Reilly-Shah.
Drafting of the manuscript: Chung, Aghaeepour, O'Reilly-Shah.
Critical review of the manuscript for important intellectual content: Chung, Fong, Walters, Yetisgen, O'Reilly-Shah.
Statistical analysis: Chung, Walters.
Obtained funding: Chung, O'Reilly-Shah.
Administrative, technical, or material support: Chung, Fong, Walters, O'Reilly-Shah.
Supervision: Walters, Aghaeepour, Yetisgen, O'Reilly-Shah.

Conflict of Interest Disclosures: Dr Walters reported receiving consulting fees from Sonosite and Philips outside the submitted work.
Dr O'Reilly-Shah reported being an equity holder of Doximity Inc outside the submitted work. No other disclosures were reported.

Funding/Support: Computational resources for this project were funded by the Microsoft Azure Cloud Compute Credits grant program from the University of Washington eScience Institute and Microsoft Azure. Financial support for this work was provided by the University of Washington Department of Anesthesiology & Pain Medicine's Bonica Scholars Program, Stanford University Research in Anesthesia Training Program (ReAP) program, and National Institutes of Health grants 5T32GM089626, and R35GM138353.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 3.

Additional Contributions: We thank University of Washington Anesthesia Department's Perioperative & Pain initiatives in Quality Safety Outcome group for assistance on data extraction and discussions in dataset and experimental design; University of Washington Department of Medicine for computational environment support; Roland Lai, BA, and Robert Fabiano, BS, from University of Washington Research IT for creating a digital research environment within the Microsoft Azure Cloud where model development and experiments were performed; and the University of Washington Biomedical Natural Language Processing group, and the Aghaeepour Laboratory at Stanford University for providing early feedback on experimental design and results. Acknowledged individuals did not receive financial compensation from the study and provided written permission to be named.

Additional Information: Code availability: code for experiments is publicly available at https://github.com/philipchung/llm-periop-prediction.

REFERENCES

1. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20*. Curran Associates Inc; 2020:1877-1901.

2. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv [csCL]*. Published online March 4, 2022. http://arxiv.org/abs/2203.02155

3. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare:instruction-tuned large language models for medical application. *arXiv [csCL]*. Published online October 23, 2023. http://arxiv.org/abs/2310.14558

4. Taori R, Gulrajani I, Zhang T, et al. Stanford alpaca: an instruction-following llama model. Accessed November 28, 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html

5. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2022:1998-2022. doi:10.18653/v1/2022.emnlp-main.130

6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

7. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding. *arXiv [csCL]*. Published online May 19, 2023. http://arxiv.org/abs/2305.12031

8. Ramachandran GK, Fu Y, Han B, et al. Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2023:385-393. doi:10.18653/v1/2023.clinicalnlp-1.41

9. Ramachandran GK, Lybarger K, Liu Y, et al. Extracting medication changes in clinical narratives using pre-trained language models. *J Biomed Inform*. 2023;139:104302. doi:10.1016/j.jbi.2023.104302

10. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking large language models for news summarization. *arXiv [csCL]*. Published online January 31, 2023. http://arxiv.org/abs/2301.13848

11. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize from human feedback. *arXiv [csCL]*. Published online September 2, 2020. http://arxiv.org/abs/2009.01325

12. Wu J, Ouyang L, Ziegler DM, et al. Recursively summarizing books with human feedback. *arXiv [csCL]*. Published online September 22, 2021. http://arxiv.org/abs/2109.10862

13. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv [csCL]*. Published online June 15, 2022. http://arxiv.org/abs/2206.07682

14. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv [csCL]*. Published online January 28, 2022. http://arxiv.org/abs/2201.11903

15. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *arXiv [csCL]*. Published online May 24, 2022. http://arxiv.org/abs/2205.11916

16. Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. *arXiv [csCL]*. Published online October 6, 2022. http://arxiv.org/abs/2210.03629

17. Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. *arXiv [csCL]*. Published online May 17, 2023. http://arxiv.org/abs/2305.10601

18. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. Accessed January 6, 2022. https://www.semanticscholar.org/paper/9405cc0d6169988371b2755e573cc28650d14dfe

19. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv [csCL]*. Published online May 16, 2023. http://arxiv.org/abs/2305.09617

20. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv [csCL]*. Published online November 28, 2023. http://arxiv.org/abs/2311.16452

21. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv [csCL]*. Published online March 20, 2023. http://arxiv.org/abs/2303.13375

22. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024. doi:10.1038/s41591-024-02855-5

23. Wang Y, Ma X, Chen W. Augmenting black-box LLMs with medical textbooks for clinical question answering. *arXiv [csCL]*. Published online September 5, 2023. http://arxiv.org/abs/2309.02233

24. Zakka C, Shad R, Chaurasia A, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*. 2024;1(2). doi:10.1056/Aloa2300068

25. Saklad M. Grading of patients for surgical procedures. *Anesthesiology*. 1941;2(3):281-284. doi:10.1097/00000542-194105000-00004

26. Mayhew D, Mendonca V, Murthy BVS. A review of ASA physical status—historical perspectives and modern developments. *Anaesthesia*. 2019;74(3):373-379. doi:10.1111/anae.14569

27. Horvath B, Kloesel B, Todd MM, Cole DJ, Prielipp RC. The evolution, current value, and future of the American Society of Anesthesiologists physical status classification system. *Anesthesiology*. 2021;135(5):904-919. doi:10.1097/ALN.0000000000003947

28. Olsson C, Elhage N, Nanda N, et al. In-context learning and induction heads. *arXiv [csLG]*. Published online September 24, 2022. http://arxiv.org/abs/2209.11895

29. Wei J, Wei J, Tay Y, et al. Larger language models do in-context learning differently. *arXiv [csCL]*. Published online March 7, 2023. http://arxiv.org/abs/2303.03846

30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63. doi:10.7326/M14-0697

31. Cuvillon P, Nouvellon E, Marret E, et al. American Society of Anesthesiologists' physical status system: a multicenter Francophone study to analyze reasons for classification disagreement. *Eur J Anaesthesiol*. 2011;28(10):742-747. doi:10.1097/EJA.0b013e328348fc9d

32. Sankar A, Johnson SR, Beattie WS, Tait G, Wijeysundera DN. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *Br J Anaesth*. 2014;113(3):424-432. doi:10.1093/bja/aeu100

33. Driess D, Xia F, Sajjadi MSM, et al. PaLM-E: an embodied multimodal language model. *arXiv*

[csLG]. Published online March 6, 2023. http://arxiv.org/abs/2303.03378

34. Belyaeva A, Cosentino J, Hormozdiari F, et al. Multimodal LLMs for health grounded in individual-specific data. *arXiv [q-bioQM]*. Published online July 18, 2023. http://arxiv.org/abs/2307.09018

35. Xu S, Yang L, Kelly C, et al. ELIXR: Toward a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv [csCV]*. Published online August 2, 2023. http://arxiv.org/abs/2308.01317

36. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *arXiv [csCL]*. Published online July 26, 2023. http://arxiv.org/abs/2307.14334

37. Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. *arXiv [csCV]*. Published online April 29, 2022. http://arxiv.org/abs/2204.14198

38. Moor M, Huang Q, Wu S, et al. Med-flamingo: a multimodal medical few-shot learner. *arXiv [csCV]*. Published online July 27, 2023. http://arxiv.org/abs/2307.15189

39. Chen X, Wang X, Changpinyo S, et al. PaLI: a jointly-scaled multilingual language-image model. *arXiv [csCV]*. Published online September 14, 2022. http://arxiv.org/abs/2209.06794

40. Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools. *arXiv [csCL]*. Published online February 9, 2023. http://arxiv.org/abs/2302.04761

41. Qin Y, Liang S, Ye Y, et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs. *arXiv [csAI]*. Published online July 31, 2023. http://arxiv.org/abs/2307.16789

42. Cai T, Wang X, Ma T, Chen X, Zhou D. Large language models as tool makers. *arXiv [csLG]*. Published online May 26, 2023. http://arxiv.org/abs/2305.17126

43. Goodell AJ, Chu SN, Rouholiman D, Chu LF. Augmentation of ChatGPT with clinician-informed tools improves performance on medical calculation tasks. *bioRxiv*. Preprint posted online December 15, 2023. doi:10.1101/2023.12.13.23299881

44. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *arXiv [csCL]*. Published online July 6, 2023. http://arxiv.org/abs/2307.03172

45. Mudumbai SC, Pershing S, Bowe T, et al. Development and validation of a predictive model for American Society of Anesthesiologists Physical Status. *BMC Health Serv Res*. 2019;19(1):859. doi:10.1186/s12913-019-4640-x

46. Graeßner M, Jungwirth B, Frank E, et al. Enabling personalized perioperative risk prediction by using a machine-learning model based on preoperative data. *Sci Rep*. 2023;13(1):7128. doi:10.1038/s41598-023-33981-8

47. Lee SW, Lee HC, Suh J, et al. Multicenter validation of machine learning model for preoperative prediction of postoperative mortality. *NPJ Digit Med*. 2022;5(1):91. doi:10.1038/s41746-022-00625-6

48. Hill BL, Brown R, Gabel E, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *Br J Anaesth*. 2019;123(6):877-886. doi:10.1016/j.bja.2019.07.030

49. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg*. 2013;217(5):833-842.e1-3. doi:10.1016/j.jamcollsurg.2013.07.385

50. Chen PF, Chen L, Lin YK, et al. Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. *JMIR Med Inform*. 2022;10(5):e38241. doi:10.2196/38241

51. Xu Z, Yao S, Jiang Z, et al. Development and validation of a prediction model for postoperative intensive care unit admission in patients with non-cardiac surgery. *Heart Lung*. 2023;62:207-214. doi:10.1016/j.hrtlng.2023.08.001

52. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical risk preoperative assessment system (SURPAS): iii. accurate preoperative prediction of 8 adverse outcomes using 8 predictor variables. *Ann Surg*. 2016;264(1):23-31. doi:10.1097/SLA.0000000000001678

53. Tully JL, Zhong W, Simpson S, et al. Machine learning prediction models to reduce length of stay at ambulatory surgery centers through case resequencing. *J Med Syst*. 2023;47(1):71. doi:10.1007/s10916-023-01966-9

54. Fang F, Liu T, Li J, et al. A novel nomogram for predicting the prolonged length of stay in postanesthesia care unit after elective operation. *BMC Anesthesiol*. 2023;23(1):404. doi:10.1186/s12871-023-02365-w

55. Gabriel RA, Waterman RS, Kim J, Ohno-Machado L. A predictive model for extended postanesthesia care unit length of stay in outpatient surgeries. *Anesth Analg*. 2017;124(5):1529-1536. doi:10.1213/ANE.0000000000001827

56. Dyas AR, Henderson WG, Madsen HJ, et al. Development and validation of a prediction model for conversion of outpatient to inpatient surgery. *Surgery*. 2022;172(1):249-256. doi:10.1016/j.surg.2022.01.025

57. Le Manach Y, Collins G, Rodseth R, et al. Preoperative score to predict postoperative mortality (POSPOM): derivation and validation. *Anesthesiology*. 2016;124(3):570-579. doi:10.1097/ALN.0000000000000972

58. Smilowitz NR, Berger JS. Perioperative Cardiovascular risk assessment and management for noncardiac surgery: a review. *JAMA*. 2020;324(3):279-290. doi:10.1001/jama.2020.7840

59. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: scaling medical pretraining for large language models. *arXiv [csCL]*. Published online November 27, 2023. http://arxiv.org/abs/2311.16079

60. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. *arXiv [csCL]*. Published online March 21, 2022. http://arxiv.org/abs/2203.11171

61. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv [csCL]*. Published online May 22, 2020. http://arxiv.org/abs/2005.11401

62. Zakka C, Chaurasia A, Shad R, et al. Almanac: retrieval-augmented language models for clinical medicine. *arXiv [csCL]*. Published online March 1, 2023. http://arxiv.org/abs/2303.01229

Invited Commentary

Travel Guide From the Brave New World of Artificial Intelligence

Daniel E. Hall, MD, MDiv, MHSc

Just over 1 year has passed since ChatGPT (OpenAI) burst onto the scene to become the fastest growing consumer software application in history.[1] It was immediately obvious that large language models (LLMs) were extraordinarily powerful, but the scope, nature, application, and implications of that power remain unclear. Chung and colleagues[2] present a fascinating article in *JAMA Surgery* demonstrating the ability of a specific LLM (GPT-4 Turbo [OpenAI]) to inspect preoperative clinician notes from the electronic health record to classify or predict several perioperative parameters such as the American Society of Anesthesiologists Physical Status (ASA) classification, hospital or intensive care unit admission, and postoperative lengths of stay. They find that the LLM not only classified cases accurately but also provided an explanation justifying why, for example, a specific patient was categorized as ASA class 3 but not class 4. By contrast, the LLM struggled to make accurate predictions of linear outcomes like length of stay.

Related article page 928

By focusing on familiar perioperative outcomes, this study serves as an approachable case study that provides the motivated reader with an accessible primer on how LLMs work, what sophisticated prompting strategies look like, the kind of outcomes that can be expected, and the way performance changes with prompting strategy. As such, it will be of interest to a wide range of surgeons and investigators eager to ex-