## Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication

Philip Chung<sup>1</sup>, Christine T. Fong<sup>2</sup>, Andrew M. Walters<sup>2</sup>, Nima Aghaeepour<sup>1</sup>, Meliha Yetisgen<sup>3,4</sup>, Vikas N. O'Reilly-Shah<sup>2</sup>

<sup>1</sup>Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University.

<sup>2</sup>Department Anesthesia & Pain Medicine, University of Washington.
<sup>3</sup>Department of Biomedical Informatics and Medical Education, University of Washington.

<sup>4</sup>Department of Linguistics, University of Washington.

#### Abstract

We investigate whether general-domain large language models such as GPT-4 Turbo can perform risk stratification and predict post-operative outcome measures using a description of the procedure and a patient's clinical notes derived from the electronic health record. We examine predictive performance on 8 different tasks: prediction of ASA Physical Status Classification, hospital admission, ICU admission, unplanned admission, hospital mortality, PACU Phase 1 duration, hospital duration, and ICU duration. Few-shot and chain-of-thought prompting improves predictive performance for several of the tasks. We achieve F1 scores of 0.50 for ASA Physical Status Classification, 0.81 for ICU admission, and 0.86 for hospital mortality. Performance on duration prediction tasks were universally poor across all prompt strategies. Current generation large language models can assist clinicians in perioperative risk stratification on classification tasks and produce high-quality natural language summaries and explanations.

**Keywords:** anesthesia, surgery, perioperative medicine, artificial intelligence, large language models, electronic health records, clinical informatics

## 1 Introduction

Instruction-tuned large language models (LLMs) have been successful at knowledge retrieval[1–4], text extraction [5–8], summarization[9–11], and reasoning[12–16] tasks without requiring domain-specific fine-tuning. Prompting has emerged as a means for task and domain specification without requiring domain-specific fine-tuning[17]. General-domain models have been shown to excel at medicine-specific tasks such as United States Medical Licensing Exam (USMLE) questions[18–20] or summarization of electronic health record text.[21]

We investigate whether general-domain LLMs can perform the task of preoperative risk stratification and prognostication—that is, given information about the patient and a surgery or procedure, assign a risk score or predict a post-operative outcome metric. This assessment of post-operative outcome assists proceduralists, surgeons, and anesthesiologists weighing the risks and benefits of proceeding with a procedure versus canceling or delaying the procedure for medical optimization.

Since there is no single post-operative outcome measure of risk, we survey LLM capabilities on 8 different tasks: (1) assignment of the American Society of Anesthesiologists Physical Status (ASA-PS) classification system[22–24], (2) prediction of post-anesthesia care unit (PACU) duration, (3) prediction of hospital admission, (4) prediction of hospital duration, (5) prediction of intensive care unit (ICU) admission, (6) prediction of ICU duration, (7) prediction of whether the patient will have an unanticipated hospital admission, (8) prediction of whether the patient will die in the hospital. We only examine outcomes where quantitative ground truth labels can be derived from the electronic health record (EHR). We also explore the effect of commonly applied prompting techniques such as in-context (few-shot) learning and chain-of-thought (CoT) reasoning, which has been shown to improve question and answering (Q&A) performance. In-context learning involves adding representative task & solution examples into the prompt prior to the actual query task to demonstrate the desired pattern of task and response [1]. CoT is a prompting strategy that instructs language models to respond with step by step reasoning prior to providing a final answer [13, 14].

While LLMs have been previously explored in medical Q&A with models attaining a passing score on the USMLE medical licensing exam, these Q&A datasets are not reflective of the real-world clinical setting[18–20]. Most of these datasets are multiple choice questions or have unambiguous answers that exist within a well-defined knowledge source such as a medical text book[25, 26]. Real-world EHRs often contain patient contexts with uncertain, incomplete, or erroneous information, and a clear answer may be elusive. It is within this real-world context that we derive our dataset and conduct our investigations to benchmark the capabilities of LLMs in perioperative risk prediction and prognostication.

## 2 Methods

This is a retrospective study of routinely collected health records data, approved by the University of Washington (UW) Institutional Review Board with a waiver of consent. The computational environment for use of protected health information (PHI) and personally identifiable information (PII) 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) guideline [27].

#### 2.1 Study Cohort and Dataset Definition

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, WA from April 1, 2021 to May 5, 2023 where the patient also had an anesthesia preoperative evaluation note associated with the case and at least one other clinician note filed in the EHR prior to the case. The notes used in our experiment included up to the last 10 clinician-written notes filed in the EHR prior to the surgery, excluding the anesthesia preoperative evaluation note associated with the case. Short notes less than 100 token lengths were excluded. PACU phase 1 timestamps and Admit/Discharge/Transfer (ADT) events for these patients were used to determine postoperative PACU, hospital, and ICU admission and length of stay duration. Unanticipated admission was determined by comparing expected patient class in the surgeon's case booking and actual postoperative patient class.

From the extracted cases, notes, and ADT events, seven datasets were created which were used for each of the eight prediction tasks (dataset for ICU admission and duration prediction are the same). For patients who had multiple cases, a single case was randomly selected. The case number was then downsampled to a target of 1250 cases and then split 80%-20% into inference and few-shot data splits for 1000 inference cases and 250 few-shot cases. The few-shot data splits are held-out

sets that are only used for crafting in-context examples for few-shot prompting. We employed the following strategies to combat class imbalance and data skew in the outcome variables: (1) The ASA-PS dataset was constructed with inverse frequency sampling of the ASA-PS; (2) PACU and hospital durations were binned to form 20 and 100 groups, respectively, and the datasets were constructed with inverse frequency sampling of the groups; (3) For hospital admission, the outer product of whether a patient was admitted and the patient class was used to form groups, and then dataset was constructed using inverse frequency sampling of the groups; (4) For ICU admission, ICU duration, unplanned admission, and hospital mortality datasets, the outcome variables were rare occurrences, so we sampled a 50%-50% mix of cases with the presence and absence of ICU admission, unplanned admission, and hospital mortality for each respective dataset.

## 2.2 Experimental Approach

An overview of the experimental apparatus is shown in Figure 1. We used GPT-4 Turbo (gpt-4-1106)[28] as our LLM through Microsoft Azure OpenAI Service with disabled content filters because of the presence of PHI/PII. Temperature and top p parameters were set to 1.0 for all experiments. A unique random seed was used for each case in the datasets. If an LLM generation was unsuccessful or the response was inappropriately formatted and cannot be extracted, the LLM generation was retried with a different random seed up to 5 times. For each of the 8 prediction tasks, we experiment with 6 prompting strategies: (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 &A using original notes, (5) zero-shot CoT Q&A using note summaries, (6) few-shot CoT Q&A using

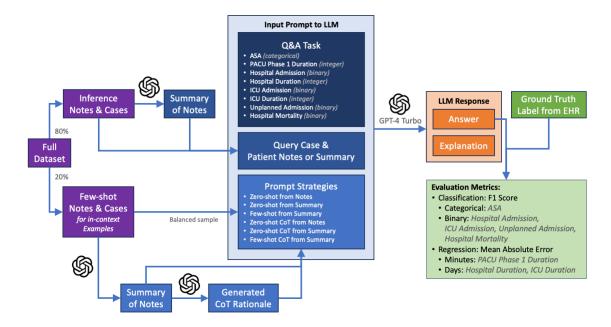


Fig. 1 Overview of the experimental apparatus. GPT-4 Turbo is used as the large language model (LLM) in all steps. Each prompt to the LLM is unique based on the task, prompt strategy and query case for which an answer and explanation is generated. Zero-shot prompt strategy is conducted with both original clinical notes and a summary of the clinical notes. Few-shot prompts utilize in-context examples derived from the few-shot dataset. Each in-context example is a question, procedure description, summary of patient notes, and answer. Summaries are generated using LLM. The few-shot chain-of-thought (CoT) prompt strategy requires a CoT rationale for each in-context example that links the question to the answer, which is also generated using LLM. Answers provided by the LLM are compared against the ground truth label derived from electronic health record (EHR) data, and either F1 score or mean absolute error (MAE) is computed, depending on whether the outcome variable for the task is categorical/binary or integer.

note summaries. For few-shot prompts, we used 5-shot, 10-shot, 20-shot, and 50-shot in-context examples. Each prompt contained an instruction, task/question, procedure information (procedure, procedure description, diagnosis, provider service), patient representation (original notes or note summaries), and instruction on how to format the response. Note summaries were generated by GPT-4 Turbo. Few-shot prompt in-context examples were drawn only from the few-shot dataset, using inverse frequency sampling to balance the representation of outcome variables in the incontext examples. For few-shot CoT Q&A prompts, each in-context example had a CoT rationale generated by GPT-4 Turbo specific to the in-context example and task. When CoT prompting was used, the LLM was instructed to provide a step-by-step explanation prior to the answer to allow the LLM to attend to the explanation while generating the answer. When CoT prompting was not used, the LLM was instructed to generate an answer and then provide an explanation. Representative text prompts used for experiments are depicted in Supplemental Figure B.1.

We applied each of the prompting strategies to the 8 tasks. For each example, the answer was compared against the ground truth label derived from the EHR. F1 score was the primary measure of performance for categorical and binary outcomes and mean absolute error (MAE) was the primary measure of performance for duration outcomes. ASA-PS prediction was a 6-class prediction task whereas hospital admission, ICU admission, unplanned admission, and hospital mortality prediction tasks were binary classification tasks. PACU Phase 1 duration was measured in minutes whereas hospital duration and ICU duration were measured in days. Reported metrics and 95% confidence intervals were estimated using 2500 bootstrap iterations. We also stratified the cases based on token length of all 10 notes into three categories—short, medium, and long—and reported the performance of each strata. Statistical significance testing was performed using Wilcoxon signed-rank test for pairwise comparison of prompt strategies.

## 3 Results

#### 3.1 Datasets

Our inclusion criteria yielded 90,844 patients undergoing 137,535 surgeries or procedures with anesthesia care, 133,500 anesthesia preoperative evaluation notes, 226,821 other clinician-written notes, and 2,253,286 admit/discharge/transfer events, which was then processed and downsampled to construct the task-specific datasets used for our experiments. The hospital mortality and unplanned admission outcomes were rare and resulted in datasets with 720 and 1186 cases prior to the 80%-20% data split. All other datasets reached the target size of 1250 cases followed by 80%-20% split for 1000 inference cases and 250 few-shot cases per task. Demographics, distribution of input and outcome variables for each dataset and data split are described in Table 1. Details to dataset creation is shown in Supplemental Figure B1. The degree of overlap of cases in each of the datasets is shown in Supplemental Figure B2. Counts for note type and author provider type are shown in Supplemental Table A1. The direct Microsoft Azure API costs of conducting these experiments was \$6000 (Supplemental Table A2).

## 3.2 Effect of Prompt Strategy on Perioperative Risk Prediction Tasks

Performance of each prompt strategy for each of the perioperative risk prediction tasks is summarized in Figure 2. A detailed breakdown of performance metrics with confidence intervals is shown in Supplemental Tables A3-A23. This includes F1, Matthew's Correlation Coefficient (MCC), Sensitivity, Specificity, PPV, NPV for classification tasks, and MAE for regression tasks. The difference between nearly all prompt strategies is statistically significant (Supplemental Figures B3-B10). GPT-4 Turbo provided a valid answer to all prompts except for 1 case in the unplanned admission dataset, which was excluded from our evaluation metrics. During dataset creation, most clinical notes written for administrative purposes such as attending attestations or for billing purposes were found to be less than 100 tokens in length and were removed. However, some of these notes were

still present in the final dataset and in cases where the sole note did not contain clinically-relevant content as in the case for the unplanned admission dataset, GPT-4 Turbo correctly identified this situation and refused to make a prediction across all 5 allotted retries.

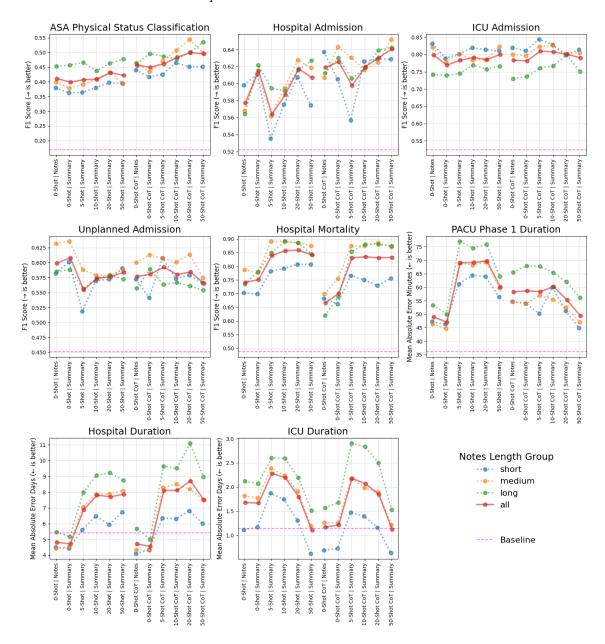


Fig. 2 Performance for the 8 perioperative prediction tasks. X-axis shows the different prompt strategies with the first six without chain-of-thought reasoning and the second six with chain-of-thought reasoning. "Notes" indicates that original clinical notes were inserted into the prompt whereas "Summary" indicates that clinical notes were first summarized using GPT-4 Turbo and then the summary was inserted into the prompt. All in-context examples for few-shot prompts used note summaries. Y-axis is F1 Score for classification tasks where higher score is better, and Mean Absolute Error for regression tasks where lower error is better. Baseline for classification tasks represent score achieved by random guessing. Baseline for regression tasks represent the MAE achieved by a regressor that always predicts the mean value in the dataset. The clinical notes are stratified into short, medium, and long length groups which represent the  $\frac{1}{3}$  shortest,  $\frac{1}{3}$  middle, and  $\frac{1}{3}$  longest notes in the dataset and performance is shown for each stratification.

ASA-PS, hospital admission, ICU admission, unplanned admission, and hospital mortality are classification tasks with performance measured using F1 score where a higher score corresponds to better performance. The baseline presented in each plot in Figure 2 for these outcomes represents the score for random guessing. All prompt strategies outperformed the random baseline for classification tasks. CoT resulted in significant performance gains for predicting ASA-PS and hospital admission, but not ICU admission, unplanned admission, and hospital mortality. Figure 3 depicts a prompt and LLM output for ASA-PS prediction using CoT, illustrating the LLM's rich understanding of the task in the context of the patient and procedure. The logical reasoning is easily understood by any clinician and can rapidly assist a clinician with perioperative evaluation of the patient. Prompts and LLM outputs for all prompt strategies are shown in Supplemental Figure B.1 to illustrate how each prompt strategy affects the LLM output. Few-shot prompting showed consistent performance gains with ASA-PS, hospital admission, and hospital mortality. Gains for CoT and few-shot prompting were synergistic for ASA-PS and hospital admission, but for hospital mortality, CoT resulted in slightly worse performance. ICU admission performance was not significantly affected by the prompt strategies, but all prompt strategies had high prediction performance, suggesting the LLM is easily able to perform this task. Zero-shot without CoT was best performing for unplanned admission and demonstrated that rationalizing does not help with all prediction tasks.

PACU phase 1 duration, hospital duration, and ICU duration are regression tasks with performance measured using MAE where a lower score corresponds to better performance. The baseline is the MAE for a dummy regressor that always predicts the mean duration in the dataset. Most prompt strategies performed worse than this baseline. For PACU phase 1 duration prediction, all

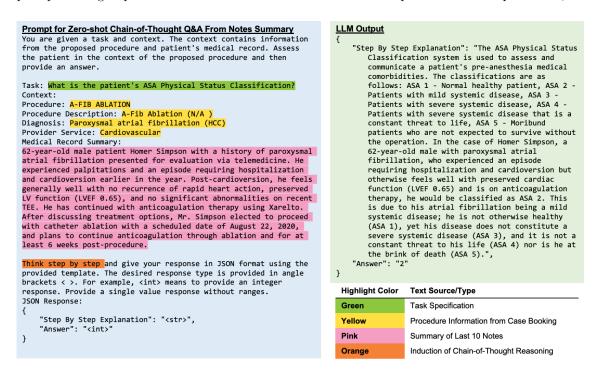


Fig. 3 Prompt and LLM Output for Zero-shot chain-of-thought Q&A from notes summary prompt strategy. Note summaries are generated from raw clinical notes from the LLM prior to insertion into the prompt. The LLM output shows that the LLM understands the definition for ASA Physical Status Classification (ASA-PS) and provides a valid rationale for which ASA-PS class the patient should be classified. All prompt strategies using this patient and procedure example are depicted in Supplemental Figure 1. While the content of this example is derived from a real patient and case from the electronic health record, all PHI and PII are removed with names obfuscated, and dates and times shifted.

prompt strategies performed worse than the dummy regressor, suggesting that the LLM struggled with this task. For hospital duration prediction, zero-shot and zero-shot CoT slightly outperformed the baseline dummy regressor with an error of 4.5 days compared to 5.4 days, which is not a clinically significant difference. Addition of few-shot prompting worsened hospital duration prediction, despite few-shot and CoT improving hospital admission classification, which is the analogous binary prediction task. For ICU duration prediction, few-shot and CoT did help with predictions, but were only able to achieve parity with the baseline dummy regressor. This is in contrast to the analogous binary prediction task, ICU admission prediction, for which the LLM far exceeded its baseline.

The best performing prompt strategies for each prediction task were:

- 1. ASA-PS: 20-shot CoT and 50-shot CoT, both which achieved F1 of 0.50.
- 2. Hospital admission: 50-shot CoT with F1 of 0.64.
- 3. ICU admission: 5-shot CoT prompt strategy with F1 score of 0.81.
- 4. Unplanned admission: Zero-shot using note summaries with F1 score of 0.61.
- 5. Hospital mortality: 10-shot and 20-shot with F1 score of 0.86.
- 6. PACU phase 1 duration: Zero-shot using original notes with MAE of 49 minutes.
- 7. Hospital duration: Zero-shot CoT using notes summary with MAE of 4.5 days.
- 8. ICU duration: Both 50-shot and 50-shot CoT with MAE of 1.1 days.

### 3.3 Effect of Summary Representation of Notes

Prior work has shown that LLM-generated summaries in the clinical domain may be preferable to human-written summaries[21]. Comparison of zero-shot prompts using original notes versus 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 resulted in a boost in performance for hospital admission, unplanned admission, and hospital mortality prediction. The magnitude of these effects were small indicating that while summaries sometimes resulted in slight loss of information useful for the predictive task, summaries may also help focus relevant information. However, a distinct advantage in using a summary representation of patient history was the ability to scale to a large number of in-context examples when using few-shot prompting, such as the 50-shot prompt strategies. In several tasks, this resulted in significantly better predictive performance, but summary representations of patient histories were necessary to compress patient notes so they fit within the input context of the LLM.

#### 3.4 Effect of Note Length on Perioperative Risk Prediction Tasks

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

#### 3.5 Numerical Prediction Tasks

Performance on numerical predictions (PACU phase 1 duration, hospital duration, icu duration) was poor. LLM predictions were often worse than simply guessing the mean value of the dataset. However, when the same task was recast as a binary prediction as in the case of hospital admission or ICU admission prediction, LLM prediction performance was significantly better. Visualizing PACU phase 1 duration predictions in Figure 4 revealed that without few-shot and CoT prompting, LLMs tend to predict quantized outputs, often with a ceiling effect. Few-shot and CoT prompting helps remove these effects, but results in worse performance. Similar findings were seen for hospital

duration prediction (Supplemental Figure B17) and ICU duration prediction (Supplemental Figure B18).

## 4 Discussion

Our investigations indicate that general-domain LLMs such as GPT-4 Turbo can be applied to clinical notes and procedure description to achieve perioperative risk assessment and prognostication. As Nori et al.[19] has observed, the addition of few-shot prompting and chain-of-thought prompting helps boost performance, though not unilaterally across all tasks.

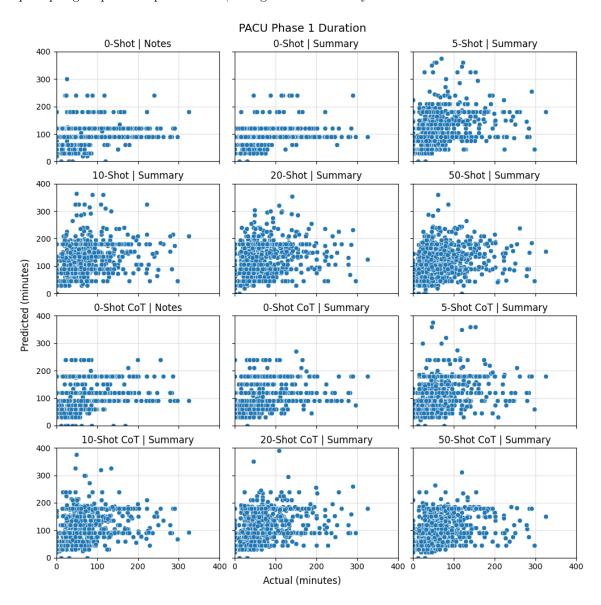


Fig. 4 Scatter plot of predicted and actual post-anesthesia care unit (PACU) Phase 1 recovery durations across all 12 prompt strategies. Without few-shot and CoT prompting, predictions are heavily quantized to specific values and exhibit a ceiling effect where the LLM rarely predicts beyond 180 minutes. The progressive addition of few-shot and CoT prompting removes this effect, but predictive performance remains poor.

We observe strong overall performance for prediction of postoperative ICU admission and hospital mortality across all prompt strategies. ICU admission prediction exhibits high sensitivity  $\geq 0.78$  and specificity  $\geq 0.74$  across all prompt strategies (Supplemental Table). Hospital mortality prediction can be enhanced using few-shot prompting techniques and exhibits high sensitivity  $\geq 0.93$  across all prompt strategies with highest specificity of 0.81 from 20-shot prompt strategy (Supplemental Table A20). We also observe good predictive performance for ASA-PS. ASA-PS assignment is known to be subjective and has only moderate interrater agreement among human anesthesiologists [29, 30], so it is unlikely that any prediction system can achieve a perfect score. In this context, a multiclass ASA-PS F1 micro score of 0.5 has meaningful clinical utility (Supplemental Table A3); confusion matrices also show that ASA-PS misclassifications made by the LLM with few-shot and CoT are almost always an adjacent ASA class (Supplemental Figure B11). Hospital admission prediction and unplanned admission prediction performance is better than random guessing, but not as impressive as outcome measures like ICU admission and hospital mortality where the illness severity of a patient is likely more apparent and it is easier for the LLM to make prediction. Still, it is remarkable that GPT-4 Turbo can achieve this kind of 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 chain-of-thought prompting reveal significant gains in predictive performance in tasks where synthesizing prior clinical knowledge is important, such as determination of ASA-PS, hospital admission prediction, and hospital mortality. These effects are additive and synergistic but the benefits of these prompting techniques do not apply to all outcomes. We believe the aforementioned prediction tasks benefit from the prompting strategies because they are heavily dependent on preoperative illness severity, which would be reflected in a patient's clinical notes. In-context examples will help the LLM compare and contrast among similar cases whereas CoT rationales will help expand upon the concepts mentioned in clinical notes, both of which will guide the LLM toward more accurate predictions. In contrast, we suspect that these gains are not seen in outcomes such as unplanned admission because factors leading to unexpected admission are predominantly due to intraoperative factors which are not accounted for in pre-operative clinical notes and procedure booking data presented to the LLM, and no amount of deliberation or rationalization would affect the outcome.

In their current incarnation, LLMs struggle with regression tasks involving prediction of continuous or numerical outcomes such as PACU phase 1 duration, hospital duration, and ICU duration. Our analysis of duration prediction tasks shows that LLMs tend to quantize values, which we suspect is due to the LLM memorizing length of stay estimations from hospital websites, textbooks, and journal articles (Figure 4, Supplemental Figure B17). In-context demonstrations and prompting the LLM to rationalize about the patient's procedure and medical history helps overcome this quantization phenomenon, but we believe the continued poor results to be attributed to the architectural design of LLMs. Namely, LLMs enforce a discrete tokenized output where each token's representation is primarily derived from text contexts. For continuous-valued outcomes, it is meaningful to be able to interpolate between numerical values but a LLM's training data and training process does not provide a robust way for the model to learn this concept. 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 [31–34]. While many visual-language [35–37] and multimodal models adopt these strategies to combine text and other data modalities such as pixel intensity in the same model, no widely available model has yet employed these solutions for general numerical predictions. Future foundational models for healthcare or EHR data should consider model architectures and pretraining routines that enable better performance for these kind of numerical prediction tasks. Another alternative is equipping LLMs with tool use [38-41], 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 capability for information understanding and synthesis to make a prediction.

Our studies in stratifying prediction performance on note length groups and comparing zeroshot predictions from original notes versus LLM-generated summaries indicate that longer input contexts do not necessarily result in better performance, though the specific effects are task-specific. This is contrary to the intuition that providing the LLM more clinical context and information would enable a more accurate prediction. We believe that this is because increased note length from the EHR may be the result of automatically copied templates that do not substantially add knowledge over a carefully crafted clinical vignette statement. Conversely, we qualitatively find LLM-generated summaries to be high-quality (Figure 3, Supplemental Figure B.1) which concords with prior studies showing LLMs to have excellent clinical summarization capabilities. Furthermore, prior work has shown that many current LLMs are biased to pay more attention to the beginning and end of prompts, which may be an artifact of training data not containing long-context prediction or retrieval tasks[42]. This would explain why some tasks perform better when note summaries are used instead of original notes. Similar to Nori et al. [19, 20], we observe GPT-4 Turbo generates high-quality CoT rationalizations and answer explanations. We also find the LLM appropriately refuses to answer when the context is irrelevant to the task, which is a desirable property in clinical use cases.

Overall, our results indicate that currently available general-domain LLMs can have direct clinical impact to perioperative risk stratification workflows in hospitals and can be used to assist in stratifying the preoperative patient population for these outcomes. In particular, we find LLMs to exhibit very good performance at ASA Physical Status Classification prediction, ICU admission prediction, and Hospital Mortality prediction. LLMs underperform dedicated classification models utilizing tabular features [43–56] and if standalone automated triage and prediction is the goal, it is unlikely current generation LLMs will perform better. However traditional machine learning learning models are rarely utilized in the clinical setting because of difficulty in interpreting a model's predictions. LLMs are unique in their ability to present natural language explanations understandable to human clinicians, and thus have significant clinical utility in summarizing the patient history and creating rationalizations against each outcome variable of interest. These explanations are valuable starting points for clinicians to perform a more comprehensive perioperative risk assessment, and are more useful clinically than standalone risk predictions.

Future research is needed to evaluate whether clinical domain-specific language models [6, 18, 57, 58], trained on clinical or EHR text shows improved performance or more advanced few-shot prompting strategies such as dynamic kNN few-shot, shuffling few-shot examples, ensembling, or retrieval-augmentation [19, 25, 26, 59, 60]. Our investigations are centered around whether LLMs can perform these tasks and use outcome-balanced datasets to enable measurement of these rare outcomes. Real-world incidence of these outcomes is rare and of the 137535 cases considered in the 2 year span from which our datasets are derived, only 0.49% of cases have associated postoperative ICU admissions, 0.43% of cases had an unanticipated hospital admission, and 0.3% of cases have associated postoperative in-hospital mortality (Figure B1. Future large-scale prospective clinical validation is necessary to verify our observed performance, especially for rare outcomes, and to compare against existing perioperative prediction algorithms [55, 56].

## 5 Conclusion

General-domain text-only large language models are capable of perioperative risk prediction and prognostication when framed as classification or binary prediction tasks, but still fall short in being able to predict continuous-valued outcomes such as PACU, hospital, and ICU length of stay. Few-shot prompting and chain-of-thought reasoning improves prediction performance for perioperative prediction tasks. Hospitals should evaluate the effectiveness of using large language models as tools to assist perioperative risk stratification.

## **Declarations**

Acknowledgments. The authors would like to acknowledge: 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 and Robert Fabiano from University of Washington Research IT for creating a digital research environment within the Microsoft Azure Cloud where experiments were performed, 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.

Funding. 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 (NIH 5T32GM089626), and NIH R35GM138353. Funding sources were not involved in design of the study, collection of data, analysis of data, interpretation of data or in writing the manuscript.

Contributions. P.C. and V.O.R conceived of the initial research question. A.M.W. and M.Y. helped refine the research question, prompting strategies, and experimental approach. P.C. and C.T.F. refined inclusion and exclusion criteria and conducted data extraction, cleaning, and transformation into final datasets. P.C. wrote code and conducted experiments, analyzed data and wrote the manuscript. N.A., A.M.W., M.Y., and V.O.R provided feedback on initial results and refinement to experimental methodology and analysis. All authors reviewed and revised the paper.

Competing Interests. None of the authors have any financial or non-financial competing interests.

Data Availability. The raw data and datasets generated in our experiments are not publicly available because they are derived from electronic health records consisting of protected health information (PHI). Data may be requested by contacting Vikas O'Reilly-Shah at voreill@uw.edu or the University of Washington Center for Perioperative & Pain Initiatives in Quality Safety Outcome (PPiQSO) at PPiQSO@uw.edu. Data access is contingent upon signing a data use agreement in accordance with UW Medicine policy.

**Code Availability.** Code for experiments and analysis results are publicly available at: https://github.com/philipchung/llm-periop-prediction

## **Tables**

## Table 1: Dataset Characteristics

Table 1: Description of patient demographics, procedure specialty, patient class, note statistics, and outcome variables for each dataset: (A) American Society of Anesthesiologists Physical Status (ASA-PS) Classification, (B) Post-Anesthesia Care Unit (PACU) Phase 1 Duration, (C) Hospital Admission, (D) Hospital Duration, (E) ICU Admission & ICU Duration, (F) Unplanned Admission, (G) Hospital Mortality.

Α.			ASA Physical St	atus Dataset
			Inference	Fewshot
<b>.</b>	Demographics	Case Counts, no. (%)	1000	250
Patient Info		Age, mean (std)	51.9 (19.9)	51.0 (21.6)
		Female	447 (44.7%)	108 (43.2%)
	Sex, no. (%)	Male	553 (55.3%)	142 (56.8%)
		Other or Unknown	_	_
	Anesthesia Type,	General	842 (84.2%)	216 (86.4%)
	no. (%)	MAC	125 (12.5%)	22 (8.8%)
		Regional	33 (3.3%)	12 (4.8%)
	Actual Patient Class, no. (%)	Deceased - Organ Donor	10 (1.0%)	2 (0.8%)
		Emergency	3 (0.3%)	1 (0.4%)
		Inpatient	513 (51.3%)	120 (48.0%)
		Observation	39 (3.9%)	10 (4.0%)
		Outpatient	295 (29.5%)	77 (30.8%)
		Surgery Admit	4 (0.4%)	_
		Surgery Overnight Stay	136 (13.6%)	40 (16.0%)
Case		Emergency	59 (5.9%)	13 (5.2%)
nfo	Expected Patient Class, no. (%)	Inpatient	265 (26.5%)	66 (26.4%)
		Outpatient	133 (13.3%)	37 (14.8%)
		Surgery Admit	175 (17.5%)	36 (14.4%)
		Surgery Overnight Stay	260 (26.0%)	69 (27.6%)
		Unspecified	108 (10.8%)	29 (11.6%)
		Burn Surgery	11 (1.1%)	_
		Cardiovascular Cath Lab	78 (7.8%)	23 (9.2%)
		Cardiovascular Surgery	35 (3.5%)	8 (3.2%)
		Gastroenterology	58 (5.8%)	13 (5.2%)
		General Surgery	157 (15.7%)	42 (16.8%)
	Proceduralist	Gynecologic Surgery	46 (4.6%)	8 (3.2%)
	Service, no. (%)	Hematology/ Oncology	1 (0.1%)	_
		Interventional Radiology	50 (5.0%)	16 (6.4%)

Table 1-A continued from previous page

Α.			ASA Physical St	atus Dataset
		Neurosurgery	82 (8.2%)	14~(5.6%)
		Obstetrics	13 (1.3%)	6 (2.4%)
		Ophthalmology	30 (3.0%)	16 (6.4%)
		Oral-Maxillofacial Surgery	25 (2.5%)	5 (2.0%)
		Organ Donor	10 (1.0%)	2 (0.8%)
		Orthopedic Surgery	219 (21.9%)	55 (22.0%)
		Otolaryngology	49 (4.9%)	16 (6.4%)
		Plastic Surgery	19 (1.9%)	6 (2.4%)
		Podiatry	7 (0.7%)	3 (1.2%)
		Thoracic Surgery	11 (1.1%)	3 (1.2%)
		Transplant Surgery	14 (1.4%)	2 (0.8%)
		Urology	58 (5.8%)	8 (3.2%)
		Vascular Surgery	27 (2.7%)	4 (1.6%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (3, 7)	4 (2, 6)
		Most Recent Clinical Notes, total number of tokens	4595 (2342, 7933)	4031 (1991, 7726)
	ASA Physical Status, no. (%)	1	218 (21.8%)	53 (21.2%)
		2	232 (23.2%)	57 (22.8%)
		3	212 (21.2%)	60 (24.0%)
		4	224 (22.4%)	56 (22.4%)
		5	105 (10.5%)	22 (8.8%)
Outcome		6	9 (0.9%)	2 (0.8%)
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	_	-
	Hospital	Yes	_	_
	Admission, no. (%)	No	_	_
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	_	_
	ICU Admission,	Yes	_	_
	no. (%)	No	_	_
	ICU Duration, median (IQR)	ICU Duration, days	-	-
	Unplanned	Yes	20 (2.0%)	1 (0.4%)
	Admission, no. (%)	No	980 (98.0%)	249 (99.6%)
	Hospital	Yes	64 (6.4%)	16 (6.4%)
	Mortality, no. (%)	No	936 (93.6%)	234 (93.6%)

В.			PACU Phase 1	Duration Dataset
			Inference	Fewshot
<b>.</b>	Demographics	Case Counts, no. (%)	100	0 250
Patient Info		Age, mean (std)	50.0 (19.5)	50.5 (19.1)
	G (M)	Female	494 (49.4%)	141 (56.4%)
	Sex, no. (%)	Male	506 (50.6%)	109 (43.6%)
		Other or Unknown	_	_
	Anesthesia Type,	General	867 (86.7%)	209 (83.6%)
	no. (%)	MAC	69 (6.9%)	18 (7.2%)
		Regional	64 (6.4%)	23 (9.2%)
		Deceased - Organ Donor	_	-
	Actual Patient	Emergency	2 (0.2%)	1 (0.4%)
	Class, no. (%)	Inpatient	333 (33.3%)	90 (36.0%)
		Observation	69 (6.9%)	16 (6.4%)
		Outpatient	373 (37.3%)	89 (35.6%)
		Surgery Admit	3 (0.3%)	1 (0.4%)
		Surgery Overnight Stay	220 (22.0%)	53 (21.2%)
Case	Expected Patient Class, no. (%)	Emergency	16 (1.6%)	5 (2.0%)
nfo		Inpatient	180 (18.0%)	52 (20.8%)
		Outpatient	180 (18.0%)	40 (16.0%)
		Surgery Admit	180 (18.0%)	45 (18.0%)
		Surgery Overnight Stay	381 (38.1%)	92 (36.8%)
		Unspecified	63 (6.3%)	16 (6.4%)
		Burn Surgery	10 (1.0%)	_
		Cardiovascular Cath Lab	25 (2.5%)	7 (2.8%)
		Cardiovascular Surgery	_	Fewshot  50.5 (19.1)  141 (56.4%)  109 (43.6%)  -  209 (83.6%)  18 (7.2%)  23 (9.2%)  -  1 (0.4%)  90 (36.0%)  16 (6.4%)  89 (35.6%)  1 (0.4%)  53 (21.2%)  5 (2.0%)  52 (20.8%)  40 (16.0%)  45 (18.0%)  92 (36.8%)  16 (6.4%)  -
		Gastroenterology	43 (4.3%)	11 (4.4%)
		General Surgery	142 (14.2%)	43 (17.2%)
	Proceduralist Service, no. (%)	Gynecologic Surgery	67 (6.7%)	21 (8.4%)
	Service, no. (70)	Hematology/ Oncology	-	-
		Interventional Radiology	20 (2.0%)	5 (2.0%)
		Neurosurgery	80 (8.0%)	14 (5.6%)
		Obstetrics	27 (2.7%)	13 (5.2%)
		Ophthalmology	56 (5.6%)	9 (3.6%)
		Oral-Maxillofacial Surgery	26 (2.6%)	6 (2.4%)
		Organ Donor	_	_
		Orthopedic Surgery	259 (25.9%)	77 (30.8%)
		Otolaryngology	71 (7.1%)	16 (6.4%)

Table 1-B continued from previous page

В.			PACU Phase 1 I	Ouration Dataset
		Plastic Surgery	20 (2.0%)	6 (2.4%)
		Podiatry	8 (0.8%)	_
		Thoracic Surgery	10 (1.0%)	2 (0.8%)
		Transplant Surgery	10 (1.0%)	_
		Urology	120 (12.0%)	19 (7.6%)
		Vascular Surgery	6 (0.6%)	1 (0.4%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (3, 7)	4 (3, 6)
		Most Recent Clinical Notes, total number of tokens	4327 (2183, 7131)	4367 (2461, 6924)
	ASA Physical Status, no. (%)	1	192 (19.2%)	42 (16.8%)
		2	494 (49.4%)	133 (53.2%)
		3	275 (27.5%)	67 (26.8%)
		4	38 (3.8%)	8 (3.2%)
		5	1 (0.1%)	_
Outcome		6	_	_
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	52 (31, 81)	52 (32, 79)
	Hospital	Yes	461 (46.1%)	122 (48.8%)
	Admission, no. (%)	No	539 (53.9%)	128 (51.2%)
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	0.00 (0.00, 2.00)	0.00 (0.00, 2.00)
	ICU Admission,	Yes	23 (2.3%)	4 (1.6%)
	no. (%)	No	977 (97.7%)	246 (98.4%)
	ICU Duration, median (IQR)	ICU Duration, days	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
	Unplanned	Yes	18 (1.8%)	4 (1.6%)
	Admission, no. (%)	No	982 (98.2%)	246 (98.4%)
	Hospital	Yes	_	1 (0.4%)
	Mortality, no. (%)	No	1000 (100.0%)	249 (99.6%)

5.			Hospital Admission Dataset		
			Inference	Fewshot	
D 4: 4	Demographics	Case Counts, no. (%)	1000	250	
Patient Info		Age, mean (std)	50.5 (19.3)	49.3 (20.8)	
	G (04)	Female	439 (43.9%)	116 (46.4%)	
	Sex, no. (%)	Male	560 (56.0%)	134 (53.6%)	
		Other or Unknown	1 (0.1%)	-	
	Anesthesia Type,	General	826 (82.6%)	210 (84.0%)	
	no. (%)	MAC	106 (10.6%)	24~(9.6%)	
		Regional	68 (6.8%)	16 (6.4%)	
		Deceased - Organ Donor	-	_	
	Actual Patient	Emergency	74 (7.4%)	18 (7.2%)	
	Class, no. (%)	Inpatient	$224\ (22.4\%)$	50 (20.0%)	
		Observation	220 (22.0%)	59 (23.6%)	
		Outpatient	251 (25.1%)	63 (25.2%)	
		Surgery Admit	82 (8.2%)	24~(9.6%)	
		Surgery Overnight Stay	149 (14.9%)	36 (14.4%)	
Case	Expected Patient Class, no. (%)	Emergency	67 (6.7%)	17 (6.8%)	
nfo		Inpatient	207 (20.7%)	59 (23.6%)	
		Outpatient	92 (9.2%)	23 (9.2%)	
		Surgery Admit	200 (20.0%)	46 (18.4%)	
		Surgery Overnight Stay	327 (32.7%)	79 (31.6%)	
		Unspecified	107 (10.7%)	26 (10.4%)	
		Burn Surgery	7 (0.7%)	1 (0.4%)	
		Cardiovascular Cath Lab	34 (3.4%)	11 (4.4%)	
		Cardiovascular Surgery	_	5 (19.3)       49.3 (20.8)         1 (43.9%)       116 (46.4%)         0 (56.0%)       134 (53.6%)         0.1%)       -         5 (82.6%)       210 (84.0%)         6 (82.6%)       16 (6.4%)         6 (8.8%)       16 (6.4%)         6 (22.4%)       50 (20.0%)         6 (22.4%)       50 (20.0%)         6 (22.0%)       59 (23.6%)         6 (25.1%)       63 (25.2%)         (8.2%)       24 (9.6%)         6 (14.9%)       36 (14.4%)         7 (20.7%)       59 (23.6%)         9 (23.6%)       17 (6.8%)         7 (20.7%)       59 (23.6%)         9 (23.6%)       23 (9.2%)         9 (20.0%)       46 (18.4%)         7 (32.7%)       79 (31.6%)         7 (10.7%)       26 (10.4%)         9 (7.7%)       1 (0.4%)         10 (1.4%)       11 (4.4%)         10 (2.5%)       8 (3.2%)         14 (15.4%)       41 (16.4%)         15 (6.0%)       15 (6.0%)         10 (2.5%)       3 (1.2%)         (2.5%)       3 (1.2%)         (3.2%)       7 (2.8%)         4 (23.8%)       68 (27.2%)	
		Gastroenterology	82 (8.2%)	18 (7.2%)	
		General Surgery	154 (15.4%)	41 (16.4%)	
	Proceduralist	Gynecologic Surgery	69 (6.9%)	15 (6.0%)	
	Service, no. (%)	Hematology/ Oncology	1 (0.1%)	_	
		Interventional Radiology	25 (2.5%)	8 (3.2%)	
		Neurosurgery	44 (4.4%)	14 (5.6%)	
		Obstetrics	22 (2.2%)	3 (1.2%)	
		Ophthalmology	75 (7.5%)	13 (5.2%)	
		Oral-Maxillofacial Surgery	32 (3.2%)	7 (2.8%)	
		Organ Donor	_	_	
		Orthopedic Surgery	238 (23.8%)	68 (27.2%)	
		Otolaryngology	81 (8.1%)	18 (7.2%)	

Table 1-C continued from previous page

C.			Hospital Admiss	ion Dataset
		Plastic Surgery	21 (2.1%)	5 (2.0%)
		Podiatry	3 (0.3%)	1 (0.4%)
		Thoracic Surgery	9 (0.9%)	_
		Transplant Surgery	_	2 (0.8%)
		Urology	93 (9.3%)	24 (9.6%)
		Vascular Surgery	10 (1.0%)	1 (0.4%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (2, 6)	4 (2, 7)
		Most Recent Clinical Notes, total number of tokens	3703 (2008, 6477)	4249 (2091, 7102)
	ASA Physical Status, no. (%)	1	187 (18.7%)	52 (20.8%)
		2	501 (50.1%)	125 (50.0%)
		3	290 (29.0%)	62 (24.8%)
		4	22 (2.2%)	11 (4.4%)
		5	_	_
Outcome		6	_	_
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	46 (24, 74)	45 (25, 74)
	Hospital	Yes	396 (39.6%)	104 (41.6%)
	Admission, no. (%)	No	604 (60.4%)	146 (58.4%)
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)
	ICU Admission,	Yes	7 (0.7%)	2 (0.8%)
	no. (%)	No	993 (99.3%)	248 (99.2%)
	ICU Duration, median (IQR)	ICU Duration, days	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
	Unplanned	Yes	10 (1.0%)	_
	Admission, no. (%)	No	990 (99.0%)	250 (100.0%)
	Hospital	Yes	_	_
	Mortality, no. (%)	No	1000 (100.0%)	250 (100.0%)

D.			Hospital Durati	ion Dataset
			Inference	Fewshot
	Demographics	Case Counts, no. (%)	1000	250
Patient Info		Age, mean (std)	56.1 (19.1)	57.9 (18.0)
	- (04)	Female	456 (45.6%)	131 (52.4%)
	Sex, no. (%)	Male	544 (54.4%)	119 (47.6%)
		Other or Unknown	_	_
	Anesthesia Type,	General	901 (90.1%)	229 (91.6%)
	no. (%)	MAC	52 (5.2%)	11 (4.4%)
		Regional	47 (4.7%)	10 (4.0%)
		Deceased - Organ Donor	_	_
	Actual Patient	Emergency	_	_
	Class, no. (%)	Inpatient	863 (86.3%)	230 (92.0%)
		Observation	37 (3.7%)	4 (1.6%)
		Outpatient	73 (7.3%)	12 (4.8%)
		Surgery Admit	_	_
		Surgery Overnight Stay	27 (2.7%)	4 (1.6%)
Case	Expected Patient Class, no. (%)	Emergency	33 (3.3%)	4 (1.6%)
nfo		Inpatient	491 (49.1%)	129 (51.6%)
		Outpatient	44 (4.4%)	9 (3.6%)
		Surgery Admit	270 (27.0%)	76 (30.4%)
		Surgery Overnight Stay	76 (7.6%)	15 (6.0%)
		Unspecified	86 (8.6%)	17 (6.8%)
		Burn Surgery	25 (2.5%)	4 (1.6%)
		Cardiovascular Cath Lab	18 (1.8%)	3 (1.2%)
		Cardiovascular Surgery	_	57.9 (18.0)  131 (52.4%)  119 (47.6%)  -  229 (91.6%)  11 (4.4%)  10 (4.0%)  -  230 (92.0%)  4 (1.6%)  12 (4.8%)  -  4 (1.6%)  129 (51.6%)  9 (3.6%)  76 (30.4%)  15 (6.0%)  17 (6.8%)  4 (1.6%)
		Gastroenterology	56 (5.6%)	10 (4.0%)
		General Surgery	122 (12.2%)	35 (14.0%)
	Proceduralist	Gynecologic Surgery	28 (2.8%)	8 (3.2%)
	Service, no. (%)	Hematology/ Oncology	_	_
		Interventional Radiology	30 (3.0%)	7 (2.8%)
		Neurosurgery	103 (10.3%)	25 (10.0%)
		Obstetrics	23 (2.3%)	6 (2.4%)
		Ophthalmology	18 (1.8%)	_
		Oral-Maxillofacial Surgery	6 (0.6%)	1 (0.4%)
		Organ Donor	_	
		Orthopedic Surgery	393 (39.3%)	105 (42.0%)
		Otolaryngology	25 (2.5%)	8 (3.2%)

Table 1-D continued from previous page

D.			Hospital Duratio	n Dataset
		Plastic Surgery	13 (1.3%)	4 (1.6%)
		Podiatry	6 (0.6%)	1 (0.4%)
		Thoracic Surgery	26 (2.6%)	8 (3.2%)
		Transplant Surgery	22 (2.2%)	5 (2.0%)
		Urology	75 (7.5%)	17 (6.8%)
		Vascular Surgery	11 (1.1%)	3 (1.2%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (3, 7)	4 (3, 7)
		Most Recent Clinical Notes, total number of tokens	5435 (2752, 8961)	5336 (2996, 8661)
	ASA Physical Status, no. (%)	1	77 (7.7%)	14 (5.6%)
		2	349 (34.9%)	89 (35.6%)
		3	500 (50.0%)	133 (53.2%)
		4	74 (7.4%)	14 (5.6%)
		5	_	_
Outcome		6	_	_
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	61 (38, 94)	65 (42, 91)
	Hospital	Yes	906 (90.6%)	238 (95.2%)
	Admission, no. (%)	No	94 (9.4%)	12 (4.8%)
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	6.00 (3.00, 10.00)	7.00 (4.00, 10.00)
	ICU Admission,	Yes	50 (5.0%)	19 (7.6%)
	no. (%)	No	950 (95.0%)	231 (92.4%)
	ICU Duration, median (IQR)	ICU Duration, days	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
	Unplanned	Yes	24 (2.4%)	8 (3.2%)
	Admission, no. (%)	No	976 (97.6%)	242 (96.8%)
	Hospital	Yes	4 (0.4%)	4 (1.6%)
	Mortality, no. (%)	No	996 (99.6%)	246 (98.4%)

Е.			ICU Admission	n & ICU Duration	Dataset
			Inference	Fewshot	
	Demographics	Case Counts, no. (%)	1000		250
Patient Info		Age, mean (std)	53.5 (18.8)	55.8 (18.5)	
	~ (~)	Female	492 (49.2%)	128 (51.2%)	
	Sex, no. (%)	Male	507 (50.7%)	122 (48.8%)	
		Other or Unknown	1 (0.1%)	_	
	Anesthesia Type,	General	862 (86.2%)	220 (88.0%)	
	no. (%)	MAC	109 (10.9%)	26 (10.4%)	
		Regional	29 (2.9%)	4 (1.6%)	
		Deceased - Organ Donor	_	_	
	Actual Patient	Emergency	3 (0.3%)	1 (0.4%)	
	Class, no. (%)	Inpatient	622 (62.2%)	160 (64.0%)	
		Observation	32 (3.2%)	7 (2.8%)	
		Outpatient	239 (23.9%)	63 (25.2%)	
		Surgery Admit	3 (0.3%)	_	
		Surgery Overnight Stay	101 (10.1%)	19 (7.6%)	
Case	Expected Patient Class, no. (%)	Emergency	31 (3.1%)	5 (2.0%)	
nfo		Inpatient	225 (22.5%)	47 (18.8%)	
		Outpatient	109 (10.9%)	29 (11.6%)	
		Surgery Admit	337 (33.7%)	89 (35.6%)	
		Surgery Overnight Stay	204 (20.4%)	50 (20.0%)	
		Unspecified	94 (9.4%)	30 (12.0%)	
		Burn Surgery	1 (0.1%)	3 (1.2%)	
		Cardiovascular Cath Lab	52 (5.2%)	13 (5.2%)	
		Cardiovascular Surgery	1 (0.1%)	1 (0.4%)	
		Gastroenterology	47 (4.7%)	10 (4.0%)	
		General Surgery	80 (8.0%)	26 (10.4%)	
	Proceduralist	Gynecologic Surgery	41 (4.1%)	10 (4.0%)	
	Service, no. (%)	Hematology/ Oncology	1 (0.1%)	-	
		Interventional Radiology	47 (4.7%)	20 (8.0%)	
		Neurosurgery	255 (25.5%)	65 (26.0%)	
		Obstetrics	16 (1.6%)	4 (1.6%)	
		Ophthalmology	46 (4.6%)	10 (4.0%)	
		Oral-Maxillofacial Surgery	15 (1.5%)	7 (2.8%)	
		Organ Donor	_	_	
		Orthopedic Surgery	193 (19.3%)	35 (14.0%)	
		Otolaryngology	68 (6.8%)	10 (4.0%)	

Table 1-E continued from previous page

Ε.			ICU Admission &	& ICU Duration Dataset
		Plastic Surgery	34 (3.4%)	8 (3.2%)
		Podiatry	5 (0.5%)	-
		Thoracic Surgery	9 (0.9%)	4 (1.6%)
		Transplant Surgery	7 (0.7%)	5 (2.0%)
		Urology	53 (5.3%)	14 (5.6%)
		Vascular Surgery	29 (2.9%)	5 (2.0%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (3, 7)	4 (2, 6)
		Most Recent Clinical Notes, total number of tokens	4318 (2390, 7542)	3716 (2168, 6662)
		1	112 (11.2%)	23 (9.2%)
	ASA Physical Status, no. (%)	2	401 (40.1%)	98 (39.2%)
		3	405 (40.5%)	106 (42.4%)
		4	78 (7.8%)	22 (8.8%)
		5	4 (0.4%)	1 (0.4%)
Outcome		6	_	-
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	58 (31, 92)	66 (34, 102)
	Hospital	Yes	681 (68.1%)	179 (71.6%)
	Admission, no. (%)	No	319 (31.9%)	71 (28.4%)
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	1.00 (0.00, 5.00)	1.00 (0.00, 5.00)
	ICU Admission,	Yes	506 (50.6%)	119 (47.6%)
	no. (%)	No	494 (49.4%)	131 (52.4%)
	ICU Duration, median (IQR)	ICU Duration, days	1.00 (0.00, 1.00)	0.00 (0.00, 1.00)
	Unplanned	Yes	22 (2.2%)	7 (2.8%)
	Admission, no. (%)	No	978 (97.8%)	243 (97.2%)
	Hospital	Yes	26 (2.6%)	6 (2.4%)
	Mortality, no. (%)	No	974 (97.4%)	244 (97.6%)

F.			Unplanned Adm	it Dataset
			Inference	Fewshot
	Demographics	Case Counts, no. (%)	949	237
Patient Info		Age, mean (std)	54.8 (18.6)	50.8 (19.2)
	G (04)	Female	471 (49.6%)	130 (54.9%)
	Sex, no. (%)	Male	478 (50.4%)	107 (45.1%)
		Other or Unknown	-	_
	Anesthesia Type,	General	812 (85.6%)	199 (84.0%)
	no. (%)	MAC	82 (8.6%)	17 (7.2%)
		Regional	55 (5.8%)	21 (8.9%)
		Deceased - Organ Donor	-	_
	Actual Patient	Emergency	_	1 (0.4%)
	Class, no. (%)	Inpatient	645 (68.0%)	155 (65.4%)
		Observation	39 (4.1%)	7 (3.0%)
		Outpatient	175 (18.4%)	47 (19.8%)
		Surgery Admit	4 (0.4%)	_
		Surgery Overnight Stay	86 (9.1%)	27 (11.4%)
Case	Expected Patient Class, no. (%)	Emergency	6 (0.6%)	2 (0.8%)
nfo		Inpatient	91 (9.6%)	24 (10.1%)
		Outpatient	163 (17.2%)	37 (15.6%)
		Surgery Admit	98 (10.3%)	23 (9.7%)
		Surgery Overnight Stay	541 (57.0%)	137 (57.8%)
		Unspecified	50 (5.3%)	14 (5.9%)
		Burn Surgery	15 (1.6%)	2 (0.8%)
		Cardiovascular Cath Lab	54 (5.7%)	15 (6.3%)
		Cardiovascular Surgery	5 (0.5%)	1 (0.4%)
		Gastroenterology	32 (3.4%)	10 (4.2%)
		General Surgery	143 (15.1%)	33 (13.9%)
	Proceduralist Service, no. (%)	Gynecologic Surgery	76 (8.0%)	21 (8.9%)
	Service, no. (%)	Hematology/ Oncology	1 (0.1%)	_
		Interventional Radiology	19 (2.0%)	4 (1.7%)
		Neurosurgery	89 (9.4%)	24 (10.1%)
		Obstetrics	27 (2.8%)	5 (2.1%)
		Ophthalmology	41 (4.3%)	4 (1.7%)
		Oral-Maxillofacial Surgery	28 (3.0%)	9 (3.8%)
		Organ Donor	_	_
		Orthopedic Surgery	193 (20.3%)	51 (21.5%)
		Otolaryngology	60 (6.3%)	27 (11.4%)

Table 1-F continued from previous page

F.			Unplanned Adm	it Dataset
		Plastic Surgery	30 (3.2%)	3 (1.3%)
		Podiatry	5 (0.5%)	3 (1.3%)
		Thoracic Surgery	6 (0.6%)	1 (0.4%)
		Transplant Surgery	18 (1.9%)	4 (1.7%)
		Urology	87 (9.2%)	18 (7.6%)
		Vascular Surgery	20 (2.1%)	2 (0.8%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (3, 7)	4 (3, 6)
		Most Recent Clinical Notes, total number of tokens	4263 (2271, 7373)	4263 (2380, 7394)
	ASA Physical Status, no. (%)	1	98 (10.3%)	31 (13.1%)
		2	413 (43.5%)	113 (47.7%)
		3	376 (39.6%)	84 (35.4%)
		4	60 (6.3%)	9 (3.8%)
		5	2 (0.2%)	_
Outcome		6	_	_
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	_	_
	Hospital	Yes	_	_
	Admission, no. (%)	No	_	_
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	-	-
	ICU Admission,	Yes	_	_
	no. (%)	No	_	_
	ICU Duration, median (IQR)	ICU Duration, days	_	_
	Unplanned	Yes	480 (50.6%)	113 (47.7%)
	Admission, no. (%)	No	469 (49.4%)	124 (52.3%)
	Hospital	Yes	6 (0.6%)	_
	Mortality, no. (%)	No	943 (99.4%)	237 (100.0%)

G.			Hospital Morta	lity Dataset
			Inference	Fewshot
	Demographics	Case Counts, no. (%)	576	3 144
Patient Info		Age, mean (std)	58.2 (18.9)	59.0 (19.7)
	~ (04)	Female	240 (41.7%)	67 (46.5%)
	Sex, no. (%)	Male	336 (58.3%)	77 (53.5%)
	Sex, no. (%)   Female	_	_	
	Anesthesia Type.	General	495 (85.9%)	124 (86.1%)
		MAC	63 (10.9%)	16 (11.1%)
		Regional	18 (3.1%)	4 (2.8%)
			_	_
	Actual Patient	Emergency	2 (0.3%)	_
		Inpatient	404 (70.1%)	92 (63.9%)
		Observation	16 (2.8%)	7 (4.9%)
		Outpatient	99 (17.2%)	29 (20.1%)
		Surgery Admit	1 (0.2%)	1 (0.7%)
			54 (9.4%)	15 (10.4%)
Case		Emergency	35 (6.1%)	8 (5.6%)
nfo	Expected Patient	Inpatient	243 (42.2%)	59 (41.0%)
Regional     Deceased -   Organ Donor     Emergency   Inpatient   Observation   Outpatient   Surgery Admit   Surgery Overnight   Stay   Inpatient   Outpatient   Outpatient   Surgery Overnight   Stay   Inpatient   Outpatient   Surgery Admit   Surgery Admit   Surgery Admit   Surgery Admit   Surgery Admit   Surgery Overnight   Stay   Unspecified   Burn Surgery   Cardiovascular   Cath Lab   Cardiovascular   Cath Lab   Cardiovascular   Surgery   Gastroenterology   General Surgery   Gastroenterology   General Surgery   Hematology   Oncology   Interventional   Radiology	Outpatient	53 (9.2%)	10 (6.9%)	
	Surgery Admit	77 (13.4%)	17 (11.8%)	
		85 (14.8%)	29 (20.1%)	
		Unspecified	83 (14.4%)	21 (14.6%)
		Burn Surgery	8 (1.4%)	1 (0.7%)
			53 (9.2%)	16 (11.1%)
			18 (3.1%)	5 (3.5%)
		Gastroenterology	35 (6.1%)	9 (6.2%)
		General Surgery	99 (17.2%)	29 (20.1%)
		Gynecologic Surgery	11 (1.9%)	6 (4.2%)
	Service, no. (%)	00,	1 (0.2%)	_
			48 (8.3%)	12 (8.3%)
		Neurosurgery	78 (13.5%)	13 (9.0%)
		Obstetrics	10 (1.7%)	2 (1.4%)
		Ophthalmology	22 (3.8%)	4 (2.8%)
			5 (0.9%)	_
		Organ Donor	_	_
		Orthopedic Surgery	98 (17.0%)	31 (21.5%)
		Otolaryngology	24 (4.2%)	7 (4.9%)

Table 1-G continued from previous page

G.			Hospital Mortali	ty Dataset
		Plastic Surgery	10 (1.7%)	_
		Podiatry	1 (0.2%)	_
		Thoracic Surgery	5 (0.9%)	1 (0.7%)
		Transplant Surgery	3 (0.5%)	_
		Urology	34 (5.9%)	6 (4.2%)
		Vascular Surgery	13 (2.3%)	2 (1.4%)
Note Statistics	Notes per Case, median (IQR)	Most Recent Clinical Notes, note count	4 (2, 7)	4 (2, 7)
		Most Recent Clinical Notes, total number of tokens	4482 (2137, 8302)	4455 (2684, 8286)
		1	45 (7.8%)	12 (8.3%)
	ASA Physical	2	140 (24.3%)	32 (22.2%)
	Status, no. (%)	3	165 (28.6%)	43 (29.9%)
		4	187 (32.5%)	49 (34.0%)
		5	39 (6.8%)	8 (5.6%)
Outcome		6	_	_
Variables	PACU Duration, median (IQR)	Phase 1 PACU Duration, minutes	_	_
	Hospital	Yes	_	_
	Admission, no. (%)	No	_	_
	Hospital Duration, median (IQR)	Hospital Admission Duration, days	_	-
	ICU Admission,	Yes	_	_
	no. (%)	No	_	_
	ICU Duration, median (IQR)	ICU Duration, days	-	_
	Unplanned	Yes	5 (0.9%)	2 (1.4%)
	Admission, no. (%)	No	571 (99.1%)	142 (98.6%)
	Hospital	Yes	292 (50.7%)	68 (47.2%)
	Mortality, no. $(\%)$	No	284 (49.3%)	76 (52.8%)

Appendix A Supplemental Tables

A.1 Supplemental Table 1: Note Type & Author Provider Type

	Dentist	Fellow	$\begin{array}{c} {\rm Nurse} \\ {\rm Anesthetist} \end{array}$	$\begin{array}{c} {\rm Nurse} \\ {\rm Practitioner} \end{array}$	Osteopath	Pharmacist	Physician	Physician Assistant	Podiatrist	Resident
Anesthesia Preprocedure Evaluation	ı	47	179	54	1	1	3346	83	1	418
Assessment & Plan Note	1	30	1	21	1	1	305	4	1	14
Brief Op Note	1	415	1	4	1	1	535	∞	1	1014
Brief Procedure Note	1	280	1	33	1	1	280	23	1	464
Consults	3	1056	1	360	ı	1	2601	999	23	5842
Discharge Summary	1	13	1	125	1	1	222	116	1	277
ED Procedure Notes	1	13	1	20	1	1	521	12	1	550
ED Provider Notes	1	20	1	402	1	1	3579	161	1	3401
H&P	06	2164	1	642	1	3	10920	1698	49	12583
H&P (View-Only)	4	391	1	561	1	1	4419	2545	187	2461
Hospital Course	1	∞	1	63	1	1	62	93	1	273
Interim Summary	1	1	1	11	1	1	28	9	1	51
Interval H&P Note	22	566	1	17	1	1	3658	822	36	4068
Op Note	23	135	1	1	1	1	2526	1	1	286
Procedures	ı	122	ı	38	1		252	14	1	323
Progress Notes	44	4687	2	12795	36	2267	87130	14103	1205	24578
Significant Event	1	13	1	10	1	1	48	11	1	94

Table A1: Count of final notes extracted from electronic health record to be considered in the construction of datasets. Only notes with the listed note type and author provider types were included. Notes with token length < 100 were also excluded as most of these notes were administrative or attending physician attestations which did not contain substantial clinical content.

## A.2 Supplemental Table 2: Experiment Costs

Task	Cost
ASA Physical Status Classification	\$765.50
Hospital Admission	\$743.02
ICU Admission	\$795.52
Unplanned Admission	\$764.09
Hospital Mortality	\$468.70
PACU Phase 1 Duration	\$831.08
Hospital Duration	\$834.44
ICU Duration	\$799.66
TOTAL	\$6,002.01

**Table A2:** Cost for API calls to GPT-4 Turbo (gpt-4-1106) using the Microsoft Azure OpenAI Service. This includes costs for both input prompt tokens as well as output LLM generated tokens.

## A.3 Supplemental Table 3: ASA Physical Status Classification - Metrics

AS	SA Physical Statu	s Classification: I	F1 Score	
	All	Long	Medium	Short
Baseline	0.17 (0.15, 0.19)	_	_	_
0-Shot — Notes	$0.41 \ (0.38, \ 0.44)$	$0.45 \ (0.40, \ 0.50)$	$0.40 \ (0.35, \ 0.45)$	0.38 (0.33, 0.43)
0-Shot — Summary	$0.40 \ (0.37, \ 0.43)$	$0.46 \ (0.41, \ 0.51)$	$0.38 \ (0.32, \ 0.43)$	0.36 (0.31, 0.42)
5-Shot — Summary	$0.41 \ (0.38, \ 0.44)$	$0.47 \ (0.41, \ 0.52)$	0.39 (0.34, 0.44)	0.36 (0.31, 0.42)
10-Shot — Summary	$0.41 \ (0.38, \ 0.44)$	0.44 (0.38, 0.49)	$0.41 \ (0.35, \ 0.46)$	0.38 (0.33, 0.43)
20-Shot — Summary	$0.43 \ (0.40, \ 0.46)$	$0.46 \ (0.41, \ 0.51)$	$0.43 \ (0.38, \ 0.49)$	0.40 (0.34, 0.45)
50-Shot — Summary	$0.42\ (0.39,\ 0.45)$	$0.48 \ (0.43, \ 0.53)$	$0.39\ (0.34,\ 0.44)$	0.39 (0.34, 0.45)
0-Shot CoT — Notes	$0.46 \ (0.43, \ 0.49)$	$0.46 \ (0.41, \ 0.52)$	$0.47 \ (0.41, \ 0.52)$	0.44 (0.39, 0.49)
0-Shot CoT — Summary	$0.45 \ (0.42, \ 0.48)$	$0.50 \ (0.44, \ 0.55)$	$0.44 \ (0.38, \ 0.49)$	$0.42 \ (0.36, \ 0.47)$
5-Shot CoT — Summary	$0.46 \ (0.43, \ 0.49)$	$0.49 \ (0.43, \ 0.54)$	$0.47 \ (0.42, \ 0.53)$	$0.43 \ (0.37, \ 0.48)$
10-Shot CoT — Summary	$0.48 \ (0.45, \ 0.52)$	$0.48 \; (0.42,  0.53)$	$0.51\ (0.45,\ 0.56)$	0.46 (0.41, 0.52)
20-Shot CoT — Summary	$0.50\ (0.47,\ 0.53)$	$0.50 \ (0.45, \ 0.56)$	$0.54 \ (0.49, \ 0.60)$	0.45 (0.40, 0.51)
50-Shot CoT — Summary	$0.50 \ (0.46, \ 0.53)$	$0.53\ (0.48,\ 0.59)$	$0.50 \ (0.44, \ 0.55)$	0.45 (0.40, 0.51)

Table A3: F1 score with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

ASA Physical Stat	us Classification: l	Matthew's Correl	ation Coefficient	(MCC)
	All	Long	Medium	Short
Baseline	0.00 (-0.03, 0.03)	_	_	_
0-Shot — Notes	$0.27 \ (0.23, \ 0.30)$	$0.29\ (0.23,\ 0.36)$	$0.25\ (0.19,\ 0.32)$	$0.24 \ (0.18, \ 0.31)$
0-Shot — Summary	$0.25 \ (0.21, \ 0.29)$	$0.31\ (0.25,\ 0.37)$	$0.22\ (0.16,\ 0.29)$	$0.22\ (0.16,\ 0.28)$
5-Shot — Summary	$0.26 \ (0.23,\ 0.30)$	$0.31\ (0.25,\ 0.38)$	$0.24\ (0.17,\ 0.31)$	$0.22\ (0.15,\ 0.29)$
10-Shot — Summary	$0.27 \ (0.23, \ 0.31)$	$0.28 \ (0.21, \ 0.34)$	$0.26\ (0.20,\ 0.33)$	0.24 (0.17, 0.30)
20-Shot — Summary	$0.30 \ (0.26, \ 0.33)$	$0.31\ (0.24,\ 0.38)$	$0.29 \ (0.22, \ 0.35)$	$0.26 \ (0.19, \ 0.33)$
50-Shot — Summary	$0.28 \ (0.24, \ 0.32)$	$0.33\ (0.26,\ 0.40)$	$0.24\ (0.17,\ 0.31)$	$0.25 \ (0.19, \ 0.32)$
0-Shot CoT — Notes	$0.32\ (0.28,\ 0.36)$	$0.30\ (0.23,\ 0.37)$	$0.33\ (0.26,\ 0.40)$	$0.31\ (0.25,\ 0.38)$
0-Shot CoT — Summary	$0.31\ (0.27,\ 0.35)$	$0.34\ (0.27,\ 0.41)$	$0.28 \ (0.22, \ 0.35)$	$0.28 \ (0.21, \ 0.35)$
5-Shot CoT — Summary	$0.32\ (0.28,\ 0.36)$	$0.34\ (0.27,\ 0.41)$	$0.32\ (0.25,\ 0.39)$	$0.27 \ (0.21, \ 0.34)$
10-Shot CoT — Summary	$0.35 \ (0.31, \ 0.39)$	$0.33\ (0.26,\ 0.40)$	$0.37 \ (0.30, \ 0.43)$	$0.32\ (0.25,\ 0.39)$
20-Shot CoT — Summary	$0.37 \ (0.33, \ 0.41)$	$0.36 \ (0.29, \ 0.43)$	$0.41\ (0.35,\ 0.48)$	$0.31\ (0.24,\ 0.37)$
50-Shot CoT — Summary	$0.36\ (0.32,\ 0.40)$	$0.40 \ (0.33, \ 0.47)$	$0.35\ (0.29,\ 0.42)$	$0.30\ (0.23,\ 0.37)$

Table A4: Matthew's Correlation Coefficient (MCC) with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

	Tri	ue Positive Rate (	True Positive Rate (TPR, Sensitivity, Recall)	Recall)		
	1	2	8	4	5	9
Baseline	0.18 (0.12, 0.24)	0.24 (0.18, 0.31)	0.22 (0.16, 0.28)	0.23 (0.17, 0.29)	0.12 (0.08, 0.17)	0.02 (0.00, 0.04)
0-Shot — Notes	0.85 (0.62, 1.00)	0.37 (0.32, 0.42)	0.36 (0.31, 0.40)	0.51 (0.44, 0.59)	0.65 (0.52, 0.77)	0.64 (0.00, 1.00)
0-Shot - Summary	0.89 (0.72, 1.00)	0.35 (0.31, 0.40)	0.35 (0.31, 0.40)	$0.52\ (0.45,\ 0.61)$	0.55 (0.41, 0.68)	0.62 (0.00, 1.00)
5-Shot — Summary	0.81 (0.61, 0.96)	0.35 (0.30, 0.40)	0.34 (0.30, 0.38)	0.57 (0.49, 0.65)	0.65 (0.52, 0.77)	0.74 (0.00, 1.00)
$10 ext{-Shot}$ — Summary	0.89 (0.72, 1.00)	0.35 (0.30, 0.40)	0.35 (0.30, 0.40)	0.53 (0.45, 0.61)	0.62 (0.50, 0.74)	$0.86\ (0.50,\ 1.00)$
$20 ext{-Shot}$ — Summary	0.78 (0.63, 0.92)	0.37 (0.33, 0.42)	0.35 (0.30, 0.40)	0.58 (0.50, 0.66)	0.63 (0.52, 0.74)	$0.90\ (0.67,\ 1.00)$
50-Shot - Summary	0.67 (0.53, 0.80)	0.37 (0.32, 0.42)	0.33 (0.28, 0.37)	0.55 (0.47, 0.64)	0.64 (0.52, 0.74)	$0.80\ (0.50,\ 1.00)$
0-Shot CoT — Notes	0.88 (0.78, 0.96)	0.40 (0.35, 0.45)	0.38 (0.33, 0.43)	0.52 (0.46, 0.59)	$1.00\ (1.00,\ 1.00)$	$0.91\ (0.67,\ 1.00)$
0-Shot CoT — Summary	0.78 (0.68, 0.88)	0.38 (0.33, 0.43)	0.39 (0.33, 0.44)	0.54 (0.47, 0.60)	0.54 (0.22, 0.86)	0.88 (0.60, 1.00)
5-Shot CoT — Summary	0.75 (0.67, 0.83)	0.40 (0.35, 0.45)	0.35 (0.30, 0.41)	0.53 (0.45, 0.60)	0.77 (0.50, 1.00)	$0.91\ (0.67,\ 1.00)$
10-Shot CoT — Summary	0.72 (0.64, 0.79)	0.42 (0.37, 0.48)	0.37 (0.32, 0.42)	0.54 (0.47, 0.62)	0.71 (0.50, 0.89)	$0.90\ (0.67,\ 1.00)$
20-Shot CoT — Summary	0.74 (0.66, 0.80)	0.45 (0.40, 0.50)	0.37 (0.32, 0.42)	0.54 (0.46, 0.61)	0.71 (0.54, 0.87)	$0.91\ (0.67,\ 1.00)$
50-Shot CoT — Summary	$0.68 \ (0.61, \ 0.75)$	$0.42\ (0.37,\ 0.47)$	$0.39\ (0.33,\ 0.44)$	0.56 (0.49, 0.64)	$0.74\ (0.59,\ 0.87)$	$0.91\ (0.67,\ 1.00)$

Table A5: True Positive Rate (Sensitivity) derived from confusion matrix.

		True Negative R	True Negative Rate (TNR, Specificity)	icity)		
	1	2	8	4	יט	9
Baseline	0.77 (0.74, 0.80)	0.77 (0.74, 0.80)	0.79 (0.76, 0.82)	0.78 (0.75, 0.81)	0.90 (0.88, 0.92)	0.99 (0.99, 1.00)
0-Shot — Notes	0.79 (0.77, 0.82)	0.85 (0.82, 0.87)	0.88 (0.86, 0.91)	0.83 (0.81, 0.86)	0.93 (0.91, 0.94)	0.99 (0.99, 1.00)
0-Shot — Summary	0.79 (0.77, 0.82)	0.84 (0.81, 0.87)	0.89 (0.86, 0.91)	0.83 (0.80, 0.85)	0.92 (0.90, 0.94)	0.99 (0.99, 1.00)
$5 ext{-Shot}$ — Summary	0.79 (0.77, 0.82)	0.83 (0.80, 0.86)	0.88 (0.85, 0.91)	0.84 (0.81, 0.86)	0.93 (0.91, 0.95)	0.99 (0.99, 1.00)
$10 ext{-Shot} -  ext{Summary}$	0.79 (0.77, 0.82)	0.83 (0.80, 0.86)	0.88 (0.85, 0.90)	0.83 (0.81, 0.86)	0.93 (0.92, 0.95)	1.00 (0.99, 1.00)
$20 ext{-Shot} -  ext{Summary}$	0.80 (0.78, 0.82)	0.85 (0.82, 0.88)	0.87 (0.85, 0.90)	0.84 (0.81, 0.86)	0.94 (0.92, 0.95)	1.00 (1.00, 1.00)
50-Shot — Summary	0.80 (0.78, 0.83)	0.84 (0.81, 0.87)	0.86 (0.83, 0.88)	0.83 (0.81, 0.86)	0.94 (0.93, 0.96)	1.00 (1.00, 1.00)
0-Shot CoT — Notes	$0.82\ (0.79,\ 0.84)$	0.87 (0.84, 0.89)	0.88 (0.86, 0.91)	0.85 (0.83, 0.88)	0.91 (0.89, 0.92)	1.00 (1.00, 1.00)
0-Shot CoT — Summary	0.82 (0.80, 0.84)	$0.85\ (0.82,\ 0.87)$	0.87 (0.84, 0.90)	0.87 (0.85, 0.90)	0.90 (0.88, 0.92)	1.00 (0.99, 1.00)
5-Shot CoT — Summary	0.85 (0.83, 0.88)	0.85 (0.83, 0.88)	0.85 (0.83, 0.88)	0.85 (0.82, 0.87)	$0.90\ (0.89,\ 0.92)$	1.00 (1.00, 1.00)
10-Shot CoT — Summary	0.87 (0.85, 0.89)	0.86 (0.84, 0.89)	0.86 (0.83, 0.89)	0.84 (0.82, 0.87)	0.91 (0.89, 0.93)	1.00(1.00, 1.00)
20-Shot CoT — Summary	0.88 (0.86, 0.90)	0.87 (0.85, 0.90)	0.86 (0.83, 0.88)	0.84 (0.81, 0.86)	0.91 (0.90, 0.93)	1.00 (1.00, 1.00)
$50 ext{-Shot CoT} -  ext{Summary}$	$0.87\ (0.85,\ 0.89)$	0.86 (0.83, 0.88)	0.86 (0.84, 0.89)	0.84 (0.82, 0.87)	$0.92\ (0.91,\ 0.94)$	$1.00\ (1.00,\ 1.00)$

	P	ositive Predictive	Positive Predictive Value (PPV, Precision)	cision)		
	1	2	8	4	ro	9
Baseline	0.04 (0.03, 0.06)	0.05 (0.04, 0.07)	0.04 (0.03, 0.06)	0.05 (0.04, 0.07)	0.02 (0.01, 0.03)	0.00 (0.00, 0.01)
0-Shot — Notes	0.01 (0.01, 0.02)	$0.18\ (0.15,\ 0.21)$	0.18 (0.15, 0.21)	0.11 (0.09, 0.14)	0.04 (0.03, 0.05)	0.00 (0.00, 0.01)
0-Shot — Summary	$0.02\ (0.01,\ 0.03)$	0.17 (0.14, 0.20)	0.19 (0.15, 0.22)	0.10 (0.08, 0.13)	0.03 (0.02, 0.04)	0.00 (0.00, 0.00)
5-Shot — Summary	0.02 (0.01, 0.03)	0.15 (0.12, 0.18)	0.18 (0.15, 0.22)	0.11 (0.09, 0.14)	0.04 (0.03, 0.06)	0.00 (0.00, 0.01)
$10 ext{-Shot} -  ext{Summary}$	0.02 (0.01, 0.03)	$0.16\ (0.13,\ 0.19)$	0.18 (0.15, 0.21)	0.11 (0.08, 0.13)	0.05 (0.03, 0.06)	0.01 (0.00, 0.01)
20-Shot — Summary	0.03 (0.02, 0.05)	$0.18\ (0.15,\ 0.21)$	0.17 (0.14, 0.20)	0.11 (0.09, 0.13)	0.05 (0.04, 0.07)	0.01 (0.00, 0.01)
50-Shot — Summary	$0.04\ (0.03,\ 0.05)$	0.17 (0.14, 0.20)	0.15 (0.13, 0.19)	0.11 (0.08, 0.13)	0.06 (0.04, 0.07)	0.01 (0.00, 0.01)
0-Shot CoT — Notes	$0.06\ (0.04,\ 0.07)$	$0.19\ (0.16,\ 0.23)$	0.17 (0.14, 0.20)	0.14 (0.11, 0.17)	0.01 (0.01, 0.02)	0.01 (0.00, 0.02)
0-Shot CoT — Summary	0.06 (0.05, 0.08)	$0.18\ (0.15,\ 0.21)$	0.16 (0.13, 0.19)	0.17 (0.14, 0.20)	0.01 (0.00, 0.01)	0.01 (0.00, 0.01)
5-Shot CoT — Summary	0.12 (0.09, 0.14)	$0.18\ (0.15,\ 0.21)$	0.14 (0.12, 0.17)	0.13 (0.11, 0.16)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)
10-Shot CoT — Summary	0.14 (0.11, 0.17)	$0.18\ (0.15,\ 0.22)$	0.15 (0.12, 0.18)	0.12 (0.10, 0.15)	0.02 (0.01, 0.03)	0.01 (0.00, 0.01)
20-Shot CoT — Summary	0.15 (0.12, 0.18)	$0.19\ (0.16,\ 0.22)$	0.15 (0.12, 0.17)	0.12 (0.09, 0.14)	0.02 (0.01, 0.04)	0.01 (0.00, 0.02)
$50 ext{-Shot CoT} -  ext{Summary}$	$0.14\ (0.12,\ 0.17)$	$0.18 \ (0.15, \ 0.21)$	0.15 (0.12, 0.18)	$0.12\ (0.10,\ 0.15)$	0.03 (0.02, 0.05)	$0.01 \ (0.00, \ 0.02)$

Table A7: Positive Predictive Value (precision) derived from confusion matrix.

		Negative Pred	Negative Predictive Value (NPV)	<u> </u>		
	1	2	3	4	יט	9
Baseline	0.81 (0.78, 0.84)	0.84 (0.81, 0.86)	0.85 (0.82, 0.87)	0.83 (0.80, 0.85)	0.83 (0.81, 0.86)	0.85 (0.82, 0.87)
$0 ext{-Shot}$ — Notes	1.00 (0.99, 1.00)	0.70 (0.67, 0.73)	0.68 (0.65, 0.71)	0.89 (0.87, 0.91)	0.98 (0.97, 0.99)	1.00 (1.00, 1.00)
0-Shot — Summary	1.00 (0.99, 1.00)	0.69 (0.66, 0.72)	0.66 (0.63, 0.69)	0.91 (0.89, 0.93)	0.98 (0.96, 0.98)	1.00 (1.00, 1.00)
5-Shot - Summary	0.99 (0.99, 1.00)	0.71 (0.68, 0.75)	0.64 (0.61, 0.68)	0.91 (0.89, 0.93)	0.98 (0.97, 0.99)	1.00 (1.00, 1.00)
$10 ext{-Shot} -  ext{Summary}$	1.00(0.99, 1.00)	0.70 (0.67, 0.73)	0.67 (0.64, 0.71)	0.91 (0.88, 0.93)	(997)	1.00 (1.00, 1.00)
$20 ext{-Shot} -  ext{Summary}$	0.99 (0.98, 1.00)	$0.71\ (0.67,\ 0.74)$	0.68 (0.65, 0.72)	0.92 (0.90, 0.94)	(997)	1.00 (1.00, 1.00)
$50 ext{-Shot}$ — Summary	$0.98\ (0.97,\ 0.99)$	0.72 (0.69, 0.75)	0.68 (0.65, 0.71)	0.91 (0.90, 0.93)	(997)	1.00 (0.99, 1.00)
0-Shot CoT — Notes	0.99 (0.99, 1.00)	0.71 (0.68, 0.74)	0.73 (0.69, 0.76)	0.87 (0.85, 0.89)	1.00 (1.00, 1.00)	1.00(1.00, 1.00)
0-Shot CoT — Summary	$0.98\ (0.97,\ 0.99)$	0.71 (0.68, 0.74)		0.86 (0.83, 0.88)	0.99 (0.99, 1.00)	1.00(1.00, 1.00)
5-Shot CoT — Summary	$0.96\ (0.95,\ 0.97)$	0.73 (0.70, 0.76)	0.74 (0.71, 0.77)	0.88 (0.86, 0.90)	1.00 (0.99, 1.00)	1.00(1.00, 1.00)
10-Shot CoT — Summary	0.95 (0.93, 0.96)	0.75 (0.72, 0.78)	0.75 (0.72, 0.78)	0.90 (0.88, 0.92)	0.99 (0.99, 1.00)	1.00 (1.00, 1.00)
20-Shot CoT — Summary	0.95 (0.93, 0.96)	0.77 (0.74, 0.80)	0.75 (0.72, 0.78)	0.90 (0.88, 0.92)	0.99 (0.98, 1.00)	$1.00\ (1.00,\ 1.00)$
50-Shot CoT — Summary	$0.93\ (0.92,\ 0.95)$	$0.76\ (0.73,\ 0.79)$	0.77 (0.74, 0.80)	0.91 (0.88, 0.93)	(0.99)	$1.00\ (1.00,\ 1.00)$

## A.4 Supplemental Table 4: Hospital Admission - Metrics

	Hospital Ad	mission: F1 Score	9	
	All	Long	Medium	Short
Baseline	0.52 (0.49, 0.55)	_	_	_
0-Shot — Notes	$0.58 \ (0.55, \ 0.61)$	$0.56 \ (0.51, \ 0.62)$	$0.57 \ (0.52, \ 0.62)$	$0.60\ (0.54,\ 0.65)$
0-Shot — Summary	$0.62\ (0.58,\ 0.65)$	$0.62\ (0.57,\ 0.67)$	$0.61\ (0.56,\ 0.66)$	$0.61\ (0.56,\ 0.66)$
5-Shot — Summary	$0.56 \ (0.53, \ 0.59)$	$0.59 \ (0.54, \ 0.65)$	$0.56 \ (0.51, \ 0.62)$	0.54 (0.48, 0.59)
10-Shot — Summary	$0.59 \ (0.56, \ 0.62)$	$0.59 \ (0.54, \ 0.64)$	$0.59 \ (0.54, \ 0.65)$	$0.58 \ (0.52, \ 0.63)$
20-Shot — Summary	$0.62\ (0.59,\ 0.65)$	$0.62 \ (0.56, \ 0.67)$	$0.63\ (0.58,\ 0.68)$	$0.61 \ (0.55, \ 0.66)$
50-Shot — Summary	$0.61\ (0.58,\ 0.64)$	$0.63\ (0.58,\ 0.68)$	$0.62\ (0.57,\ 0.67)$	$0.57 \ (0.52, \ 0.63)$
0-Shot CoT — Notes	$0.62\ (0.59,\ 0.65)$	$0.61\ (0.56,\ 0.66)$	$0.61\ (0.56,\ 0.66)$	$0.64 \ (0.59, \ 0.69)$
0-Shot CoT — Summary	$0.63\ (0.59,\ 0.66)$	$0.63\ (0.58,\ 0.68)$	$0.64 \ (0.59, \ 0.69)$	$0.61\ (0.55,\ 0.66)$
5-Shot CoT — Summary	$0.60\ (0.57,\ 0.63)$	$0.61\ (0.55,\ 0.65)$	$0.63\ (0.58,\ 0.68)$	$0.56 \ (0.50, \ 0.61)$
10-Shot CoT — Summary	$0.62\ (0.59,\ 0.65)$	$0.62\ (0.57,\ 0.67)$	$0.62\ (0.56,\ 0.67)$	$0.63 \ (0.57, \ 0.68)$
20-Shot CoT — Summary	$0.63 \ (0.60, \ 0.66)$	$0.64 \ (0.59, \ 0.69)$	$0.62\ (0.57,\ 0.67)$	$0.63 \ (0.57, \ 0.68)$
50-Shot CoT — Summary	$0.64 \ (0.61, \ 0.67)$	0.64 (0.59, 0.69)	0.65 (0.60, 0.70)	0.63 (0.57, 0.68)

**Table A9:** F1 score with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Hospital A	dmission: Matthey	w's Correlation C	oefficient (MCC)	
	All	Long	Medium	Short
Baseline	0.05 (-0.01, 0.11)	_	_	_
0-Shot — Notes	$0.28 \ (0.22, \ 0.33)$	$0.20\ (0.09,\ 0.29)$	$0.28 \ (0.18, \ 0.36)$	$0.31\ (0.22,\ 0.40)$
0-Shot — Summary	$0.27 \ (0.21, \ 0.33)$	$0.29 \ (0.19, \ 0.39)$	$0.26 \ (0.16, \ 0.36)$	0.22 (0.11, 0.32)
5-Shot — Summary	$0.25 \ (0.20, \ 0.31)$	$0.28 \ (0.19, \ 0.37)$	$0.24 \ (0.14, \ 0.34)$	0.22 (0.11, 0.31)
10-Shot — Summary	$0.28 \ (0.23, \ 0.34)$	$0.27 \ (0.17, \ 0.36)$	$0.28 \ (0.19, \ 0.37)$	0.27 (0.17, 0.36)
20-Shot — Summary	$0.33\ (0.27,\ 0.38)$	$0.30\ (0.21,\ 0.39)$	$0.33\ (0.23,\ 0.42)$	0.31 (0.21, 0.40)
50-Shot — Summary	$0.29 \ (0.23,\ 0.34)$	$0.30\ (0.20,\ 0.40)$	$0.32\ (0.23,\ 0.41)$	0.23 (0.12, 0.33)
0-Shot CoT — Notes	$0.30\ (0.24,\ 0.36)$	$0.26 \ (0.16, \ 0.36)$	$0.27 \ (0.17, \ 0.37)$	$0.36 \ (0.27, \ 0.44)$
0-Shot CoT — Summary	$0.33\ (0.27,\ 0.38)$	$0.32\ (0.22,\ 0.41)$	$0.34 \ (0.25, \ 0.44)$	0.29 (0.19, 0.38)
5-Shot CoT — Summary	$0.31\ (0.26,\ 0.36)$	$0.29 \ (0.19, \ 0.38)$	$0.37 \ (0.29, \ 0.45)$	0.24 (0.14, 0.34)
10-Shot CoT — Summary	$0.34\ (0.29,\ 0.39)$	$0.31\ (0.22,\ 0.40)$	$0.34\ (0.24,\ 0.42)$	0.34 (0.24, 0.43)
20-Shot CoT — Summary	$0.35\ (0.30,\ 0.40)$	$0.35 \ (0.26, \ 0.44)$	$0.34\ (0.25,\ 0.43)$	$0.33\ (0.23,\ 0.42)$
50-Shot CoT — Summary	0.35 (0.30, 0.40)	$0.34 \ (0.25, \ 0.43)$	$0.36\ (0.27,\ 0.45)$	0.32 (0.22, 0.41)

Table A10: Matthew's Correlation Coefficient (MCC) with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Hospital Admission					
	True Positive Rate (TPR, Sensitivity, Recall)	True Negative Rate (TNR, Specificity)	Positive Predictive Value (PPV, Precision)	Negative Predictive Value (NPV)	
Baseline	0.42 (0.38, 0.46)	0.63 (0.59, 0.67)	0.36 (0.30, 0.41)	0.51 (0.47, 0.55)	
0-Shot — Notes	$0.48 \ (0.44, \ 0.52)$	$0.82\ (0.77,\ 0.86)$	$0.57 \ (0.50, \ 0.65)$	$0.38 \ (0.35, \ 0.42)$	
0-Shot — Summary	$0.51\ (0.47,\ 0.55)$	$0.76 \ (0.72, \ 0.80)$	$0.49 \ (0.43, \ 0.56)$	$0.53 \ (0.49, \ 0.57)$	
5-Shot — Summary	$0.47 \ (0.44, \ 0.51)$	$0.81\ (0.76,\ 0.85)$	$0.57 \ (0.49, \ 0.65)$	$0.37 \ (0.33, \ 0.40)$	
10-Shot — Summary	$0.49 \ (0.45, \ 0.52)$	$0.82\ (0.77,\ 0.86)$	$0.56 \ (0.49, \ 0.64)$	$0.41\ (0.37,\ 0.45)$	
20-Shot — Summary	$0.51\ (0.47,\ 0.55)$	$0.83\ (0.79,\ 0.86)$	$0.56 \ (0.49, \ 0.63)$	0.46 (0.43, 0.50)	
50-Shot — Summary	$0.50 \ (0.46, \ 0.54)$	$0.79 \ (0.75, \ 0.83)$	$0.53 \ (0.46, \ 0.60)$	0.48 (0.44, 0.52)	
0-Shot CoT — Notes	$0.51\ (0.47,\ 0.55)$	$0.79 \ (0.75, \ 0.83)$	$0.53 \ (0.46, \ 0.60)$	0.50 (0.46, 0.54)	
0-Shot CoT — Summary	$0.52\ (0.48,\ 0.55)$	$0.81\ (0.77,\ 0.85)$	$0.54 \ (0.48, \ 0.62)$	0.49 (0.45, 0.53)	
5-Shot CoT — Summary	$0.50 \ (0.46, \ 0.53)$	0.84 (0.80, 0.88)	$0.58 \; (0.51,  0.65)$	0.41 (0.37, 0.45)	
10-Shot CoT — Summary	$0.51\ (0.48,\ 0.55)$	$0.85 \ (0.81, \ 0.88)$	$0.57 \ (0.50, \ 0.65)$	0.45 (0.41, 0.49)	
20-Shot CoT — Summary	$0.52\ (0.48,\ 0.56)$	0.84 (0.80, 0.88)	$0.57 \ (0.50, \ 0.64)$	0.48 (0.44, 0.52)	

Table A11: True positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) derived from confusion matrix.

# A.5 Supplemental Table 5: Intensive Care Unit (ICU) Admission - Metrics

ICU Admission: F1 Score				
	All	Long	Medium	Short
Baseline	0.52 (0.49, 0.56)	_	_	_
0-Shot — Notes	$0.80\ (0.77,\ 0.82)$	$0.74\ (0.70,\ 0.79)$	$0.82\ (0.78,\ 0.86)$	$0.83 \ (0.79, \ 0.87)$
0-Shot — Summary	$0.77 \ (0.74, \ 0.80)$	$0.74\ (0.69,\ 0.78)$	$0.78 \ (0.73, \ 0.82)$	$0.79 \ (0.75, \ 0.84)$
5-Shot — Summary	$0.78 \ (0.76, \ 0.81)$	$0.75 \ (0.70, \ 0.79)$	$0.80\ (0.76,\ 0.84)$	$0.80\ (0.76,\ 0.84)$
10-Shot — Summary	$0.79\ (0.77,\ 0.82)$	$0.77 \ (0.72, \ 0.81)$	$0.78 \ (0.74, \ 0.83)$	$0.82\ (0.78,\ 0.86)$
20-Shot — Summary	$0.79\ (0.76,\ 0.81)$	$0.76 \ (0.71, \ 0.80)$	$0.78 \ (0.74, \ 0.83)$	$0.81\ (0.77,\ 0.86)$
50-Shot — Summary	$0.80\ (0.77,\ 0.82)$	$0.77 \ (0.72, \ 0.81)$	$0.82\ (0.78,\ 0.86)$	$0.81\ (0.77,\ 0.85)$
0-Shot CoT — Notes	$0.78 \ (0.76, \ 0.81)$	$0.73 \ (0.68, \ 0.78)$	$0.80\ (0.76,\ 0.84)$	$0.82\ (0.78,\ 0.86)$
0-Shot CoT — Summary	$0.78 \ (0.76, \ 0.81)$	$0.74\ (0.69,\ 0.78)$	$0.80\ (0.75,\ 0.84)$	$0.81\ (0.77,\ 0.85)$
5-Shot CoT — Summary	$0.81\ (0.78,\ 0.83)$	0.76 (0.71, 0.80)	$0.82\ (0.78,\ 0.86)$	0.84 (0.81, 0.88)
10-Shot CoT — Summary	$0.81\ (0.78,\ 0.83)$	$0.77 \ (0.72, \ 0.81)$	$0.83\ (0.79,\ 0.86)$	$0.83 \ (0.79, \ 0.87)$
20-Shot CoT — Summary	$0.80\ (0.78,\ 0.83)$	$0.80\ (0.76,\ 0.84)$	$0.80\ (0.76,\ 0.84)$	0.80 (0.75, 0.84)
50-Shot CoT — Summary	$0.79 \ (0.77, \ 0.81)$	0.75 (0.70, 0.80)	$0.81\ (0.76,\ 0.85)$	0.81 (0.77, 0.86)

Table A12: F1 score with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

ICU Admission: Matthew's Correlation Coefficient (MCC)				
	All	Long	Medium	Short
Baseline	0.05 (-0.01, 0.11)	_	_	_
0-Shot — Notes	$0.60 \ (0.55, \ 0.65)$	$0.46 \ (0.36, \ 0.55)$	$0.64 \ (0.55, \ 0.72)$	$0.65 \ (0.56, \ 0.73)$
0-Shot — Summary	0.54 (0.49, 0.59)	$0.45 \ (0.35, \ 0.55)$	$0.56 \ (0.48, \ 0.65)$	$0.56 \ (0.47, \ 0.65)$
5-Shot — Summary	$0.57 \ (0.51, \ 0.62)$	$0.47 \ (0.36, \ 0.56)$	$0.61\ (0.52,\ 0.69)$	0.59 (0.49, 0.68)
10-Shot — Summary	$0.58 \ (0.53, \ 0.63)$	$0.52\ (0.42,\ 0.60)$	$0.57 \ (0.49, \ 0.66)$	$0.62 \ (0.54, \ 0.71)$
20-Shot — Summary	$0.57 \ (0.52, \ 0.62)$	$0.49 \ (0.39, \ 0.59)$	$0.57 \ (0.48, \ 0.66)$	$0.61\ (0.52,\ 0.70)$
50-Shot — Summary	$0.60 \ (0.55, \ 0.65)$	$0.51\ (0.42,\ 0.60)$	$0.65 \ (0.56, \ 0.72)$	$0.61\ (0.52,\ 0.70)$
0-Shot CoT — Notes	$0.57 \ (0.52, \ 0.62)$	$0.44 \ (0.34, \ 0.53)$	$0.61\ (0.52,\ 0.69)$	$0.62 \ (0.54, \ 0.71)$
0-Shot CoT — Summary	$0.56 \ (0.51, \ 0.61)$	$0.45 \ (0.35, \ 0.54)$	0.59 (0.51, 0.68)	0.61 (0.51, 0.69)
5-Shot CoT — Summary	$0.62 \ (0.57, \ 0.67)$	$0.50 \ (0.40, \ 0.59)$	$0.65 \ (0.56, \ 0.72)$	0.68 (0.59, 0.76)
10-Shot CoT — Summary	$0.61\ (0.57,\ 0.67)$	$0.51\ (0.41,\ 0.60)$	$0.66 \ (0.58, \ 0.73)$	0.64 (0.56, 0.72)
20-Shot CoT — Summary	$0.60 \ (0.56, \ 0.65)$	$0.58 \ (0.49, \ 0.66)$	$0.62\ (0.53,\ 0.69)$	0.58 (0.49, 0.67)
50-Shot CoT — Summary	$0.59\ (0.53,\ 0.63)$	0.49 (0.39, 0.58)	0.62 (0.54, 0.70)	0.61 (0.53, 0.70)

Table A13: Matthew's Correlation Coefficient (MCC) with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

ICU Admission				
	True Positive Rate (TPR, Sensitivity, Recall)	True Negative Rate (TNR, Specificity)	Positive Predictive Value (PPV, Precision)	Negative Predictive Value (NPV)
Baseline 0-Shot — Notes 0-Shot — Summary	0.42 (0.38, 0.46) 0.48 (0.44, 0.52) 0.51 (0.47, 0.55)	0.63 (0.59, 0.67) 0.82 (0.77, 0.86) 0.76 (0.72, 0.80)	0.36 (0.30, 0.41) 0.57 (0.50, 0.65) 0.49 (0.43, 0.56)	0.51 (0.47, 0.55) 0.38 (0.35, 0.42) 0.53 (0.49, 0.57)

5-Shot — Summary	$0.47 \; (0.44,  0.51)$	$0.81\ (0.76,\ 0.85)$	$0.57 \ (0.49, \ 0.65)$	$0.37 \ (0.33, \ 0.40)$
10-Shot — Summary	$0.49 \ (0.45, \ 0.52)$	$0.82\ (0.77,\ 0.86)$	$0.56 \ (0.49, \ 0.64)$	$0.41\ (0.37,\ 0.45)$
20-Shot — Summary	$0.51\ (0.47,\ 0.55)$	$0.83\ (0.79,\ 0.86)$	$0.56 \ (0.49, \ 0.63)$	$0.46 \ (0.43, \ 0.50)$
50-Shot — Summary	$0.50 \ (0.46, \ 0.54)$	$0.79 \ (0.75, \ 0.83)$	$0.53 \ (0.46, \ 0.60)$	$0.48 \; (0.44,  0.52)$
0-Shot CoT — Notes	$0.51\ (0.47,\ 0.55)$	$0.79 \ (0.75, \ 0.83)$	$0.53 \ (0.46, \ 0.60)$	$0.50 \ (0.46, \ 0.54)$
0-Shot CoT — Summary	$0.52\ (0.48,\ 0.55)$	$0.81\ (0.77,\ 0.85)$	$0.54\ (0.48,\ 0.62)$	$0.49 \ (0.45, \ 0.53)$
5-Shot CoT — Summary	$0.50 \ (0.46, \ 0.53)$	$0.84\ (0.80,\ 0.88)$	$0.58 \ (0.51, \ 0.65)$	$0.41\ (0.37,\ 0.45)$
10-Shot CoT — Summary	$0.51\ (0.48,\ 0.55)$	$0.85 \ (0.81, \ 0.88)$	$0.57 \ (0.50, \ 0.65)$	$0.45 \ (0.41, \ 0.49)$
20-Shot CoT — Summary	$0.52\ (0.48,\ 0.56)$	$0.84\ (0.80,\ 0.88)$	$0.57 \ (0.50, \ 0.64)$	$0.48 \; (0.44,  0.52)$
50-Shot CoT — Summary	$0.53 \ (0.49, \ 0.57)$	$0.83\ (0.79,\ 0.86)$	$0.55 \ (0.48, \ 0.62)$	$0.51\ (0.47,\ 0.55)$

Table A14: True positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) derived from confusion matrix

## A.6 Supplemental Table 6: Unplanned Admission - Metrics

Unplanned Admission: F1 Score				
	All	Long	Medium	Short
Baseline	0.45 (0.42, 0.48)	_	_	_
0-Shot — Notes	$0.60\ (0.57,\ 0.63)$	$0.59 \ (0.53, \ 0.64)$	$0.63 \ (0.58, \ 0.68)$	$0.58 \ (0.53, \ 0.64)$
0-Shot — Summary	$0.61\ (0.58,\ 0.64)$	$0.59 \ (0.53, \ 0.64)$	$0.64\ (0.59,\ 0.69)$	$0.60 \ (0.55, \ 0.66)$
5-Shot — Summary	$0.55 \ (0.52, \ 0.59)$	$0.56 \ (0.50, \ 0.61)$	$0.59 \ (0.53, \ 0.64)$	$0.52 \ (0.46, \ 0.57)$
10-Shot — Summary	$0.57 \ (0.54, \ 0.60)$	$0.57 \ (0.52, \ 0.63)$	$0.58 \ (0.53, \ 0.63)$	$0.57 \ (0.51, \ 0.63)$
20-Shot — Summary	$0.58 \ (0.54, \ 0.61)$	$0.58 \ (0.52, \ 0.63)$	$0.58 \ (0.53, \ 0.63)$	$0.57 \ (0.52, \ 0.63)$
50-Shot — Summary	$0.58 \ (0.55, \ 0.61)$	$0.57 \ (0.52, \ 0.63)$	0.59 (0.53, 0.64)	0.59 (0.54, 0.64)
0-Shot CoT — Notes	$0.58 \ (0.55, \ 0.61)$	$0.56 \ (0.50, \ 0.61)$	$0.60 \ (0.55, \ 0.66)$	0.57 (0.52, 0.63)
0-Shot CoT — Summary	$0.58 \ (0.55, \ 0.61)$	$0.59 \ (0.53, \ 0.64)$	$0.61\ (0.56,\ 0.66)$	0.54 (0.49, 0.59)
5-Shot CoT — Summary	$0.59 \ (0.56, \ 0.62)$	$0.56 \ (0.51, \ 0.62)$	$0.61\ (0.55,\ 0.66)$	0.61 (0.56, 0.66)
10-Shot CoT — Summary	$0.58 \; (0.55,  0.61)$	$0.57 \ (0.51, \ 0.62)$	$0.60 \ (0.55, \ 0.65)$	0.57 (0.52, 0.63)
20-Shot CoT — Summary	$0.58 \ (0.55, \ 0.61)$	$0.56 \ (0.51, \ 0.61)$	$0.61\ (0.56,\ 0.67)$	$0.58 \ (0.53, \ 0.63)$
50-Shot CoT — Summary	$0.57 \ (0.53, \ 0.60)$	$0.55 \ (0.50, \ 0.61)$	$0.57 \ (0.52, \ 0.63)$	0.57 (0.51, 0.62)

Table A15: F1 score with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Unplanned Admission: Matthew's Correlation Coefficient (MCC)				
	All	Long	Medium	Short
Baseline	-0.10 (-0.16, -0.03)	_	_	_
0-Shot — Notes	$0.22 \ (0.15, \ 0.28)$	$0.20\ (0.09,\ 0.30)$	$0.28 \ (0.18, \ 0.38)$	$0.17 \ (0.07, \ 0.28)$
0-Shot — Summary	$0.22 \ (0.16, \ 0.28)$	$0.19 \ (0.08, \ 0.30)$	$0.27 \ (0.17, \ 0.37)$	$0.20\ (0.10,\ 0.31)$
5-Shot — Summary	$0.11\ (0.05,\ 0.17)$	$0.12\ (0.01,\ 0.23)$	$0.17 \ (0.06, \ 0.28)$	0.04 (-0.07, 0.15)
10-Shot — Summary	$0.15 \ (0.09, \ 0.21)$	$0.16 \ (0.05, \ 0.26)$	$0.15 \ (0.05, \ 0.26)$	$0.14 \ (0.03, \ 0.25)$
20-Shot — Summary	$0.16 \ (0.09, \ 0.22)$	$0.17 \ (0.06, \ 0.27)$	$0.16 \ (0.05, \ 0.26)$	$0.15 \ (0.04, \ 0.26)$
50-Shot — Summary	$0.17 \ (0.11, \ 0.23)$	$0.15 \ (0.04, \ 0.26)$	$0.17 \ (0.07, \ 0.28)$	$0.19 \ (0.09, \ 0.29)$
0-Shot CoT — Notes	$0.16 \ (0.09, \ 0.21)$	$0.11\ (0.01,\ 0.22)$	$0.21\ (0.10,\ 0.31)$	$0.15 \ (0.04, \ 0.25)$
0-Shot CoT — Summary	$0.16 \ (0.10, \ 0.22)$	$0.18 \ (0.07, \ 0.29)$	$0.23 \ (0.12, \ 0.33)$	0.08 (-0.03, 0.19)
5-Shot CoT — Summary	$0.19 \ (0.12, \ 0.25)$	$0.13 \ (0.02, \ 0.24)$	$0.21\ (0.10,\ 0.32)$	$0.22\ (0.11,\ 0.32)$
10-Shot CoT — Summary	$0.16 \ (0.10, \ 0.22)$	$0.13 \ (0.03, \ 0.24)$	$0.20\ (0.09,\ 0.30)$	$0.15 \ (0.04, \ 0.25)$
20-Shot CoT — Summary	$0.17 \ (0.11, \ 0.23)$	$0.12\ (0.01,\ 0.23)$	$0.23 \ (0.12, \ 0.34)$	$0.16 \ (0.05, \ 0.27)$
50-Shot CoT — Summary	$0.13\ (0.07,\ 0.20)$	0.11 (-0.00, 0.23)	$0.15 \ (0.04, \ 0.26)$	$0.13\ (0.03,\ 0.24)$

Table A16: Matthew's Correlation Coefficient (MCC) with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Unplanned Admission				
	True Positive Rate (TPR, Sensitivity, Recall)	True Negative Rate (TNR, Specificity)	Positive Predictive Value (PPV, Precision)	Negative Predictive Value (NPV)
Baseline	$0.46 \ (0.41, \ 0.50)$	$0.45 \ (0.40, \ 0.49)$	$0.47 \ (0.39, \ 0.55)$	0.45 (0.40, 0.49)
$0 ext{-Shot}$ — Notes	$0.57 \ (0.53, \ 0.61)$	$0.66 \ (0.61,\ 0.72)$	$0.82\ (0.72,\ 0.94)$	$0.39 \ (0.35, \ 0.43)$
0-Shot — Summary	$0.60 \ (0.56, \ 0.64)$	$0.63 \ (0.58, \ 0.67)$	$0.71\ (0.61,\ 0.81)$	$0.52 \ (0.48, \ 0.57)$
5-Shot — Summary	$0.55 \ (0.51, \ 0.59)$	$0.56 \ (0.51,\ 0.61)$	$0.67 \ (0.57, \ 0.77)$	$0.45 \ (0.41, \ 0.50)$
10-Shot — Summary	$0.56 \ (0.52, \ 0.60)$	$0.60 \ (0.54, \ 0.64)$	$0.72\ (0.62,\ 0.82)$	$0.44 \ (0.40, \ 0.48)$
20-Shot — Summary	$0.56 \ (0.52, \ 0.60)$	$0.60\ (0.55,\ 0.65)$	$0.73\ (0.62,\ 0.84)$	$0.44 \ (0.40, \ 0.48)$
50-Shot — Summary	$0.57 \ (0.53, \ 0.61)$	$0.61\ (0.56,\ 0.66)$	$0.74\ (0.64,\ 0.85)$	$0.44 \ (0.39, \ 0.48)$
0-Shot CoT — Notes	$0.59 \ (0.54, \ 0.64)$	$0.57 \ (0.52, \ 0.61)$	$0.54\ (0.45,\ 0.63)$	$0.63 \ (0.59, \ 0.67)$
0-Shot CoT — Summary	$0.58 \ (0.54, \ 0.63)$	$0.58 \ (0.53, \ 0.62)$	$0.60\ (0.52,\ 0.70)$	$0.57 \ (0.53, \ 0.61)$
5-Shot CoT — Summary	$0.59 \ (0.54, \ 0.63)$	$0.60\ (0.55,\ 0.65)$	$0.66 \ (0.57, \ 0.76)$	$0.54 \ (0.49, \ 0.58)$
10-Shot CoT — Summary	$0.58 \ (0.54, \ 0.63)$	$0.58 \ (0.53, \ 0.62)$	$0.60\ (0.51,\ 0.69)$	$0.58 \ (0.53, \ 0.62)$
20-Shot CoT — Summary	$0.59 \ (0.55, \ 0.63)$	$0.58 \ (0.54, \ 0.62)$	$0.59\ (0.50,\ 0.68)$	$0.59 \ (0.55, \ 0.64)$
50-Shot CoT — Summary	$0.58 \ (0.53, \ 0.62)$	$0.56 \ (0.51,\ 0.60)$	$0.53\ (0.45,\ 0.61)$	$0.62\ (0.57,\ 0.66)$

Table A17: True positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) derived from confusion matrix.

## A.7 Supplemental Table 7: Hospital Mortality - Metrics

Hospital Mortality: F1 Score				
	All	Long	Medium	Short
Baseline	0.49 (0.45, 0.53)	_	_	_
0-Shot — Notes	$0.74 \ (0.70, \ 0.78)$	$0.73 \ (0.67, \ 0.80)$	$0.79 \ (0.73, \ 0.84)$	0.70 (0.64, 0.77)
0-Shot — Summary	0.75 (0.72, 0.79)	$0.78 \ (0.72, \ 0.84)$	$0.78 \ (0.72, \ 0.83)$	$0.70 \ (0.63, \ 0.76)$
5-Shot — Summary	0.84 (0.81, 0.87)	$0.85 \ (0.80, \ 0.90)$	$0.89 \ (0.84, \ 0.93)$	0.78 (0.72, 0.84)
10-Shot — Summary	$0.86 \ (0.83, \ 0.89)$	$0.89 \ (0.84, \ 0.93)$	$0.89\ (0.85,\ 0.93)$	0.79 (0.73, 0.85)
20-Shot — Summary	$0.86 \ (0.83, \ 0.89)$	$0.89 \ (0.84, \ 0.93)$	$0.89 \ (0.84, \ 0.93)$	0.81 (0.75, 0.86)
50-Shot — Summary	0.84 (0.81, 0.87)	$0.84\ (0.79,\ 0.90)$	$0.88 \ (0.83, \ 0.92)$	0.81 (0.75, 0.86)
0-Shot CoT — Notes	$0.67 \ (0.63, \ 0.70)$	$0.62\ (0.55,\ 0.69)$	$0.70 \ (0.64, \ 0.76)$	0.68 (0.61, 0.74)
0-Shot CoT — Summary	$0.70 \ (0.66, \ 0.74)$	$0.69 \ (0.62, \ 0.76)$	0.76 (0.69, 0.81)	0.66 (0.59, 0.72)
5-Shot CoT — Summary	$0.83 \ (0.80, \ 0.86)$	$0.85 \ (0.80, \ 0.90)$	$0.88 \ (0.83, \ 0.92)$	0.77 (0.70, 0.82)
10-Shot CoT — Summary	$0.83 \ (0.80, \ 0.86)$	$0.88 \ (0.83, \ 0.93)$	$0.88 \ (0.83, \ 0.92)$	0.75 (0.69, 0.81)
20-Shot CoT — Summary	$0.83 \ (0.80, \ 0.86)$	$0.88 \ (0.83, \ 0.93)$	$0.89 \ (0.84, \ 0.93)$	0.73 (0.67, 0.79)
50-Shot CoT — Summary	0.83(0.80, 0.86)	0.87 (0.82, 0.92)	0.87 (0.82, 0.92)	0.76 (0.69, 0.82)

**Table A18**: F1 score with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Hospital Mortality: Matthew's Correlation Coefficient (MCC)				
	All	Long	Medium	Short
Baseline	-0.02 (-0.11, 0.06)	_	_	_
0-Shot — Notes	$0.56 \ (0.51, \ 0.61)$	$0.58 \ (0.50, \ 0.66)$	$0.61\ (0.52,\ 0.70)$	$0.47 \ (0.38, \ 0.56)$
0-Shot — Summary	$0.58 \ (0.53, \ 0.63)$	$0.64 \ (0.56, \ 0.72)$	$0.59 \ (0.50,  0.68)$	0.46 (0.37, 0.55)
5-Shot — Summary	$0.70 \ (0.65, \ 0.75)$	$0.72 \ (0.63, \ 0.81)$	0.79(0.71, 0.87)	0.59 (0.48, 0.68)
10-Shot — Summary	$0.73 \ (0.68, \ 0.78)$	$0.79 \ (0.71, \ 0.87)$	$0.78 \ (0.69, \ 0.86)$	0.60 (0.50, 0.70)
20-Shot — Summary	$0.73 \ (0.67, \ 0.78)$	$0.78 \ (0.68, \ 0.86)$	$0.77 \ (0.68, \ 0.86)$	$0.62 \ (0.52, \ 0.73)$
50-Shot — Summary	$0.70 \ (0.65, \ 0.75)$	$0.71\ (0.62,\ 0.80)$	$0.76 \ (0.67, \ 0.84)$	0.62 (0.51, 0.72)
0-Shot CoT — Notes	0.44 (0.39, 0.49)	0.44 (0.37, 0.51)	$0.44 \ (0.35, \ 0.54)$	0.43 (0.33, 0.52)
0-Shot CoT — Summary	$0.49 \ (0.43, \ 0.54)$	$0.50 \ (0.41, \ 0.59)$	$0.54 \ (0.44, \ 0.64)$	0.38 (0.28, 0.48)
5-Shot CoT — Summary	$0.69 \ (0.63, \ 0.74)$	0.72 (0.63, 0.81)	0.76 (0.68, 0.84)	0.55 (0.44, 0.66)
10-Shot CoT — Summary	0.70 (0.65, 0.75)	0.77(0.69, 0.85)	0.76 (0.68, 0.84)	0.53 (0.42, 0.64)
20-Shot CoT — Summary	$0.69 \ (0.64, \ 0.74)$	$0.77 \ (0.69, \ 0.85)$	$0.78 \ (0.70, \ 0.86)$	0.48 (0.37, 0.59)
50-Shot CoT — Summary	0.69 (0.64, 0.74)	0.76 (0.66, 0.84)	0.75 (0.66, 0.83)	0.54 (0.44, 0.64)

#### Table A19 continued from previous page $\,$

#### Hospital Mortality: Matthew's Correlation Coefficient (MCC)

Table A19: Matthew's Correlation Coefficient (MCC) with 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the score from a dummy classification model that randomly guesses the outcome variable.

Hospital Mortality				
	True Positive Rate (TPR, Sensitivity, Recall)	True Negative Rate (TNR, Specificity)	Positive Predictive Value (PPV, Precision)	Negative Predictive Value (NPV)
Baseline	0.50 (0.44, 0.55)	0.48 (0.42, 0.54)	0.52 (0.42, 0.63)	0.47 (0.41, 0.53)
0-Shot — Notes	0.99(0.97, 1.00)	$0.66 \ (0.61, \ 0.70)$	$0.51\ (0.42,\ 0.62)$	0.99 (0.98, 1.00)
0-Shot — Summary	0.99 (0.97, 1.00)	$0.67 \ (0.62, \ 0.71)$	$0.53 \ (0.44, \ 0.64)$	0.99 (0.98, 1.00)
5-Shot — Summary	$0.95 \ (0.92, \ 0.98)$	0.77 (0.73, 0.81)	$0.74\ (0.62,\ 0.88)$	0.96 (0.94, 0.98)
10-Shot — Summary	$0.95 \ (0.92, \ 0.97)$	$0.79 \ (0.75, \ 0.84)$	$0.79 \ (0.66, \ 0.93)$	0.96 (0.93, 0.98)
20-Shot — Summary	$0.93\ (0.90,\ 0.96)$	$0.81\ (0.76,\ 0.85)$	$0.81\ (0.68,\ 0.96)$	0.94 (0.91, 0.97)
50-Shot — Summary	$0.93\ (0.90,\ 0.96)$	$0.78 \ (0.73, \ 0.82)$	$0.76 \ (0.64, \ 0.91)$	0.95 (0.92, 0.97)
0-Shot CoT — Notes	$0.98 \ (0.95, 1.00)$	$0.60 \ (0.55, \ 0.64)$	$0.36\ (0.29,\ 0.44)$	0.99 (0.98, 1.00)
0-Shot CoT — Summary	0.95 (0.92, 0.99)	$0.63 \ (0.58, \ 0.67)$	$0.45 \ (0.36, \ 0.54)$	0.98 (0.96, 0.99)
5-Shot CoT — Summary	$0.95 \ (0.92, \ 0.98)$	$0.76 \ (0.72, \ 0.80)$	0.73 (0.61, 0.86)	0.96 (0.94, 0.98)
10-Shot CoT — Summary	0.97 (0.94, 0.99)	0.76 (0.72, 0.80)	0.72 (0.60, 0.85)	0.98 (0.96, 0.99)
20-Shot CoT — Summary	0.95 (0.92, 0.98)	0.76 (0.71, 0.80)	0.73 (0.61, 0.86)	0.96 (0.94, 0.98)
50-Shot CoT — Summary	0.95(0.92, 0.98)	0.76(0.72, 0.80)	0.73 (0.61, 0.87)	0.96 (0.94, 0.98)

Table A20: True positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) derived from confusion matrix.

## A.8 Supplemental Table 8: Post-Anesthesia Care Unit (PACU) - Phase 1 Duration - Metrics

Phase 1 PACU Duration: Mean Absolute Error (MAE) in minutes				
	All	Long	Medium	Short
Baseline	36.12 (34.02, 38.28)	_	_	_
0-Shot — Notes	49.01 (46.89, 51.15)	53.30 (49.11, 57.59)	46.37 (42.93, 50.03)	47.24 (44.04, 50.67)
0-Shot — Summary	47.03 (45.11, 48.98)	49.99 (46.56, 53.68)	44.64 (41.46, 47.86)	46.35 (43.23, 49.87)
5-Shot — Summary	69.06 (65.87, 72.25)	76.98 (71.67, 82.52)	68.89 (63.42, 74.61)	61.24 (56.83, 65.85)
10-Shot — Summary	69.14 (66.09, 72.27)	74.55 (69.20, 80.12)	68.24 (62.82, 73.70)	64.43 (59.63, 69.49)
20-Shot — Summary	69.74 (66.78, 72.86)	75.90 (70.69, 81.18)	69.23 (63.85, 74.85)	63.96 (59.25, 68.74)
50-Shot — Summary	60.09 (57.19, 63.06)	63.94 (58.47, 69.51)	59.72 (54.91, 64.80)	56.42 (51.73, 61.18)
0-Shot CoT — Notes	58.29 (55.80, 61.01)	65.49 (60.90, 70.23)	54.62 (50.36, 58.89)	54.68 (50.31, 59.21)
0-Shot CoT — Summary	58.66 (56.14, 61.17)	67.88 (63.33, 72.48)	53.88 (49.82, 58.14)	54.02 (49.84, 58.45)
5-Shot CoT — Summary	58.35 (55.73, 61.24)	67.74 (62.80, 72.62)	56.97 (52.01, 62.22)	50.23 (46.19, 54.38)
10-Shot CoT — Summary	60.21 (55.16, 67.92)	65.37 (60.37, 70.75)	55.43 (50.75, 60.35)	60.01 (47.80, 80.37)
20-Shot CoT — Summary	55.24 (52.67, 57.88)	61.99 (56.98, 67.32)	52.44 (48.23, 56.77)	51.21 (46.97, 55.70)
50-Shot CoT — Summary	49.38 (46.88, 51.90)	56.13 (51.58, 60.86)	$47.01 \ (43.21, \ 51.03)$	44.91 (40.71, 49.11)

Table A21: Mean absolute error (MAE) minutes and 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the MAE from a dummy regression model that always predicts the mean duration in the dataset.

## A.9 Supplemental Table 9: Hospital Duration - Metrics

Hospital Duration:	Mean Absolute Er	ror (MAE) in days	
All	Long	Medium	Short

Table A22 continued from previous page

Hospita	al Duration: Mean	Absolute Error (N	MAE) in days	
Baseline	5.42 (5.04, 5.85)	_	_	_
0-Shot — Notes	4.80 (4.40, 5.26)	5.47 (4.66, 6.45)	4.43 (3.84, 5.11)	4.49 (3.87, 5.23)
0-Shot — Summary	4.72(4.31, 5.17)	5.17(4.37, 6.14)	4.55 (3.98, 5.20)	4.42 (3.80, 5.18)
5-Shot — Summary	6.87 (6.39, 7.39)	7.98 (6.99, 9.08)	7.03 (6.24, 7.87)	5.60 (4.89, 6.39)
10-Shot — Summary	$7.81\ (7.30,\ 8.35)$	9.06 (8.04, 10.11)	$7.87 \ (7.02, 8.75)$	6.47 (5.72, 7.35)
20-Shot — Summary	7.70 (7.16, 8.24)	9.22 (8.17, 10.31)	7.89 (7.00, 8.82)	5.95 (5.19, 6.84)
50-Shot — Summary	$7.87 \ (7.35, \ 8.42)$	8.74 (7.83, 9.81)	8.09 (7.24, 8.98)	6.72 (5.91, 7.65)
0-Shot CoT — Notes	$4.70 \ (4.31, \ 5.17)$	5.67 (4.83, 6.70)	4.33 (3.80, 4.92)	4.09 (3.53, 4.81)
0-Shot CoT — Summary	4.55 $(4.18, 4.98)$	$5.01 \ (4.28, \ 5.91)$	4.30 (3.75, 4.89)	4.32 (3.74, 5.03)
5-Shot CoT — Summary	8.10 (7.57, 8.67)	9.64 (8.58, 10.78)	8.29 (7.40, 9.18)	6.35 (5.58, 7.22)
10-Shot CoT — Summary	8.11 (7.56, 8.69)	9.51 (8.44, 10.58)	$8.50\ (7.58,\ 9.46)$	$6.30 \ (5.44, \ 7.27)$
20-Shot CoT — Summary	8.71 (8.15, 9.30)	11.09 (9.97, 12.27)	8.19 (7.30, 9.14)	6.81 (5.93, 7.77)
50-Shot CoT — Summary	7.51 (7.01, 8.02)	8.96 (7.98, 9.95)	$7.52 \ (6.70,\ 8.34)$	$6.01\ (5.28,\ 6.85)$

Table A22: Mean absolute error (MAE) days and 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the MAE from a dummy regression model that always predicts the mean duration in the dataset.

## A.10 Supplemental Table 10: Intensive Care Unit (ICU) Duration - Metrics

ICU Duration: Mean Absolute Error (MAE) in days				
	All	Long	Medium	Short
Baseline	1.15 (0.99, 1.33)	_	_	_
0-Shot — Notes	1.68 (1.51, 1.86)	$2.12\ (1.80,\ 2.46)$	$1.82\ (1.47,\ 2.27)$	1.12 (0.96, 1.30)
0-Shot — Summary	1.67 (1.51, 1.84)	2.07 (1.78, 2.38)	1.77 (1.42, 2.20)	1.17 (1.01, 1.34)
5-Shot — Summary	2.28 (2.12, 2.45)	$2.60\ (2.31,\ 2.90)$	$2.39\ (2.05,\ 2.80)$	1.87 (1.67, 2.10)
10-Shot — Summary	$2.19\ (2.00,\ 2.38)$	2.59 (2.26, 2.94)	2.24 (1.89, 2.67)	1.75 (1.53, 1.99)
20-Shot — Summary	1.80 (1.62, 1.99)	$2.20\ (1.90,\ 2.53)$	1.91 (1.55, 2.33)	1.31 (1.10, 1.54)
50-Shot — Summary	1.11 (0.94, 1.29)	1.51 (1.20, 1.86)	1.19 (0.85, 1.61)	0.62 (0.51, 0.74)
0-Shot CoT — Notes	1.17 (1.02, 1.34)	1.57 (1.30, 1.87)	1.27 (0.95, 1.68)	0.69 (0.58, 0.81)
0-Shot CoT — Summary	1.22 (1.07, 1.39)	1.68 (1.41, 1.98)	1.26 (0.94, 1.66)	0.73 (0.61, 0.85)
5-Shot CoT — Summary	2.18 (1.98, 2.38)	2.90 (2.56, 3.26)	2.18 (1.80, 2.61)	1.48 (1.25, 1.71)
10-Shot CoT — Summary	2.07 (1.88, 2.27)	2.84 (2.49, 3.20)	1.99 (1.62, 2.41)	1.40 (1.17, 1.64)
20-Shot CoT — Summary	$1.85\ (1.67,\ 2.05)$	$2.50\ (2.18,\ 2.85)$	1.90 (1.53, 2.33)	1.17 (0.96, 1.38)
50-Shot CoT — Summary	$1.12 \ (0.97, \ 1.29)$	1.53 (1.28, 1.84)	$1.21\ (0.88,\ 1.62)$	$0.64 \ (0.50, \ 0.79)$

Table A23: Mean absolute error (MAE) days and 95% confidence interval for all cases as well as short, medium, long note length strata. Baseline is the MAE from a dummy regression model that always predicts the mean duration in the dataset.

## Appendix B Supplemental Figures

### B.1 Supplemental Figure 1: Representative Text Prompts

B.1.1 to B.1.9 are representative examples of all prompts used in experiments. The prompts are given followed by the LLM output after the dashed line. Prompts are presented with color-coded highlights to assist identification of text source and salient portions of the prompt. We show all prompts and corresponding LLM outputs using the same patient case and procedure for B.1.1 to B.1.8 to enable comparison between the different prompt strategies. Few-shot prompts in B.1.4 and B.1.7 are illustrated with 4 in-context examples, but in our experiments we use 5, 10, 20, and 50 in-context examples. The prompt in B.1.9 uses a different patient case and procedure because it is only used for generating Chain-of-Thought rationale for the in-context examples depicted in prompt for B.1.7 "Few-shot CoT Q&A From Notes Summary".

While the content of these examples derived from a real patient and case from the electronic health record, all PHI and PII are removed with names obfuscated, and dates and times shifted.

Highlight Color	Text Source/Type
Green	Task Specification
Yellow	Procedure Information from Case Booking
$\operatorname{Red}$	Note Metadata
${f Purple}$	Note Text for Query Case
Pink	Summarized Note Text for Query Case
Blue	In-Context Examples

## B.1.1 System Message

#### System Message

Orange

Brown

You are a physician working in a hospital surgery center who is assessing patients to determine their outcome after a procedure.

In-Context Examples Chain-of-Thought Rationale

Modifications to Prompt to Induce Chain-of-Thought Reasoning

## B.1.2 Zero-shot Q&A From Original Notes

#### Prompt & LLM Output

You are given a task and context. The context contains information from the proposed procedure and patient's medical record. Assess the patient in the context of the proposed procedure and then provide an answer.

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Notes:

Progress Notes written by Physician at 2020-07-02 12:23:00:

ELECTROPHYSIOLOGY CLINIC INITIAL CONSULT REASON FOR CONSULT: Atrial fibrillation REFERRING: Self PRIMARY CARE: No primary care provider on file. Distant Site Telemedicine Encounter I conducted this encounter from Marvin Monroe

Memorial Hospital via secure, live, face-to-face video conference with the patient. Homer was located at home with his wife. I reviewed the risks and benefits of telemedicine as pertinent to this visit and the patient agreed to proceed. PROBLEM LIST 1. Atrial fibrillation HPI: Mr. Simpson is a 62 year old male referred for evaluation of atrial fibrillation. He has a long standing history of episodic palpitations for several years. After an episode of the same earlier in 2020 he was seen by his PCP who noted that 12 lead ECG showed atrial fibrillation. These episodes were paroxysmal in nature without clear provoking triggers or palliating factors. He had a more sustained episode of palpitations and light headedness on a flight to Springfield where he ultimately was admitted to a hospital for evaluation. ECG and telemetry showed persistent atrial fibrillation and he underwent a TEE cardioversion. He was started on Xarelto and advised to continue for 4 weeks. He had a TTE showing normal LV function and no valvular heart disease. Imaging also showed possible cholecystitis though HIDA scan was most in keeping with biliary dyskinesia. Today he joins me via telemedicine. He has felt generally well since his cardioversion and discharge. He has not had a recurrence of rapid heart action. He has continued Xarelto. He is an avid outdoorsman and has been avoiding higher risk physical activity in light of his anticoagulation use. He denies pre-syncope or syncope. He has not had chest pain. He denies peripheral edema, orthopnea or PND. ECG June 19, 2020 (personally reviewed) shows atrial fibrillation with IVCD. CARDIOLOGY STUDIES TEE [June 2020]: NO LAA thrombus, LVEF 0.65, trace TR. PAST MEDICAL HISTORY Patient Active Problem List Diagnosis Date Noted • Paroxysmal atrial fibrillation (HCC) [148.0] 07/02/2020 Added automatically from request for surgery 127730 SOCIAL HISTORY: Married. Non-smoker. Rare EtOH. Works for local tech company. FAMILY HISTORY: No heart failure, SCD. ALLERGIES: Patient has no allergy information on record. Current Outpatient Medications Medication Sig Dispense Refill • rivaroxaban (Xarelto) 20 MG tablet Take 1 tablet (20 mg) by mouth daily. 90 tablet 0 No current facility-administered medications for this visit. REVIEW OF SYSTEMS Other than HPI, all other systems are negative. PHYSICAL EXAM VITAL SIGNS: There were no vitals taken for this visit. Appears well IMPRESSION: Mr. Simpson is a 62 year old male with the following salient arrhythmic issues: 1. Atrial fibrillation We discussed atrial fibrillation in detail, including the risks factors, triggers, natural history, stroke risk prediction and treatment options. Specifically, we discussed a rate control strategy with medications versus a rhythm control strategy. For many patients, rate control is sufficient to maintain quality of life and cardiovascular function. We discussed the antiarrhythmic drugs in detail, including class Ic and class III AADs. We also discussed the importance of lifestyle changes in improving both ablative and non-ablative approaches to AF, namely weight reduction and exercise. Given his clearly symptomatic atrial fibrillation we discussed the rationale for rhythm control. Given his low resting heart rate and background significant physical activity, we discussed that AAD options are not likely to be tolerated or less well tolerated. I therefore discussed ablation. I explained that the chances of success with this procedure is about 75% but may need more than one procedure. I explained how a CT or MRI of his heart is completed ahead of time to assess his pulmonary vein anatomy. The procedure is completed under general anesthesia, possibly with a transesophageal echocardiogram. I explained that a transseptal puncture is completed and a wide-area circumferential ablation around the pulmonary veins is completed. I explained that he may go home the same day or may spend one night in the hospital and go home the next day. I said the risks of the procedure included, but were not limited to bleeding, infection, cardiac perforation and tamponade, 1 in 200 chance of a stroke, gastroparesis, pulmonary vein stenosis, atrialesophageal fistula, and 1 in a 1000 chance of death. After discussion, he would like to proceed to catheter ablation. PLAN: 1. AF ablation August 22, 2020 2. Cardiac CT for PV anatomy 3. Continue Xarelto through ablation and minimum 6 weeks after FOLLOW-UP: Post ablation per protocol I spent a total of 60 minutes for the patient's care on the date of the service including chart review, history taking, physical exam, counseling, referrals and communication with other providers, and documentation. Julius M Hibbert, MD Cardiac Electrophysiology Section Division of Cardiology

Give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Answer": "<int>",
    "Explanation": "<str>"
}
{
    "Answer": "2",
    "Explanation": "The patient's ASA Physical Status Classification appears
        to be ASA II. According to the ASA classification system, an ASA II
        patient is defined as a patient with mild systemic disease. In this
        case, the patient has a history of paroxysmal atrial fibrillation,
        which is a chronic condition, but he has normal left ventricular
        function, no valvular heart disease, and is otherwise healthy without
        significant lifestyle constraints. He reports feeling generally well
        since the cardioversion and continues his regular activities as an
        outdoorsman, albeit with caution due to anticoagulation. There are no
        other significant comorbidities or systemic diseases mentioned in the
        medical record that would elevate his classification to ASA III or
        beyond. Meanwhile, his condition is not limited to a specific organ
        system or seems to have a substantive impact on his daily functioning,
        so he would not be classified as ASA I. Therefore, ASA II is the most
        fitting classification."
}
```

### B.1.3 Zero-shot Q&A From Notes Summary

#### Prompt & LLM Output

You are given a task and context. The context contains information from the proposed procedure and patient's medical record. Assess the patient in the context of the proposed procedure and then provide an answer.

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Summary:

62-year-old male patient Homer Simpson with a history of paroxysmal atrial fibrillation presented for evaluation via telemedicine. He experienced palpitations and an episode requiring hospitalization and cardioversion earlier in the year. Post-cardioversion, he feels generally well with no recurrence of rapid heart action, preserved LV function (LVEF 0.65), and no significant abnormalities on recent TEE. He has continued with anticoagulation therapy using

Xarelto. After discussing treatment options, Mr. Simpson elected to proceed with catheter ablation with a scheduled date of August 22, 2020, and plans to continue anticoagulation through ablation and for at least 6 weeks post-procedure.

Give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Answer": "<int>",
    "Explanation": "<str>"
}
{
    "Answer": "2".
    "Explanation": "The ASA Physical Status Classification for this patient
        would be ASA II. The classification ASA II is for patients with mild
        systemic disease. Even though Mr. Simpson experienced palpitations and
        required hospitalization and cardioversion, his current state includes
        preserved left ventricular function, no significant findings on recent
        transesophageal echocardiogram (TEE), and he is stable on
        anticoagulation therapy with Xarelto. There is no mention of any other
        systemic diseases that are poorly controlled. Thus, reflecting a
        patient with a mild systemic disease without substantive functional
        limitations fits the ASA II category."
}
```

### B.1.4 Few-shot Q&A From Notes Summary

#### Prompt & LLM Output

You are given examples of task, context, and answer. The context contains information from the proposed procedure and patient's medical record which can be used to determine the answer.

Task: What is the patient's ASA Physical Status Classification?

#### Context

Procedure: HYSTERECTOMY, TOTAL, VAGINAL, WITH VAGINAL VAULT SUSPENSION

Procedure Description: LAPAROSCOPIC ASSISTED VAGINAL HYSTERECTOMY REMOVAL BILATERALTUBES AND OVARIES WITH UTEROSACRAL LIGAMENT SUSPENSION AND CYSTOSCOPY. (N/A Vagina)

Diagnosis: Uterine prolapse—Cystocele, midline

Provider Service: Gynecology

#### Medical Record Summary:

Marge Simpson is a 61-year-old female with a history of uterine prolapse treated with a pessary for 10-15 years, endometrial hyperplasia managed with progesterone cream, bladder infections, pelvic pressure, lipidemia, and thyroid disease. She is scheduled for a LAVH, BSO, USLS with a preoperative Caprini score of 3 indicating a high risk for VTE, for which heparin and SCDs are planned. She takes levothyroxine and liothyronine for thyroid disease and has documented allergies to nitrofurantoin and levofloxacin, causing GI upset. Physical

examination reveals a well-nourished woman with stable vital signs and no signs of anxiety or agitation.

```
JSON Response: {
    "Answer": "2"
}
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: MICROLARYNGOSCOPY, USING LASER, WITH BRONCHOSCOPY Procedure Description: MICROLARYNGOSCOPY, DILATIN, STEROID INJECTION, WITH BRONCHOSCOPY (Bronchus)

Diagnosis: Subglottic stenosis—Dyspnea, unspecified type—Dysphonia

Provider Service: Otolaryngology

Medical Record Summary:

Selma Bouvier, a 59-year-old female with a long history of subglottic stenosis, has experienced recurrent symptoms of dyspnea, dysphonia, and mucus build-up leading to multiple interventions. Most recently, she underwent a microlaryngoscopy using CO2 laser with bronchoscopy. Postoperative reports indicate significant improvement in breathing and ability to carry out activities with intermittent post-nasal drip and productive cough attributed to allergies. Vital signs are stable, and recent physical exams show slight stridor, no acute distress, and a patent airway with 50-60% subglottic stenosis. Current plan includes consideration of further dilation and potential steroid injections for long-term management. She is not interested in in-office balloon dilation or open tracheal resection at this time.

```
JSON Response: {
    "Answer": "3"
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: LAPAROTOMY, EXPLORATORY

Procedure Description: LAPAROTOMY, EXPLORATORY, PORTAL VEIN REPAIR,

TEMPORARY ABDOMINAL CLOSURE (N/A Abdomen)

Diagnosis: Intraabdominal hemorrhage Provider Service: General Surgery

Medical Record Summary:

Patty Bouvier is a 63-year-old G2P2 female with a history of breast cancer treated with tamoxifen, presenting with post-menopausal bleeding (PMB). Initial assessment revealed likely endometrial polyp and a simple paratubal cyst, with a negative endometrial biopsy for hyperplasia and malignancy. Patient subsequently underwent hysteroscopic polypectomy with pathology showing complex endometrial hyperplasia without atypia. A Mirena IUD was inserted for management of the hyperplasia with endometrial sampling planned for follow-up. Patient later requested IUD removal after discontinuing tamoxifen and reports minimal spotting. Advised to have repeat endometrial biopsy in 4-6 months. The patient's medical history also includes spinal fusion, appendectomy, and surgeries due to an accident; family history includes heart disease and stroke. Notable allergies include Lymphazurin, Vancomycin, Dexamethasone, Lamisil, and Pcn. At the age of 67, she experienced intraabdominal hemorrhage, suspected to be related to a pancreatic head mass with adjacent hemorrhage.

```
JSON Response:
{
    "Answer": "5"
}
Task: What is the patient's ASA Physical Status Classification?
Context:
```

Procedure: SURGICAL PROCUREMENT, ORGAN

Procedure Description: SURGICAL PROCUREMENT, ORGAN / LIVER AND KIDNEYS

(N/A Abdomen)Di Other accident Provider Service: Organ Donor

Medical Record Summary:

Mr. Doe, a 60-year-old male with unknown past medical history, suffered a severe traumatic brain injury from a fall from a ladder. He had an initial Glasgow Coma Scale of 3, was intubated at the scene, and showed fixed and dilated pupils. CT scan revealed a left-sided acute subdural hematoma with midline shift and a diffuse subarachnoid hemorrhage. He was clinically consistent with brain death upon hospital examination, and after discussion with his wife, was admitted to the trauma ICU with DNR status for medical futility. Organ donation was considered, but the family eventually declined brain death testing. He ultimately underwent a formal brain death examination, resulting in a confirmed brain death diagnosis, and was pronounced deceased.

```
JSON Response: {
    "Answer": "6"
}
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Summary:

62-year-old male patient Homer Simpson with a history of paroxysmal atrial fibrillation presented for evaluation via telemedicine. He experienced palpitations and an episode requiring hospitalization and cardioversion earlier in the year. Post-cardioversion, he feels generally well with no recurrence of rapid heart action, preserved LV function (LVEF 0.65), and no significant abnormalities on recent TEE. He has continued with anticoagulation therapy using Xarelto. After discussing treatment options, Mr. Simpson elected to proceed with catheter ablation with a scheduled date of August 22, 2020, and plans to continue anticoagulation through ablation and for at least 6 weeks post-procedure.

Give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Answer": "<int>",
    "Explanation": "<str>"
}
```

```
{
    "Answer": "2",
    "Explanation": "The patient's ASA Physical Status Classification is likely
    2, indicating a patient with mild systemic disease. In this case,
    although the patient has a history of paroxysmal atrial fibrillation
    and an episode requiring cardioversion, he is currently feeling
    generally well with no recurrence of symptoms, preserved left
    ventricular function, and has been stable on anticoagulation therapy."
}
```

### B.1.5 Zero-shot Chain-of-Thought Q&A From Original Notes

#### Prompt & LLM Output

You are given a task and context. The context contains information from the proposed procedure and patient's medical record. Assess the patient in the context of the proposed procedure and then provide an answer.

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Notes:

Progress Notes written by Physician at 2020-07-02 12:23:00:

ELECTROPHYSIOLOGY CLINIC INITIAL CONSULT REASON FOR CONSULT: Atrial fibrillation REFERRING: Self PRIMARY CARE: No primary care provider on file. Distant Site Telemedicine Encounter I conducted this encounter from Marvin Monroe Memorial Hospital via secure, live, face-to-face video conference with the patient. Homer was located at home with his wife. I reviewed the risks and benefits of telemedicine as pertinent to this visit and the patient agreed to proceed. PROBLEM LIST 1. Atrial fibrillation HPI: Mr. Simpson is a 62 year old male referred for evaluation of atrial fibrillation. He has a long standing history of episodic palpitations for several years. After an episode of the same earlier in 2020 he was seen by his PCP who noted that 12 lead ECG showed atrial fibrillation. These episodes were paroxysmal in nature without clear provoking triggers or palliating factors. He had a more sustained episode of palpitations and light headedness on a flight to Springfield where he ultimately was admitted to a hospital for evaluation. ECG and telemetry showed persistent atrial fibrillation and he underwent a TEE cardioversion. He was started on Xarelto and advised to continue for 4 weeks. He had a TTE showing normal LV function and no valvular heart disease. Imaging also showed possible cholecystitis though HIDA scan was most in keeping with biliary dyskinesia. Today he joins me via telemedicine. He has felt generally well since his cardioversion and discharge. He has not had a recurrence of rapid heart action. He has continued Xarelto. He is an avid outdoorsman and has been avoiding higher risk physical activity in light of his anticoagulation use. He denies pre-syncope or syncope. He has not had chest pain. He denies peripheral edema, orthopnea or PND. ECG June 19, 2020 (personally reviewed) shows atrial fibrillation with IVCD. CARDIOLOGY STUDIES TEE [June 2020]: NO LAA thrombus, LVEF 0.65, trace TR. PAST MEDICAL HISTORY Patient Active Problem List Diagnosis Date Noted • Paroxysmal atrial fibrillation (HCC) [148.0] 07/02/2020 Added automatically from request

for surgery 127730 SOCIAL HISTORY: Married. Non-smoker. Rare EtOH. Works for local tech company. FAMILY HISTORY: No heart failure, SCD. ALLERGIES: Patient has no allergy information on record. Current Outpatient Medications Medication Sig Dispense Refill • rivaroxaban (Xarelto) 20 MG tablet Take 1 tablet (20 mg) by mouth daily. 90 tablet 0 No current facility-administered medications for this visit. REVIEW OF SYSTEMS Other than HPI, all other systems are negative. PHYSICAL EXAM VITAL SIGNS: There were no vitals taken for this visit. Appears well IMPRESSION: Mr. Simpson is a 62 year old male with the following salient arrhythmic issues: 1. Atrial fibrillation We discussed atrial fibrillation in detail, including the risks factors, triggers, natural history, stroke risk prediction and treatment options. Specifically, we discussed a rate control strategy with medications versus a rhythm control strategy. For many patients, rate control is sufficient to maintain quality of life and cardiovascular function. We discussed the antiarrhythmic drugs in detail, including class Ic and class III AADs. We also discussed the importance of lifestyle changes in improving both ablative and non-ablative approaches to AF, namely weight reduction and exercise. Given his clearly symptomatic atrial fibrillation we discussed the rationale for rhythm control. Given his low resting heart rate and background significant physical activity, we discussed that AAD options are not likely to be tolerated or less well tolerated. I therefore discussed ablation. I explained that the chances of success with this procedure is about 75% but may need more than one procedure. I explained how a CT or MRI of his heart is completed ahead of time to assess his pulmonary vein anatomy. The procedure is completed under general anesthesia, possibly with a transesophageal echocardiogram. I explained that a transseptal puncture is completed and a wide-area circumferential ablation around the pulmonary veins is completed. I explained that he may go home the same day or may spend one night in the hospital and go home the next day. I said the risks of the procedure included, but were not limited to bleeding, infection, cardiac perforation and tamponade, 1 in 200 chance of a stroke, gastroparesis, pulmonary vein stenosis, atrialesophageal fistula, and 1 in a 1000 chance of death. After discussion, he would like to proceed to catheter ablation. PLAN: 1. AF ablation August 22, 2020 2. Cardiac CT for PV anatomy 3. Continue Xarelto through ablation and minimum 6 weeks after FOLLOW-UP: Post ablation per protocol I spent a total of 60 minutes for the patient's care on the date of the service including chart review, history taking, physical exam, counseling, referrals and communication with other providers, and documentation. Julius M Hibbert, MD Cardiac Electrophysiology Section Division of Cardiology

Think step by step and give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Step By Step Explanation": "<str>",
    "Answer": "<int>"
}
```

```
{
    "Step By Step Explanation": "The ASA Physical Status Classification system
        is used to assess the preoperative health of patients before they
        undergo a procedure. It ranges from ASA I, a normal healthy patient,
        to ASA VI, a brain-dead patient awaiting organ removal. Based on the
        information provided, Mr. Simpson has a history of paroxysmal atrial
        fibrillation, which qualifies as at least a systemic disease without
        substantive functional limitations (ASA II). He is not described as
        having a severe systemic disease (ASA III) since he does not have a
        constant threat to life, nor is he incapacitated (ASA IV). There is no
        mention of emergency surgery (ASA E) or organ donation (ASA VI). Addit |
        ionally, he has normal LV function, no valvular heart disease, and has
        been described as feeling generally well since his cardioversion, with
        no other systemic diseases noted, and is not incapacitated. Even thoug
        h he has been started on Xarelto (anticoagulant therapy) due to A-Fib,
        this does not automatically increase his ASA status if he is otherwise
        healthy. Therefore, he seems to fit the description of ASA II.",
    "Answer": "2"
}
```

### B.1.6 Zero-shot Chain-of-Thought Q&A From Notes Summary

#### Prompt & LLM Output

You are given a task and context. The context contains information from the proposed procedure and patient's medical record. Assess the patient in the context of the proposed procedure and then provide an answer.

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Summary:

62-year-old male patient Homer Simpson with a history of paroxysmal atrial fibrillation presented for evaluation via telemedicine. He experienced palpitations and an episode requiring hospitalization and cardioversion earlier in the year. Post-cardioversion, he feels generally well with no recurrence of rapid heart action, preserved LV function (LVEF 0.65), and no significant abnormalities on recent TEE. He has continued with anticoagulation therapy using Xarelto. After discussing treatment options, Mr. Simpson elected to proceed with catheter ablation with a scheduled date of August 22, 2020, and plans to continue anticoagulation through ablation and for at least 6 weeks post-procedure.

Think step by step and give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Step By Step Explanation": "<str>",
    "Answer": "<int>"
}
```

```
{
    "Step By Step Explanation": "The ASA Physical Status Classification system
        is used to assess and communicate a patient's pre-anesthesia medical
        comorbidities. The classifications are as follows: ASA 1 - Normal
        healthy patient, ASA 2 - Patients with mild systemic disease, ASA 3 -
        Patients with severe systemic disease, ASA 4 - Patients with severe
        systemic disease that is a constant threat to life, ASA 5 - Moribund
        patients who are not expected to survive without the operation. In the
        case of Homer Simpson, a 62-year-old male with paroxysmal atrial
        fibrillation, who experienced an episode requiring hospitalization and
        cardioversion but otherwise feels well with preserved cardiac function
        (LVEF 0.65) and is on anticoagulation therapy, he would be classified
        as ASA 2. This is due to his atrial fibrillation being a mild systemic
        disease; he is not otherwise healthy (ASA 1), yet his disease does not
        constitute a severe systemic disease (ASA 3), and it is not a constant
        threat to his life (ASA 4) nor is he at the brink of death (ASA 5).",
    "Answer": "2"
}
```

### B.1.7 Few-shot Chain-of-Thought Q&A From Notes Summary

#### Prompt & LLM Output

You are given examples of task, context, logical reasoning, and answer. The context contains information from the proposed procedure and patient's medical record which can be used to determine the answer. The logical reasoning contains a step by step explanation leading to the answer.

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: HYSTERECTOMY, TOTAL, VAGINAL, WITH VAGINAL VAULT SUSPENSION

Procedure Description: LAPAROSCOPIC ASSISTED VAGINAL HYSTERECTOMY REMOVAL BILATERALTUBES AND OVARIES WITH UTEROSACRAL LIGAMENT SUSPENSION AND CYSTOSCOPY. (N/A Vagina)

Diagnosis: Uterine prolapse—Cystocele, midline

Provider Service: Gynecology

Medical Record Summary:

Marge Simpson is a 61-year-old female with a history of uterine prolapse treated with a pessary for 10-15 years, endometrial hyperplasia managed with progesterone cream, bladder infections, pelvic pressure, lipidemia, and thyroid disease. She is scheduled for a LAVH, BSO, USLS with a preoperative Caprini score of 3 indicating a high risk for VTE, for which heparin and SCDs are planned. She takes levothyroxine and liothyronine for thyroid disease and has documented allergies to nitrofurantoin and levofloxacin, causing GI upset. Physical examination reveals a well-nourished woman with stable vital signs and no signs of anxiety or agitation.

JSON Response:

"Step By Step Explanation": "The ASA Physical Status Classification System is a system used by anesthesiologists to assess the fitness of patients before surgery. It is important

because it helps predict the risk of complications during or after surgery. The ASA physical status classification is as follows:

- 1. ASA 1: A normal healthy patient.
- 2. ASA 2: A patient with mild systemic disease.
- 3. ASA 3: A patient with severe systemic disease.
- 4. ASA 4: A patient with severe systemic disease that is a constant threat to life.
- 5. ASA 5: A moribund patient who is not expected to survive without the operation.
- 6. ASA 6: A declared brain-dead patient whose organs are being removed for donor purposes. Given the information from the patient's medical record, Marge Simpson has systemic diseases that are well controlled. She has endometrial hyperplasia managed with progesterone cream, bladder infections, pelvic pressure, thyroid disease for which she is taking medications, and lipidemia. While she has multiple medical issues, none are described as 'severe.' Her vital signs are stable, and there is no indication in the provided information that her systemic diseases are a constant threat to her life or a severe systemic disease. Therefore, she does not fit the criteria for ASA 3 or higher.

Additionally, the patient does not qualify for ASA 1, as she does have systemic diseases, albeit mild. The patient's preoperative Caprini score of 3 indicates a high risk for venous thromboembolism (VTE), which is being managed with prophylactic measures such as heparin and sequential compression devices (SCDs). An ASA 2 status is typically assigned to patients who have a mild systemic disease that does not limit their activity.

Based on her controlled medical conditions and her physical examination, which does not reveal signs of severe systemic disease, the correct ASA Physical Status Classification for Marge Simpson is ASA 2.",

```
"Answer": "2"
```

Task: What is the patient's ASA Physical Status Classification?

### Context:

Procedure: MICROLARYNGOSCOPY, USING LASER, WITH BRONCHOSCOPY Procedure Description: MICROLARYNGOSCOPY, DILATIN, STEROID INJECTION, WITH BRONCHOSCOPY (Bronchus)

Diagnosis: Subglottic stenosis—Dyspnea, unspecified type—Dysphonia

Provider Service: Otolaryngology

### Medical Record Summary:

Selma Bouvier, a 59-year-old female with a long history of subglottic stenosis, has experienced recurrent symptoms of dyspnea, dysphonia, and mucus build-up leading to multiple interventions. Most recently, she underwent a microlaryngoscopy using CO2 laser with bronchoscopy. Postoperative reports indicate significant improvement in breathing and ability to carry out activities with intermittent post-nasal drip and productive cough attributed to allergies. Vital signs are stable, and recent physical exams show slight stridor, no acute distress, and a patent airway with 50-60% subglottic stenosis. Current plan includes consideration of further dilation and potential steroid injections for long-term management. She is not interested in in-office balloon dilation or open tracheal resection at this time.

## JSON Response:

"Step By Step Explanation": "The ASA Physical Status Classification System is used to assess and communicate a patient's pre-anesthesia medical comorbidities. This classification ranges from ASA 1 to ASA 6. The ASA 1 indicates a normal healthy patient. ASA 2 indicates a patient with a mild systemic disease. ASA 3 is assigned to a patient with severe systemic disease. ASA 4 is for a patient with a severe systemic disease that is a constant

threat to life. ASA 5 indicates a moribund patient who is not expected to survive without the operation. ASA 6 is declared for a patient who is declared brain-dead and whose organs are being removed for donor purposes. Selma Bouvier's classification as ASA 3 likely stems from several factors mentioned in her medical record:

- 1. She has a 'long history of subglottic stenosis,' which is a chronic condition that affects the airway. This is a significant piece of medical history indicative of a severe systemic disease that has required multiple interventions.
- 2. Ms. Bouvier had 'recurrent symptoms of dyspnea and dysphonia,' symptoms that can limit function and indicate that her systemic disease significantly impacts her health.
- 3. Ms. Bouvier's recent physical exam showed 'slight stridor,' and a 'patent airway with 50-60% subglottic stenosis,' which means that despite previous interventions, there is still notable impairment which could be considered severe.

These points suggest that Ms. Bouvier's medical condition significantly impacts her daily life, thereby fitting the criteria for an ASA 3 classification. Ms. Bouvier does not appear to be at a constant threat to life nor moribund, which would be required for an ASA 4 or ASA 5 classification, respectively.",

```
"Answer": "3"
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: LAPAROTOMY, EXPLORATORY

Procedure Description: LAPAROTOMY, EXPLORATORY, PORTAL VEIN REPAIR,

TEMPORARY ABDOMINAL CLOSURE (N/A Abdomen)

Diagnosis: Intraabdominal hemorrhage Provider Service: General Surgery

Medical Record Summary:

Patty Bouvier is a 63-year-old G2P2 female with a history of breast cancer treated with tamoxifen, presenting with post-menopausal bleeding (PMB). Initial assessment revealed likely endometrial polyp and a simple paratubal cyst, with a negative endometrial biopsy for hyperplasia and malignancy. Patient subsequently underwent hysteroscopic polypectomy with pathology showing complex endometrial hyperplasia without atypia. A Mirena IUD was inserted for management of the hyperplasia with endometrial sampling planned for follow-up. Patient later requested IUD removal after discontinuing tamoxifen and reports minimal spotting. Advised to have repeat endometrial biopsy in 4-6 months. The patient's medical history also includes spinal fusion, appendectomy, and surgeries due to an accident; family history includes heart disease and stroke. Notable allergies include Lymphazurin, Vancomycin, Dexamethasone, Lamisil, and Pcn. At the age of 67, she experienced intraabdominal hemorrhage, suspected to be related to a pancreatic head mass with adjacent hemorrhage.

## JSON Response:

"Step By Step Explanation": "To determine Patty Bouvier's American Society of Anesthesiologists (ASA) Physical Status Classification, we assess the information given about her health status and the urgency and severity of her medical condition. The ASA classification is a system used to evaluate the fitness of patients before surgery. The scale ranges from 1 (a healthy patient) to 6 (a declared brain-dead patient whose organs are being removed for donor purposes). The ASA Physical Status Classification defines class 5 as 'a moribund patient who is not expected to survive without the operation.' This classification applies to patients facing an immediate threat to life. In this context, Ms. Bouvier has suffered an

intraabdominal hemorrhage, which is a life-threatening condition. The procedure planned is an exploratory laparotomy with portal vein repair and temporary abdominal closure, which indicates imminent surgery to control hemorrhage and is likely an emergency situation. Given this information and considering that intraabdominal hemorrhage puts Ms. Bouvier's life in immediate danger, the correct classification would be ASA 5. The other medical history such as breast cancer, endometrial hyperplasia, and previous surgeries also corroborate a more complicated medical background which further supports the assignment of a higher ASA class.",

```
"Answer": "5"
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: SURGICAL PROCUREMENT, ORGAN

Procedure Description: SURGICAL PROCUREMENT, ORGAN / LIVER AND KIDNEYS

(N/A Abdomen)

Diagnosis: Other accident Provider Service: Organ Donor

Medical Record Summary:

Mr. Doe, a 60-year-old male with unknown past medical history, suffered a severe traumatic brain injury from a fall from a ladder. He had an initial Glasgow Coma Scale of 3, was intubated at the scene, and showed fixed and dilated pupils. CT scan revealed a left-sided acute subdural hematoma with midline shift and a diffuse subarachnoid hemorrhage. He was clinically consistent with brain death upon hospital examination, and after discussion with his wife, was admitted to the trauma ICU with DNR status for medical futility. Organ donation was considered, but the family eventually declined brain death testing. He ultimately underwent a formal brain death examination, resulting in a confirmed brain death diagnosis, and was pronounced deceased.

### JSON Response:

{

"Step By Step Explanation": "The ASA Physical Status Classification System is a way for anesthesiologists to categorize a patient's preoperative health. This system ranges from ASA 1, denoting a normal healthy patient, to ASA 6, indicating a declared brain-dead patient whose organs are being removed for donor purposes. In this particular case, we can determine the ASA classification by following these logical steps:

- 1. Identify the patient's current medical status from the medical record summary provided. Mr. Doe has been confirmed to be brain-dead following a severe traumatic brain injury and a conclusive formal brain death examination.
- 2. Recognize the type of procedure described. The text mentions 'SURGICAL PROCURE-MENT, ORGAN / LIVER AND KIDNEYS' which indicates organ retrieval for transplantation purposes.
- 3. Understand that Mr. Doe is designated as an organ donor ('Provider Service: Organ Donor'), and the procedure is taking place posthumously after he was pronounced deceased.
- 4. Recall that the ASA classification of '6' corresponds specifically to patients who are declared brain-dead and are organ donors, according to the definition of ASA 6.
- 5. Conclude that the correct ASA Physical Status Classification for Mr. Doe is ASA 6, based on his brain-dead status and the fact that his organs are being procured for donation. This aligns with the provided answer.",

```
"Answer": "6"
```

Task: What is the patient's ASA Physical Status Classification?

Context:

Procedure: A-FIB ABLATION

Procedure Description: A-Fib Ablation (N/A) Diagnosis: Paroxysmal atrial fibrillation (HCC)

Provider Service: Cardiovascular

Medical Record Summary:

62-year-old male patient Homer Simpson with a history of paroxysmal atrial fibrillation presented for evaluation via telemedicine. He experienced palpitations and an episode requiring hospitalization and cardioversion earlier in the year. Post-cardioversion, he feels generally well with no recurrence of rapid heart action, preserved LV function (LVEF 0.65), and no significant abnormalities on recent TEE. He has continued with anticoagulation therapy using Xarelto. After discussing treatment options, Mr. Simpson elected to proceed with catheter ablation with a scheduled date of August 22, 2020, and plans to continue anticoagulation through ablation and for at least 6 weeks post-procedure.

Think step by step and give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

```
JSON Response:
{
    "Step By Step Explanation": "<str>",
    "Answer": "<int>"
}
```

{ "Step By Step Explanation": "The ASA Physical Status Classification System is a way to assess and communicate a patient's pre-anesthesia medical co-morbidities. The categories are as follows: \n\n- ASA 1: A normal healthy patient.\n- ASA 2: A patient with mild systemic disease.\n-ASA 3: A patient with severe systemic disease.\n- ASA 4: A patient with severe systemic disease that is a constant threat to life.\n- ASA 5: A moribund patient who is not expected to survive without the operation.\n- ASA 6: A declared brain-dead patient whose organs are being removed for donor purposes.\n\nHomer Simpson is a 62-year-old male with a history of paroxysmal atrial fibrillation. This condition is considered a systemic disease which can affect multiple systems in the body; however, it is noted that his left ventricular ejection fraction (LVEF) is preserved at 0.65 and there are no significant abnormalities on recent transesophageal echocardiogram (TEE). Additionally, there has been no recurrence of rapid heart action since cardioversion, and he has been managing his condition with anticoagulation therapy.\n\nAtrial fibrillation by itself can vary in severity, but given that the patient feels generally well, has preserved LV function, and is currently stable on medication, it indicates a 'mild systemic disease'. No other systemic diseases or significant health issues are mentioned. The patient's ability to actively participate in telemedicine consultation and make an informed treatment choice further suggests that his systemic disease does not incapacitate him. \n\nAccording to the details provided, he does not have a severe systemic disease that dramatically limits his activity or poses a constant threat to life, which excludes classifications ASA 3 and ASA 4. Therefore, the correct ASA classification for Mr. Simpson would be ASA 2. This classification reflects a patient who has a mild systemic disease without substantive functional limitations.", "Answer": "2" }

### B.1.8 Summary Generation

### Prompt & LLM Output

You are given information from the patient's medical record. Summarize this information, making sure to include the most important positive clinical findings.

### Progress Notes written by Physician at 2020-07-02 12:23:00:

ELECTROPHYSIOLOGY CLINIC INITIAL CONSULT REASON FOR CONSULT: Atrial fibrillation REFERRING: Self PRIMARY CARE: No primary care provider on file. Distant Site Telemedicine Encounter I conducted this encounter from Marvin Monroe Memorial Hospital via secure, live, face-to-face video conference with the patient. Homer was located at home with his wife. I reviewed the risks and benefits of telemedicine as pertinent to this visit and the patient agreed to proceed. PROBLEM LIST 1. Atrial fibrillation HPI: Mr. Simpson is a 62 year old male referred for evaluation of atrial fibrillation. He has a long standing history of episodic palpitations for several years. After an episode of the same earlier in 2020 he was seen by his PCP who noted that 12 lead ECG showed atrial fibrillation. These episodes were paroxysmal in nature without clear provoking triggers or palliating factors. He had a more sustained episode of palpitations and light headedness on a flight to Springfield where he ultimately was admitted to a hospital for evaluation. ECG

and telemetry showed persistent atrial fibrillation and he underwent a TEE cardioversion. He was started on Xarelto and advised to continue for 4 weeks. He had a TTE showing normal LV function and no valvular heart disease. Imaging also showed possible cholecystitis though HIDA scan was most in keeping with biliary dyskinesia. Today he joins me via telemedicine. He has felt generally well since his cardioversion and discharge. He has not had a recurrence of rapid heart action. He has continued Xarelto. He is an avid outdoorsman and has been avoiding higher risk physical activity in light of his anticoagulation use. He denies pre-syncope or syncope. He has not had chest pain. He denies peripheral edema, orthopnea or PND. ECG June 19, 2020 (personally reviewed) shows atrial fibrillation with IVCD. CARDIOLOGY STUDIES TEE [June 2020]: NO LAA thrombus, LVEF 0.65, trace TR. PAST MEDICAL HISTORY Patient Active Problem List Diagnosis Date Noted • Paroxysmal atrial fibrillation (HCC) [148.0] 07/02/2020 Added automatically from request for surgery 127730 SOCIAL HISTORY: Married. Non-smoker. Rare EtOH. Works for local tech company. FAMILY HISTORY: No heart failure, SCD. ALLERGIES: Patient has no allergy information on record. Current Outpatient Medications Medication Sig Dispense Refill • rivaroxaban (Xarelto) 20 MG tablet Take 1 tablet (20 mg) by mouth daily. 90 tablet 0 No current facility-administered medications for this visit. REVIEW OF SYSTEMS Other than HPI, all other systems are negative. PHYSICAL EXAM VITAL SIGNS: There were no vitals taken for this visit. Appears well IMPRESSION: Mr. Simpson is a 62 year old male with the following salient arrhythmic issues: 1. Atrial fibrillation We discussed atrial fibrillation in detail, including the risks factors, triggers, natural history, stroke risk prediction and treatment options. Specifically, we discussed a rate control strategy with medications versus a rhythm control strategy. For many patients, rate control is sufficient to maintain quality of life and cardiovascular function. We discussed the antiarrhythmic drugs in detail, including class Ic and class III AADs. We also discussed the importance of lifestyle changes in improving both ablative and non-ablative approaches to AF, namely weight reduction and exercise. Given his clearly symptomatic atrial fibrillation we discussed the rationale for rhythm control. Given his low resting heart rate and background significant physical activity, we discussed that AAD options are not likely to be tolerated or less well tolerated. I therefore discussed ablation. I explained that the chances of success with this procedure is about 75% but may need more than one procedure. I explained how a CT or MRI of his heart is completed ahead of time to assess his pulmonary vein anatomy. The procedure is completed under general anesthesia, possibly with a transesophageal echocardiogram. I explained that a transseptal puncture is completed and a wide-area circumferential ablation around the pulmonary veins is completed. I explained that he may go home the same day or may spend one night in the hospital and go home the next day. I said the risks of the procedure included, but were not limited to bleeding, infection, cardiac perforation and tamponade, 1 in 200 chance of a stroke, gastroparesis, pulmonary vein stenosis, atrialesophageal fistula, and 1 in a 1000 chance of death. After discussion, he would like to proceed to catheter ablation. PLAN: 1. AF ablation August 22, 2020 2. Cardiac CT for PV anatomy 3. Continue Xarelto through ablation and minimum 6 weeks after FOLLOW-UP: Post ablation per protocol I spent a total of 60 minutes for the patient's care on the date of the service including chart review, history taking, physical exam, counseling, referrals and communication with other providers, and documentation. Julius M Hibbert, MD Cardiac Electrophysiology Section Division of Cardiology

Give your response in JSON format using the provided template. The desired response type is provided in angle brackets < >. For example, <int> means to provide an integer response. Provide a single value response without ranges.

JSON Response:

```
{
    "Brief Patient Summary": "<str>"
}

{
    "Brief Patient Summary": "62-year-old male patient Homer Simpson with a history of paroxysmal atrial fibrillation presented for evaluation via telemedicine. He experienced palpitations and an episode requiring hospitalization and cardioversion earlier in the year. Post-cardiovers ion, he feels generally well with no recurrence of rapid heart action, preserved LV function (LVEF 0.65), and no significant abnormalities on recent TEE. He has continued with anticoagulation therapy using Xarelto. After discussing treatment options, Mr. Simpson elected to proceed with catheter ablation with a scheduled date of August 22, 2020, and plans to continue anticoagulation through ablation and for at least 6 weeks post-procedure."

}
```

### B.1.9 Chain-of-thought Rationale Generation

#### Prompt & LLM Output

You are given a task, the answer, and context which contains information from the proposed procedure and patient's medical record. Provide the logical reasoning steps that lead to the answer using information from the proposed procedure and patient's medical record.

Task: What is the patient's ASA Physical Status Classification?

Answer: 5
Context:

Procedure: REPAIR, ANEURYSM, AORTA, THORACIC, ASCENDING

Procedure Description: REPAIR, ANEURYSM, AORTA, THORACIC, ASCENDING

(Chest)

Diagnosis: Dissection of ascending aorta (HCC)

Provider Service: Cardiovascular Surgery

Medical Record Summary:

Barney Gumble is a 56-year-old male with a history of alcohol use disorder who suffered a ground-level fall resulting in a right frontal lobe contusion, subarachnoid hemorrhage, and a Type A aortic dissection. He was found with elevated blood alcohol level, hypotension, and altered mental status necessitating intubation. Imaging revealed aneurysmal dilatation of the aortic root and dissection involving major arteries. He was admitted to ICU for stabilization and impulse control with IV antihypertensives, and was evaluated for potential surgical intervention for the aortic dissection. The patient showed neurological improvement with response to commands while sedated, and had stable vitals without evidence of malperfusion or ischemic complications from the dissection at the time. There is an incidental finding of a right upper lobe nodule and small bowel intussusception.

```
JSON Response:
{
    "Step By Step Explanation": "<str>"
}
```

"Step By Step Explanation": "The ASA Physical Status Classification System classifies a patient's preoperative health to help assess anesthetic risk before surgery. The system ranges from ASA 1 (a normal healthy patient) to ASA 6 (a brain-dead patient whose organs are being removed for donor purposes). ASA 5 is classified as a moribund patient who is not expected to survive without the operation. Here are the logical reasoning steps that lead to the answer, ASA 5, for Barney Gumble based on the information provided in the medical record summary:\n\nStep 1: Identify the patient's medical condition. The patient has suffered a Type A aortic dissection, which is a life-threatening condition that can lead to death if not treated promptly and effectively. \n\nStep 2: Assess the urgency and severity of the underlying condition. A Type A aortic dissection involves the ascending aorta and is often considered a surgical emergency due to the risk of rapid deterioration and fatal outcomes.\n\nStep 3: Evaluate the patient's overall health and comorbidities. Barney has a history of alcohol use disorder. He has also suffered a fall resulting in a right frontal lobe contusion and a subarachnoid hemorrhage. These additional health issues add to his operative risk.\n\nStep 4: Consider the patient's current status. At the time of evaluation, Barney had been admitted to ICU for stabilization, was intubated due to altered mental status which further suggests severe systemic disease and an unstable condition.\n\nStep 5: Determine if the patient is expected to survive without the operation. Given the life-threatening nature of a Type A aortic dissection and the fact that Barney was in the ICU for impulse control with IV antihypertensives, it can be inferred that without the surgery, Barney's chance of survival is poor.\n\nStep 6: Confirm the lack of less acute findings that could suggest a less critical state. The incidental finding of a right upper lobe nodule and small bowel intussusception do not mitigate the acute and critical nature of the patient's condition.\n\nStep 7: Assign the appropriate ASA classification. Based on Barney's critical state, need for urgent surgery, life-threatening diagnosis, and the expectation that he would not survive without the operation, he fits the criteria for ASA 5. Therefore, the assignment of ASA 5 is justified."

{

}

### B.2 Supplemental Figure 2: Dataset Creation Flow Diagram

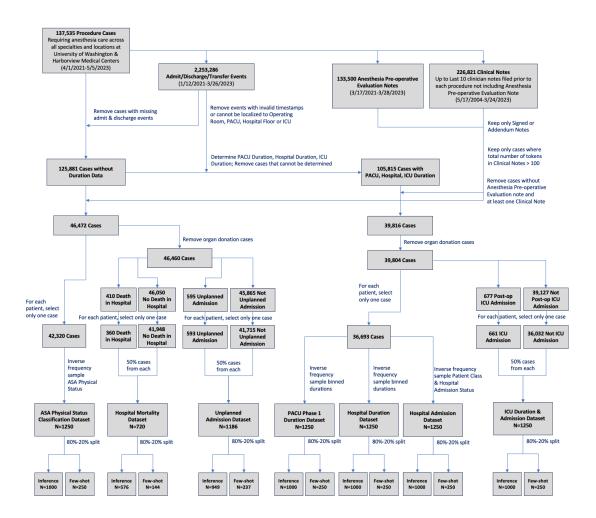


Fig. B1 Flow diagram showing how the task-specific datasets were constructed from Electronic Health Record data. The natural occurrence of certain outcomes such as ICU admission, unplanned admission, and hospital mortality are rare, so datasets are constructed to balance the task label. If patients have multiple procedure cases, only a single case for that patient was included in the final dataset. "Clinical Notes" refers to up to the last 10 clinician-written notes filed prior to each procedure, excluding notes directly associated with the procedure itself. Due to the rarity of ICU admission, the datasets for ICU Duration and Admission tasks are identical.

### B.3 Supplemental Figure 3: Dataset Case Overlap

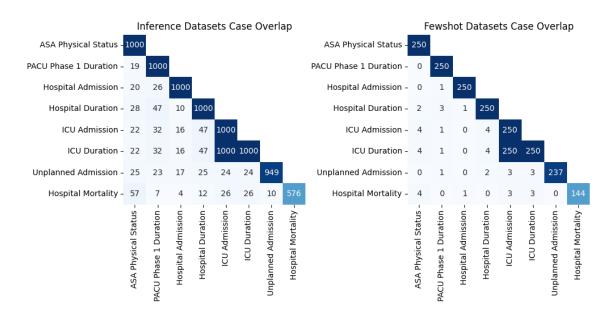


Fig. B2 Number of cases that overlap between any of the task-specific datasets in the inference and few-shot data splits. Datasets for ICU admission and ICU duration prediction tasks are identical. Most datasets have <5% overlap in cases.

## B.4 Supplemental Figure 4: ASA Physical Status - Statistical Significance

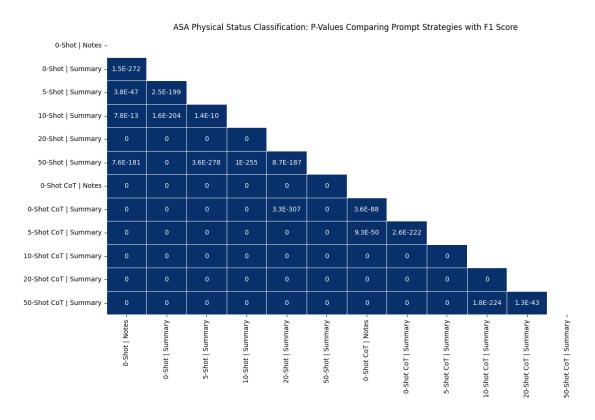


Fig. B3 P-values comparing prompt strategies with F1 score for ASA physical status classification prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

# B.5 Supplemental Figure 5: Hospital Admission - Statistical Significance

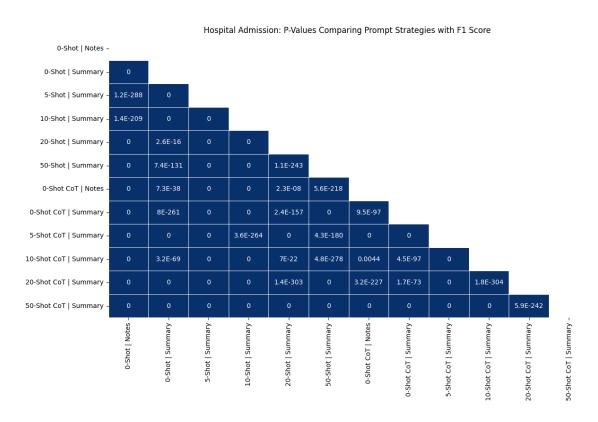


Fig. B4 P-values comparing prompt strategies with F1 score for hospital admission prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

## B.6 Supplemental Figure 6: ICU Admission - Statistical Significance

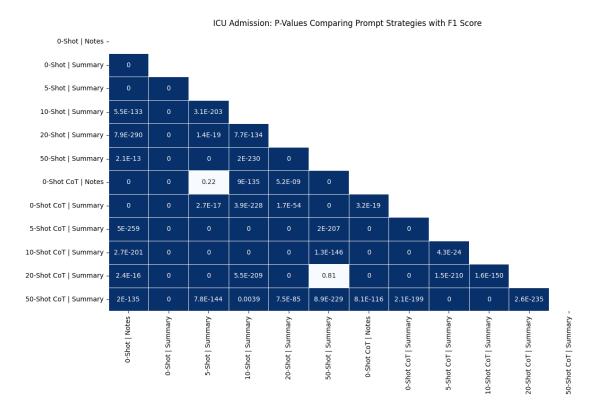


Fig. B5 P-values comparing prompt strategies with F1 score for ICU admission prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

## B.7 Supplemental Figure 7: Unplanned Admission - Statistical Significance

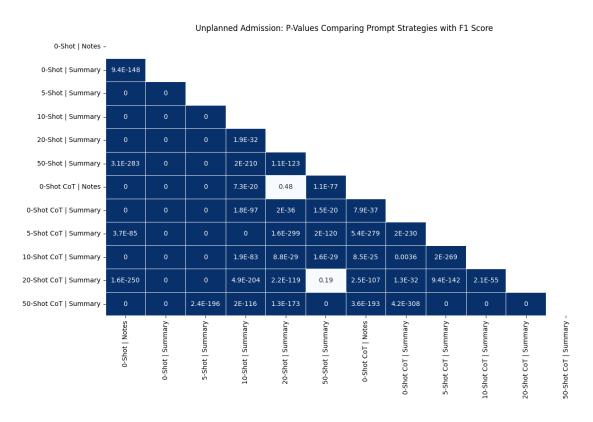


Fig. B6 P-values comparing prompt strategies with F1 score for unplanned admission prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

## B.8 Supplemental Figure 8: Hospital Mortality - Statistical Significance

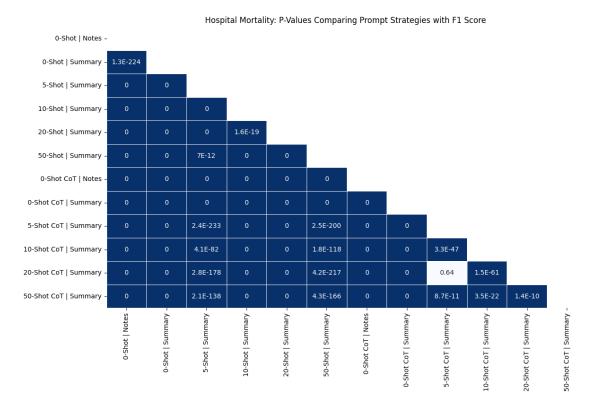


Fig. B7 P-values comparing prompt strategies with F1 score for hospital mortality prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

## B.9 Supplemental Figure 9: Post-Anesthesia Care Unit (PACU) - Phase 1 Duration - Statistical Significance

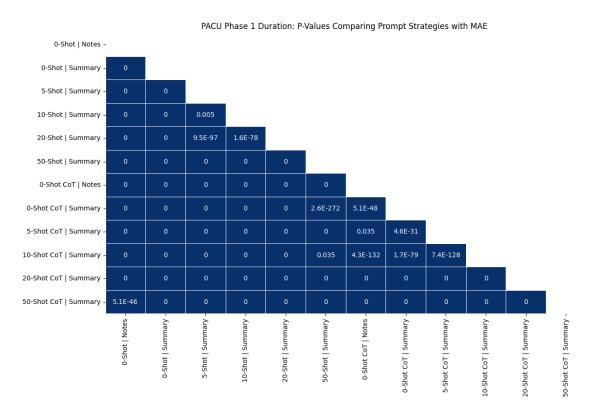


Fig. B8 P-values comparing prompt strategies with mean absolute error (minutes) for PACU phase 1 duration prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

# B.10 Supplemental Figure 10: Hospital Duration - Statistical Significance

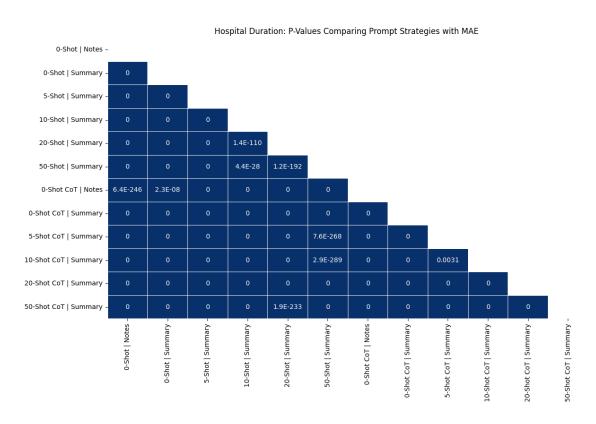


Fig. B9 P-values comparing prompt strategies with mean absolute error (days) for hospital duration prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha = 0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

# B.11 Supplemental Figure 11: Intensive Care Unit (ICU) Duration - Statistical Significance

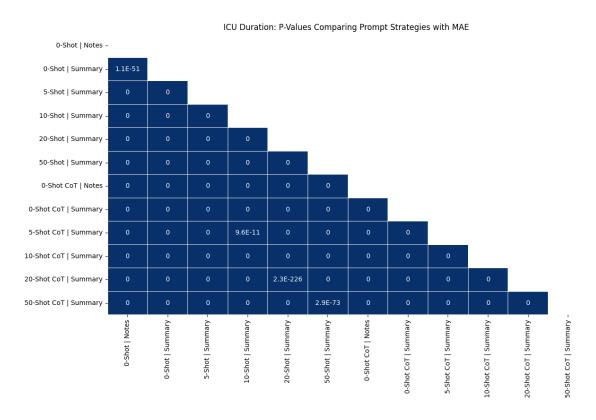


Fig. B10 P-values comparing prompt strategies with mean absolute error (days) for ICU duration prediction. P-values are computed using Wilcoxon signed-rank test with Benjamini-Hochberg procedure for false-discovery rate control level  $\alpha=0.05$ . Statistically significant p-values < 0.05 are depicted in cells with a dark blue background.

## B.12 Supplemental Figure 12: ASA Physical Status - Predicted vs. Actual for Each Prompt Strategy

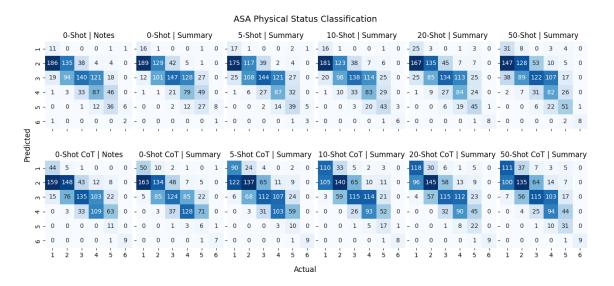


Fig. B11 Confusion matrix of predicted and actual ASA Physical Status (ASA-PS) across all 12 prompt strategies. Without few-shot and CoT prompting, predictions are heavily localized to predicting ASA 2 & 3 (the most common classifications) and the LLM rarely predicts ASA 1 and ASA 6. The addition of few-shot and CoT prompting helps the LLM correctly predict ASA 1 & 6. When the LLM does make a mis-classification error, it generally predicts an adjacent ASA-PS class and rarely makes egregious prediction errors.

## B.13 Supplemental Figure 13: Hospital Admission - Predicted vs. Actual for Each Prompt Strategy

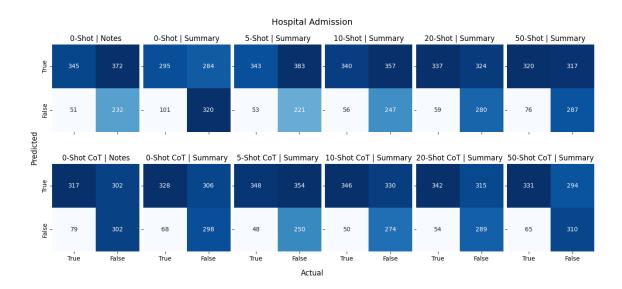


Fig. B12 Confusion matrix of predicted and actual postoperative hospital admission across all 12 prompt strategies. All prompt strategies exhibit a high number of false positives, though this is reduced with the addition of few-shot and CoT prompting. The number of false negatives remains low across all prompt strategies.

## B.14 Supplemental Figure 14: ICU Admission - Predicted vs. Actual for Each Prompt Strategy

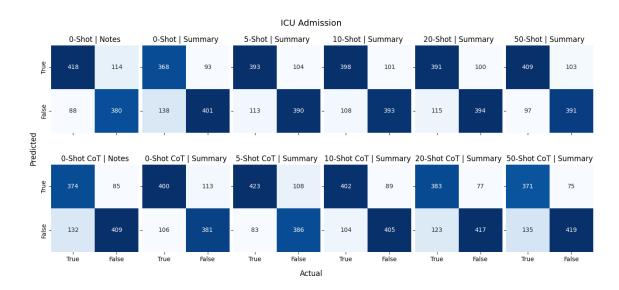


Fig. B13 Confusion matrix of predicted and actual postoperative ICU admission across all 12 prompt strategies. All prompt strategies exhibit a high number of true positive and true negatives relative to false positive and false negatives.

# B.15 Supplemental Figure 15: Unplanned Admission - Predicted vs. Actual for Each Prompt Strategy

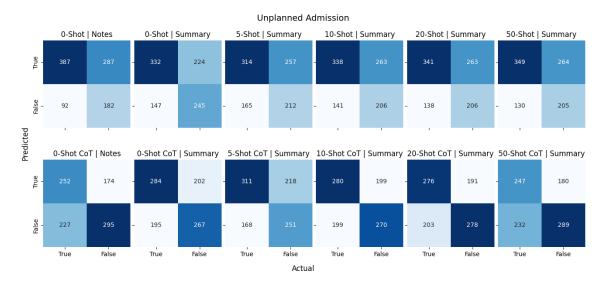


Fig. B14 Confusion matrix of predicted and actual postoperative unplanned admissions across all 12 prompt strategies. Zero-shot approaches without CoT have the best performance with low false negatives, but high false positives. Few-shot and CoT strategies steer the LLM from predicting unplanned admissions and increase the number of "False" predictions. Even though this improves the true negative rate, it results in a significant increase in false negatives and worsens overall prediction performance.

# B.16 Supplemental Figure 16: Hospital Mortality - Predicted vs. Actual for Each Prompt Strategy

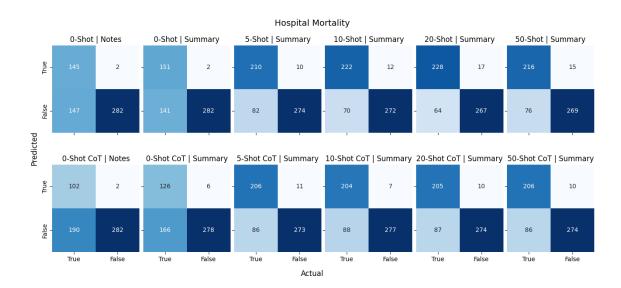


Fig. B15 Confusion matrix of predicted and actual postoperative hospital mortality across all 12 prompt strategies. The LLM rarely makes a false positive prediction, regardless of prompt strategy. The addition of few-shot and CoT dramatically reduces the number of false negatives and results in improved overall prediction performance.

## B.17 Supplemental Figure 17: PACU Phase 1 Duration - Predicted vs. Actual for Each Prompt Strategy

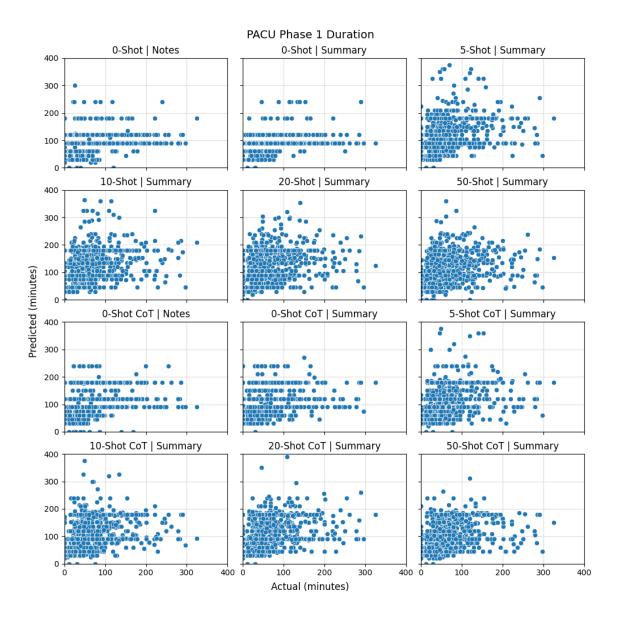


Fig. B16 This is the same figure as Figure 4, replicated here for completeness. Scatter plot of predicted and actual post-anesthesia care unit (PACU) Phase 1 recovery durations across all 12 prompt strategies. Without few-shot and CoT prompting, predictions are heavily quantized to specific values and exhibit a ceiling effect where the LLM rarely predicts beyond 180 minutes. The progressive addition of few-shot and CoT prompting removes this effect, but predictive performance remains poor.

## B.18 Supplemental Figure 18: Hospital Duration - Predicted vs. Actual for Each Prompt Strategy

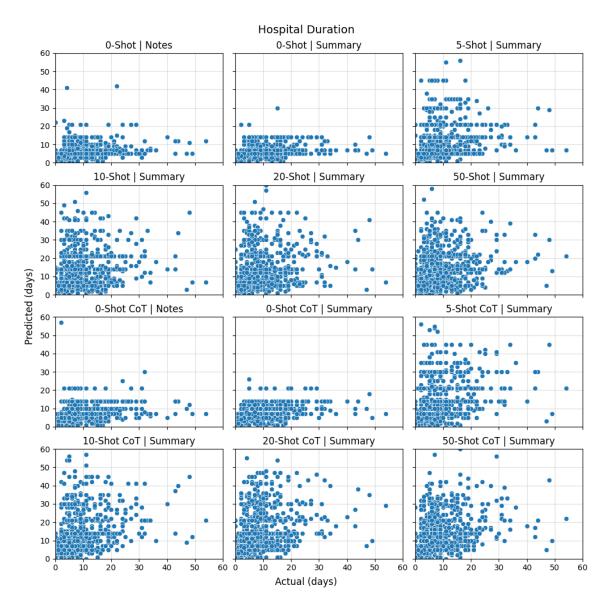


Fig. B17 Scatter plot of predicted versus actual postoperative hospital admission durations across all 12 prompt strategies. Without few-shot and CoT prompting, predictions are heavily quantized to specific values and exhibit a ceiling effect where the LLM rarely predicts beyond 14 days. The progressive addition of few-shot and CoT prompting removes this effect, but predictive performance remains poor.

## B.19 Supplemental Figure 19: ICU Duration - Predicted vs. Actual for Each Prompt Strategy

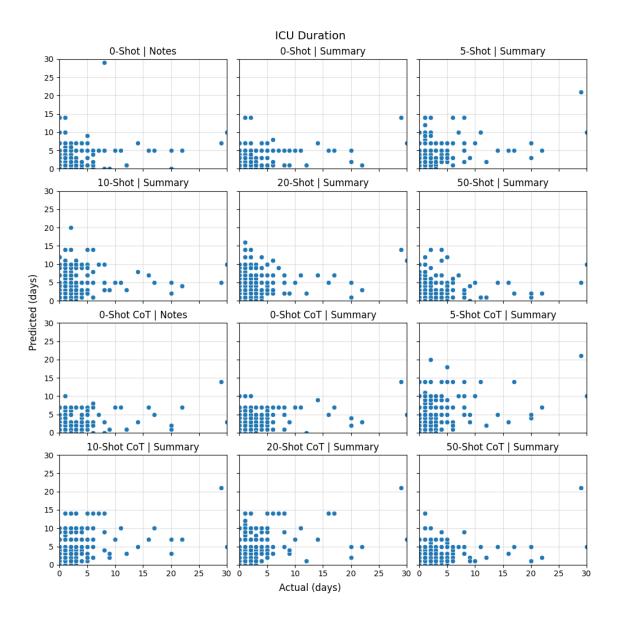


Fig. B18 Scatter plot of predicted versus actual postoperative ICU admission durations across all 12 prompt strategies. Without few-shot prompting, predictions exhibit a ceiling effect where the LLM rarely predicts beyond 10 days. The progressive addition of few-shot prompting removes this effect, but predictive performance remains poor.

## References

- [1] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20, pp. 1877–1901. Curran Associates Inc., Red Hook, NY, USA (2020)
- [2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022) arXiv:2203.02155 [cs.CL]
- [3] Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., Petzold, L.R.: AlpaCare:Instruction-tuned large language models for medical application (2023) arXiv:2310.14558 [cs.CL]
- [4] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://crfm.stanford.edu/2023/03/13/alpaca.html (2023)
- [5] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large language models are Few-Shot clinical information extractors (2022) arXiv:2205.12689 [cs.CL]
- [6] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera Y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. Nature 620(7972), 172–180 (2023)
- [7] Ramachandran, G.K., Lybarger, K., Liu, Y., Mahajan, D., Liang, J.J., Tsou, C.-H., Yetisgen, M., Uzuner, Ö.: Extracting medication changes in clinical narratives using pre-trained language models. J. Biomed. Inform. 139, 104302 (2023)
- [8] Ramachandran, G.K., Fu, Y., Han, B., Lybarger, K., Dobbins, N., Uzuner, O., Yetisgen, M.: Prompt-based extraction of social determinants of health using few-shot learning. In: Proceedings of the 5th Clinical Natural Language Processing Workshop, pp. 385–393. Association for Computational Linguistics, Toronto, Canada (2023)
- [9] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B.: Benchmarking large language models for news summarization (2023) arXiv:2301.13848 [cs.CL]
- [10] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (2020) arXiv:2009.01325 [cs.CL]
- [11] Wu, J., Ouyang, L., Ziegler, D.M., Stiennon, N., Lowe, R., Leike, J., Christiano, P.: Recursively summarizing books with human feedback (2021) arXiv:2109.10862 [cs.CL]
- [12] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models (2022) arXiv:2206.07682 [cs.CL]

- [13] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought prompting elicits reasoning in large language models (2022) arXiv:2201.11903 [cs.CL]
- [14] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are Zero-Shot reasoners (2022) arXiv:2205.11916 [cs.CL]
- [15] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models (2022) arXiv:2210.03629 [cs.CL]
- [16] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models (2023) arXiv:2305.10601 [cs.CL]
- [17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. undefined (2019)
- [18] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Sara Mahdavi, S., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., Natarajan, V.: Towards Expert-Level medical question answering with large language models (2023) arXiv:2305.09617 [cs.CL]
- [19] Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S.M., Ness, R.O., Poon, H., Qin, T., Usuyama, N., White, C., Horvitz, E.: Can generalist foundation models outcompete Special-Purpose tuning? case study in medicine (2023) arXiv:2311.16452 [cs.CL]
- [20] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of GPT-4 on medical challenge problems (2023) arXiv:2303.13375 [cs.CL]
- [21] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Collins, W., Ahuja, N., Langlotz, C.P., Hom, J., Gatidis, S., Pauly, J., Chaudhari, A.S.: Clinical text summarization: Adapting large language models can outperform human experts (2023) arXiv:2309.07430 [cs.CL]
- [22] Saklad, M.: Grading of patients for surgical procedures. Anesthesiology 2(3), 281–284 (1941)
- [23] Mayhew, D., Mendonca, V., Murthy, B.V.S.: A review of ASA physical status historical perspectives and modern developments. Anaesthesia **74**(3), 373–379 (2019)
- [24] Horvath, B., Kloesel, B., Todd, M.M., Cole, D.J., Prielipp, R.C.: The evolution, current value, and future of the american society of anesthesiologists physical status classification system. Anesthesiology 135(5), 904–919 (2021)
- [25] Wang, Y., Ma, X., Chen, W.: Augmenting black-box LLMs with medical textbooks for clinical question answering (2023) arXiv:2309.02233 [cs.CL]
- [26] Zakka, C., Chaurasia, A., Shad, R., Dalal, A.R., Kim, J.L., Moor, M., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Nelson, J., Hiesinger, W.: Almanac: Retrieval-Augmented language models for clinical medicine (2023) arXiv:2303.01229 [cs.CL]
- [27] Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD

- statement. BMC Med. 13, 1 (2015)
- [28] OpenAI: GPT-4 technical report. arXiv [cs.CL] (2023)
- [29] Cuvillon, P., Nouvellon, E., Marret, E., Albaladejo, P., Fortier, L.-P., Fabbro-Perray, P., Malinovsky, J.-M., Ripart, J.: American society of anesthesiologists' physical status system: a multicentre francophone study to analyse reasons for classification disagreement. Eur. J. Anaesthesiol. 28(10), 742–747 (2011)
- [30] Sankar, A., Johnson, S.R., Beattie, W.S., Tait, G., Wijeysundera, D.N.: Reliability of the american society of anesthesiologists physical status scale in clinical practice. Br. J. Anaesth. 113(3), 424–432 (2014)
- [31] Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-E: An embodied multimodal language model (2023) arXiv:2303.03378 [cs.LG]
- [32] Belyaeva, A., Cosentino, J., Hormozdiari, F., Eswaran, K., Shetty, S., Corrado, G., Carroll, A., McLean, C.Y., Furlotte, N.A.: Multimodal LLMs for health grounded in individual-specific data (2023) arXiv:2307.09018 [q-bio.QM]
- [33] Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.-H., Kiraly, A., Kazemzadeh, S., Melamed, Z., Park, J., Strachan, P., Liu, Y., Lau, C., Singh, P., Chen, C., Etemadi, M., Kalidindi, S.R., Matias, Y., Chou, K., Corrado, G.S., Shetty, S., Tse, D., Prabhakara, S., Golden, D., Pilgrim, R., Eswaran, K., Sellergren, A.: ELIXR: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders (2023) arXiv:2308.01317 [cs.CV]
- [34] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., Virmani, S., Semturs, C., Sara Mahdavi, S., Green, B., Dominowska, E., Arcas, B., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Singhal, K., Florence, P., Karthikesalingam, A., Natarajan, V.: Towards generalist biomedical AI (2023) arXiv:2307.14334 [cs.CL]
- [35] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for Few-Shot learning (2022) arXiv:2204.14198 [cs.CV]
- [36] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E.P., Rajpurkar, P., Leskovec, J.: Med-Flamingo: a multimodal medical few-shot learner (2023) arXiv:2307.15189 [cs.CV]
- [37] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: PaLI: A Jointly-Scaled multilingual Language-Image model (2022) arXiv:2209.06794 [cs.CV]
- [38] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda,

- N., Scialom, T.: Toolformer: Language models can teach themselves to use tools (2023) arXiv:2302.04761 [cs.CL]
- [39] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., Sun, M.: ToolLLM: Facilitating large language models to master 16000+ real-world APIs (2023) arXiv:2307.16789 [cs.AI]
- [40] Cai, T., Wang, X., Ma, T., Chen, X., Zhou, D.: Large language models as tool makers (2023) arXiv:2305.17126 [cs.LG]
- [41] Goodell, A.J., Chu, S.N., Rouholiman, D., Chu, L.F.: Augmentation of ChatGPT with clinician-informed tools improves performance on medical calculation tasks (2023)
- [42] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts (2023) arXiv:2307.03172 [cs.CL]
- [43] Mudumbai, S.C., Pershing, S., Bowe, T., Kamal, R.N., Sears, E.D., Finlay, A.K., Eisenberg, D., Hawn, M.T., Weng, Y., Trickey, A.W., Mariano, E.R., Harris, A.H.S.: Development and validation of a predictive model for american society of anesthesiologists physical status. BMC Health Serv. Res. 19(1), 859 (2019)
- [44] Graeßner, M., Jungwirth, B., Frank, E., Schaller, S.J., Kochs, E., Ulm, K., Blobner, M., Ulm, B., Podtschaske, A.H., Kagerbauer, S.M.: Enabling personalized perioperative risk prediction by using a machine-learning model based on preoperative data. Sci. Rep. 13(1), 7128 (2023)
- [45] Lee, S.W., Lee, H.-C., Suh, J., Lee, K.H., Lee, H., Seo, S., Kim, T.K., Lee, S.-W., Kim, Y.-J.: Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. NPJ Digit Med 5(1), 91 (2022)
- [46] Hill, B.L., Brown, R., Gabel, E., Rakocz, N., Lee, C., Cannesson, M., Baldi, P., Olde Loohuis, L., Johnson, R., Jew, B., Maoz, U., Mahajan, A., Sankararaman, S., Hofer, I., Halperin, E.: An automated machine learning-based model predicts postoperative mortality using readilyextractable preoperative electronic health record data. Br. J. Anaesth. 123(6), 877–886 (2019)
- [47] Bilimoria, K.Y., Liu, Y., Paruch, J.L., Zhou, L., Kmiecik, T.E., Ko, C.Y., Cohen, M.E.: 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. 217(5), 833–84213 (2013)
- [48] Chen, P.-F., Chen, L., Lin, Y.-K., Li, G.-H., Lai, F., Lu, C.-W., Yang, C.-Y., Chen, K.-C., Lin, T.-Y.: Predicting postoperative mortality with deep neural networks and natural language processing: Model development and validation. JMIR Med Inform **10**(5), 38241 (2022)
- [49] Xu, Z., Yao, S., Jiang, Z., Hu, L., Huang, Z., Zeng, Q., Liu, X.: Development and validation of a prediction model for postoperative intensive care unit admission in patients with non-cardiac surgery. Heart Lung **62**, 207–214 (2023)
- [50] Meguid, R.A., Bronsert, M.R., Juarez-Colunga, E., Hammermeister, K.E., Henderson, W.G.: Surgical risk preoperative assessment system (SURPAS): III. accurate preoperative prediction of 8 adverse outcomes using 8 predictor variables. Ann. Surg. 264(1), 23–31 (2016)
- [51] Tully, J.L., Zhong, W., Simpson, S., Curran, B.P., Macias, A.A., Waterman, R.S., Gabriel, R.A.: Machine learning prediction models to reduce length of stay at ambulatory surgery

- centers through case resequencing. J. Med. Syst. 47(1), 71 (2023)
- [52] Fang, F., Liu, T., Li, J., Yang, Y., Hang, W., Yan, D., Ye, S., Wu, P., Hu, Y., Hu, Z.: A novel nomogram for predicting the prolonged length of stay in post-anesthesia care unit after elective operation. BMC Anesthesiol. **23**(1), 404 (2023)
- [53] Gabriel, R.A., Waterman, R.S., Kim, J., Ohno-Machado, L.: A predictive model for extended postanesthesia care unit length of stay in outpatient surgeries. Anesth. Analg. 124(5), 1529– 1536 (2017)
- [54] Dyas, A.R., Henderson, W.G., Madsen, H.J., Bronsert, M.R., Colborn, K.L., Lambert-Kerzner, A., McIntyre, R.C. Jr, Meguid, R.A.: Development and validation of a prediction model for conversion of outpatient to inpatient surgery. Surgery 172(1), 249–256 (2022)
- [55] Le Manach, Y., Collins, G., Rodseth, R., Le Bihan-Benjamin, C., Biccard, B., Riou, B., Devereaux, P.J., Landais, P.: Preoperative score to predict postoperative mortality (POSPOM): Derivation and validation. Anesthesiology 124(3), 570–579 (2016)
- [56] Smilowitz, N.R., Berger, J.S.: Perioperative cardiovascular risk assessment and management for noncardiac surgery: A review. JAMA 324(3), 279–290 (2020)
- [57] Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., Bosselut, A.: MEDITRON-70B: Scaling medical pretraining for large language models (2023) arXiv:2311.16079 [cs.CL]
- [58] Toma, A., Lawler, P.R., Ba, J., Krishnan, R.G., Rubin, B.B., Wang, B.: Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding (2023) arXiv:2305.12031 [cs.CL]
- [59] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-Consistency improves chain of thought reasoning in language models (2022) arXiv:2203.11171 [cs.CL]
- [60] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented generation for Knowledge-Intensive NLP tasks (2020) arXiv:2005.11401 [cs.CL]