TEACHING CONSENSUS RULES (AND EXCEPTIONS) TO LLMS FOR TRUSTWORTHY MEDICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning for early prediction in medicine has recently shown breakthrough performance, however, the focus on improving prediction accuracy has led to a neglect of faithful explanations that are required to gain the trust of medical practitioners. The goal of this paper is to teach LLMs to follow medical consensus guidelines step-by-step in their reasoning and prediction process. Since consensus guidelines are ubiquitous in medicine, instantiations of verbalized medical inference rules to electronic health records provide data for fine-tuning LLMs to learn consensus rules and possible exceptions thereof for many medical areas. Consensus rules also enable an automatic evaluation of the model's inference process regarding its derivation correctness (evaluating correct and faithful deduction of a conclusion from given premises) and value correctness (comparing predicted values against real-world measurements). We exemplify our work using the complex Sepsis-3 consensus definition. Our experiments show that small fine-tuned models outperform one-shot learning of considerably larger LLMs that are prompted with the explicit definition and models that are trained on medical texts including consensus definitions. Since fine-tuning on verbalized rule instantiations of a specific medical area yields nearly perfect derivation correctness for rules (and exceptions) on unseen patient data in that area, the bottleneck for early prediction is not out-of-distribution generalization, but the orthogonal problem of generalization into the future by forecasting sparsely and irregularly sampled clinical variables. We show that the latter results can be improved by integrating the output representations of a time series forecasting model with the LLM in a multimodal setup.

1 INTRODUCTION

Medical consensus definitions are guidelines stated by a representative group of experts on how to diagnose and treat a disease based on clinical evidence. From the perspective of logical inference, consensus guidelines include deductive and inductive inference rules. Deductive rules have the form of if-then relations where the conclusion is certain given the correctness of the premise, for example, in mapping thresholds on clinical measurements to step functions of diseases. If used for early prediction purposes (a.k.a. prognosis), inductive inference rules are required where the if-then relation between premise and conclusion is probabilistic. An illustrative example for a complex consensus guideline is the Sepsis-3 definition that identifies an organ dysfunction as an acute change in total SOFA score ≥ 2 points consequent to an infection (Singer et al., 2016; Seymour et al., 2016). The SOFA (Sepsis-Related Organ Failure Assessment) score itself constitutes a consensus definition that is based on definitions for six organ systems, each defining thresholds on particular clinical variables observed during a 24h window ((Vincent et al., 1996), see Table 5 in Appendix A.1). These calculations comprise a logical rule system where time series forecasting (TSF) of clinical variables represents inductive rules, which are composed with deductive rules that calculate extrema over time, map clinical measurements onto step functions, and calculate changes over time (see Figure 1).

The goal of this work is to teach LLMs to follow the deductive and inductive rules of medical consensus guidelines step-by-step in their prediction and reasoning process[1], in order to foster the trust of medical practitioners in the generated diagnosis or prognosis. We exemplify our work

[1]In the following, in order to avoid an anthropomorphic attribution of "reasoning" capabilities to LLMs, we will speak of medical "rules" that are taught, and of an "inference process" that is being evaluated for the LLM.

using the complex Sepsis-3 consensus definition. Instantiations of verbalized medical inference rules to clinical patient data allow the model to learn the consensus rule and possible exceptions to the rule from patient examples in many medical areas (see Figure 2). Our work transfers learning of compositional inference tasks from mathematical problems to medical inference, with several advantages.

First, consensus definitions are ubiquitous in medicine, ranging from guidelines for mental disorders (American Psychiatric Association, 2013) to neurological (McDonald et al., 2001) and physiological diseases (KDIGO Acute Kidney Injury Work Group, 2012). For each area, verbalization of consensus rules can be done automatically by using templates (which can themselves be generated automatically by using LLMs) that describe each step of an application of a consensus rule system to patient data. Furthermore, human curation can be integrated by annotating verbalized inference chains with possible exceptions and corrections to the rule. In contrast to prompting LLMs, supervised fine-tuning allows learning of rules and exceptions from example instantiations to patient data.

Second, the use of consensus rules enables an exact and automatic evaluation of the model's inference process against a trusted medical gold standard. We present an evaluation setup that differentiates the derivation correctness of a trained model — measuring a model's ability to learn to correctly deduce a conclusion from a given premise — from a model's value correctness — comparing the numerical value predicted in each inference step against real-world measurements. An evaluation of derivation correctness also can be seen to measure the faithfulness of the model's inference process in the sense of checking for an accurate representation of the reasoning process behind the model's prediction.

Our results show that small fine-tuned models (LLaMA 8B parameters) outperform one-shot learning of considerably larger LLMs (LLaMA 70B parameters) that are given the explicit definition in the prompt, and LLMs that are trained on medical texts including the original consensus definitions (Me-LLaMA 8B parameters) under all evaluation metrics. In particular, fine-tuned LLMs show nearly perfect generalization to unseen patient data with respect to derivation correctness of rules and exceptions. This demonstrates that an adaptation of an LLM to verbalized rule instantiations for a specific medical area likely guarantees consistent results in that area, whereas training or prompting models with the abstract definition texts is not sufficient. We conjecture that generalization across consensus definitions in areas as different as psychiatry (based on interviews (American Psychiatric Association, 2013)), neurology (based partially on magnetic resonance imaging (McDonald et al., 2001)), or physiological diseases (based on vital signals and lab tests (KDIGO Acute Kidney Injury Work Group, 2012; Cederholm et al., 2019; Singer et al., 2016; Seymour et al., 2016)) is difficult to bridge without supervision. The bottleneck is thus not out-of-distribution generalization (Chu et al., 2025), but the orthogonal problem of generalization into the future, i.e., the inherent complexity of time series forecasting of sparsely and irregularly sampled clinical variables. The latter results can be improved by moving from a text-based encoding of clinical measurements to a multi-modal approach where the output representations of a TSF model are fed as additional input to the LLM.

2 Related Work

Reasoning in LLMs Recent evaluations of the "reasoning" process of LLMs have shown that they are able to correctly solve sub-tasks of multi-step inference tasks, but fail to compose them into a fully correct inference path, especially in extrapolation to more complex out-of-distribution data (Dziri et al., 2023; Zhang et al., 2023; Saparov & He, 2023; Yang et al., 2024; Mondorf & Plank, 2024). For compositional inference problems in mathematics or logics, fine-tuning on so-called "scratchpads" has been shown to yield consistently strong generalization (Nye et al., 2021; Hochlehnert et al., 2025). To our knowledge, our work is the first one to apply an automatic generation of scratchpads for fine-tuning and step-by-step evaluation to the inference process of medical LLMs. Recent works have addressed the aspect of faithfulness of chain-of-thought explanations of LLMs, asking whether they accurately represent the underlying inference process of the model (Jacovi & Goldberg, 2020). Faithfulness in chain-of-thought reasoning has been enforced by incorporating deterministic solvers into the inference process (Lyu et al., 2023; Xu et al., 2024). Evaluation of faithfulness has been done manually for LLMs with explicit (Turpin et al., 2023) and implicit biases (Arcuschin et al., 2025), or by defining automatic evaluation metrics of chain-of-thought inference processes (Lanham et al., 2023; Wang et al., 2025; Chen et al., 2025). The latter metrics are based on accuracy or confidence of

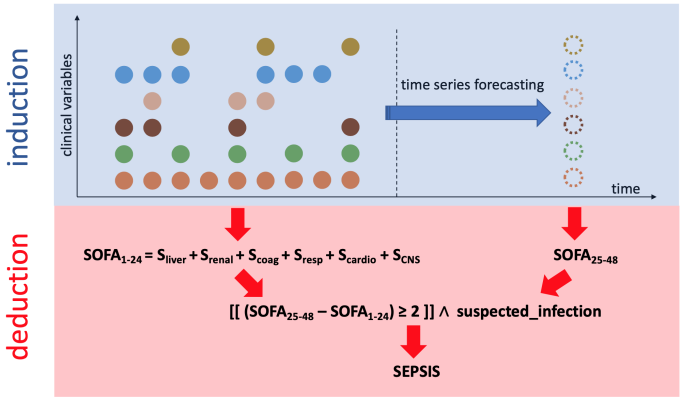

Figure 1: Inductive and deductive inference rules in the Sepsis-3 Consensus Definition (Singer et al., 2016; Seymour et al., 2016). Deductive rules calculate extrema over time, map thresholds onto step functions for SOFA scores, and calculate total SOFA and changes over time. Inductive rules involve time series forecasting of clinical variables 24 hours into the future.

the generated LLM outputs, and can thus be considered "reference-free" evaluation metrics, whereas a consensus rule can be used as a unique gold standard reference for evaluation.

Medical Reasoning Medical reasoning has been characterized to include two contrasting mechanisms, termed analytic versus non-analytic strategies (Eva, 2005), causal versus example-based knowledge (Norman, 2005), or contrasting external clinical evidence with individual clinical expertise (Sackett et al., 1996). To our knowledge, no attempt has been made so far to differentiate these aspects in teaching and evaluating LLMs for healthcare. Instead, most evaluations are based on medical question-answering (Liu et al., 2023; Singhal et al., 2025), while it has been shown that the vast majority of questions in medical QA benchmarks can be answered by factual recall, not requiring multi-step inference at all (Thapa et al., 2025). While excelling in medical question-answering tasks, few-shot LLMs suffer from prompt brittleness (Hochlehnert et al., 2025; Reese et al., 2024) and fail in basic TSF tasks (Merrill et al., 2024; Tan et al., 2024) and lack adherence to consensus guidelines (Hager et al., 2024). This discrepancy calls for evaluations of medical LLMs that go beyond the final result and instead evaluate the inference process itself. Various approaches have been presented to teach LLMs medical inference rules, e.g., by pre-training or fine-tuning on medical texts including consensus guidelines (Chen et al., 2023; Xie et al., 2025), or by incorporating rule-based knowledge in knowledge graphs (Wu et al., 2025), reward models (Yun et al., 2025; Chen et al., 2024; Zhang et al., 2025), or by manual annotation by clinical experts (Liu et al., 2025). However, with the exception of the last work which relies on manual evaluation, all works are evaluated on standard medical question-answering benchmarks. Our work uses trusted consensus rules as training data and, instantiated to different patients, as gold standard references for evaluation, and obviates the need for manual evaluation.

Early Prediction of Sepsis Machine learning for early prediction of sepsis starts by labeling clinical data[2], by applying a Sepsis consensus definition [3] to future clinical observations that are unseen to the model, and then training models to predict the label based on clinical measurements observed several hours before. Examples are the machine learning approaches taken by the 104 participants in the PhysioNet Challenge on "Early Prediction of Sepsis from Clinical Data" (Reyna et al., 2019), or the 21 approaches described in a recent overview (Moor et al., 2021). Machine learning for early prediction of sepsis includes a broad variety of learning algorithms, from linear learners to deep learning (see Reyna et al. (2019) and Moor et al. (2021)), up to most recent works using Transformers (Choi et al., 2024), graph neural networks (Yin et al., 2025), and pretrained LLMs (Li et al., 2024). The setup of predicting a label that is the outcome of a consensus definition can be abstractly viewed as the prediction of an effect that is caused by applying a consensus definition to future values of

[2]E.g., MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), or eICU (Pollard et al., 2018).
[3]Sepsis-3 or Sepsis-2 (Levy et al., 2003; Dellinger et al., 2013) in earlier works.

clinical variables. Another option is to directly predict the causes by forecasting clinical variables, and determine the effect by applying the consensus definition to the forecasted values (Staniek et al., 2024). Our approach is most similar to the latter, with the important difference that our approach not only learns rules, but also exceptions thereof from patient data.

3 Teaching LLMs to Diagnose According to the Sepsis-3 Definition

Deductive Inference Sepsis-3 contains several deductive inference steps: A first step is a calculation of "worst" values (minima or maxima) of clinical variables over 24 hours. The crucial deductive task in Sepsis-3 can be seen as a deductive syllogism consisting of rules (the major premises) that are instantiated to thresholds on clinical variables (the minor premises) that are mapped to step functions for six SOFA subscores, yielding scores from $0-4$ (the conclusions). Each SOFA subscore corresponds to an organ system, namely the central nervous system, and the cardiovascular, respiratory, coagulation, liver, and renal organ systems. The rules themselves are abbreviated as S_{CNS}, S_{cardio}, S_{resp}, S_{coag}, S_{liver}, and S_{renal} in the lower part of Figure 1 and shown in the columns in Table 5 in Appendix A.1. The total SOFA score is calculated by summing these six subscores to a score ranging from $0-24$, once for clinical measurements during the first 24 hours (SOFA_{1-24}), and for predicted measurements in the next 24 hours (SOFA_{25-48}). A further deductive inference step is the computation of a change ≥ 2 as

$$\text{SOFA}_{\text{diff}} := [\![(\text{SOFA}_{25:48} - \text{SOFA}_{1:24}) \geq 2]\!], \tag{1}$$

where $[\![a]\!] = 1$ if a is true, 0 otherwise. Finally, a binary Sepsis label is assessed by combining the indicator function in Equation 1 with a binary indicator suspected_infection[4], yielding

$$\text{SEPSIS} := \text{SOFA}_{\text{diff}} \wedge \text{suspected\_infection}. \tag{2}$$

Inductive Inference Sepsis-3 is laid out to detect a life-threatening organ dysfunction by an acute change in the total SOFA score. Measuring an increase in SOFA score over time requires TSF of future values of a set of clinical features F (see upper part of Figure 1), and an application of the SOFA definition to current and future clinical values. Formally, given a representation x of an input time series, an output vector of predicted clinical values $\hat{y}_t \in \mathbb{R}^{|F|}$ is produced. The features used in our experiments are 131 clinical measurements extracted from the MIMIC-III database (Johnson et al., 2016). A full list is given in Appendix A.4. The TSF task in our experiments is defined to predict the next 24 hours from a history of preceding 24 hours. Dedicated models to perform the TSF task are described in Appendix A.2.

Teaching Verbalized Consensus Rules to Autoregressive LLMs An approach to teaching LLMs to generate a chain of inference rules that adheres to a certain definition is to deploy an autoregressive architecture where next-token prediction is based on a history of previously predicted tokens, and token-wise errors on a target inference chain can be backpropagated through the system. This procedure coincides with standard supervised fine-tuning, where an input question, a gold-standard verbalization of the inference process, and the answer, is provided to the model during training. An example for fine-tuning data including inference rules following the Sepsis-3 definition is shown in Figure 2, left column. The data starts with a verbalization of a sparse multivariate input time series of measurements for a 24 hour window, followed by an instruction to classify the patient as septic in the next 24 hours, given information about suspected infection[5]. The third text block in the fine-tuning data consists of a gold standard inference chain explaining the scores of the six SOFA systems from the corresponding clinical measurements, and computing their sum, given the measurements of the first 24 hour window. The fourth text block includes forecasts of the clinical variables relevant for SOFA for the next 24 hours, and the corresponding inference about the six SOFA subscores, and their sum. The last text block contains a gold standard inference about the Sepsis prediction and the responsible SOFA system, if applicable.

[4]Following Singer et al. (2016); Seymour et al. (2016), a suspected infection is defined as a combination of antibiotics treatment and blood cultures, starting within the first 24 hours after admission.

[5]The onset of suspicion of infection is thus given prior to the onset of salient organ dysfunction in our approach, corresponding to labeling scheme H1 of Cohen et al. (2024), and consistent with other work on early prediction of sepsis where suspected infection is included in the list of input features (Nemati et al., 2018).

Patient is 75.0 years old and is male. Given all the information in this text, answer the question at the end. Here are the measurements: DBP at time -22.37: 49.0, SBP at time -22.37: 105.0, DBP at time -20.37: 52.0, GCS_eye at time -20.37: 4.0, GCS_motor at time -20.37: 6.0, GCS_verbal at time -20.37: 1.0, SBP at time -20.37: 117.0, DBP at time -19.37: 56.0, FiO2 at time -19.37: 0.5, SBP at time -19.37: 127.0, DBP at time -18.37: 43.0,...

Now answer the following question: The doctors suspect an infection, based on this information and the other information in this text, will the patient be classified as septic tomorrow?

First we need to calculate the SOFA scores given the extracted values. The SOFA scores for the current time are the following: The minimum value of GCS_eye is 4.0, GCS_motor is 6.0 and GCS_verbal is 1.0, this produces the sum 11.0 and means the CNS SOFA is 2. Because minimum MAP is 55.333, max Dopamine is 0, max Dobutamine is 0, max Epinephrine is 0 and max Norepinephrine is 0 with a patient weight of 62.8 kg, the cardiovascular SOFA is 1. Given that minimum PO2 is 100.0 and minimum FiO2 is 0.5 the calculated PAO2FIO2 is 200.0, this means the respiratory SOFA is 2. Because the minimum Platelet count is 310.0 the coagulation SOFA is 0. The maximum Bilirubin (Total) is 1 leading to a liver SOFA of 0. Because total Urine output is 1095.0 and maximum creatinine in the blood is 0.4 the renal SOFA is 0. To summarize: the patient has a total SOFA score of 5.

Now we need to calculate the SOFA scores with forecasted values. The SOFA scores in the future based on the forecasted values are the following: The minimum value of GCS_eye will be 4.0, GCS_motor will be 6.0 and GCS_verbal will be 1.0, this produces the sum 11.0 and means the CNS SOFA will be 2. Because future minimum MAP will be 65.333, future max Dopamine will be 0, future max Dobutamine will be 0, future max Epinephrine will be 0 and future max Norepinephrine will be 0 with a patient weight of 62.8 kg, the cardiovascular SOFA will be 1. Given that minimum PO2 will be 100.0 and minimum FiO2 will be 0.5 the forecasted PAO2FIO2 will be 200.0, this means the respiratory SOFA will be 2. Because the Platelet count will be 310.0 the coagulation SOFA is going to be 0. The maximum Bilirubin (Total) will be 1 leading to a liver SOFA of 0. Because Urine output will be 150.0 and maximum creatinine in the blood will be 0.4 the renal SOFA will be 4. **To summarize: the patient will have a future total SOFA score of 9.**

This calculation means that the patient will likely experience a kidney failure since SOFA increased by 4. The patient will develop sepsis in the next 24 hours, because total SOFA increased by 4 and infection is suspected.

Patient is 75.0 years old and is male. Given all the information in this text, answer the question at the end. Here are the measurements: DBP at time -22.37: 49.0, SBP at time -22.37: 105.0, DBP at time -20.37: 52.0, GCS_eye at time -20.37: 4.0, GCS_motor at time -20.37: 6.0, GCS_verbal at time -20.37: 1.0, SBP at time -20.37: 117.0, DBP at time -19.37: 56.0, FiO2 at time -19.37: 0.5, SBP at time -19.37: 127.0, DBP at time -18.37: 43.0,...

Now answer the following question: **The patient has an existing precondition given by the ICD-10 code N18.9.** The doctors suspect an infection, based on this information and the other information in this text, will the patient be classified as septic tomorrow?

First we need to calculate the SOFA scores given the extracted values. The SOFA scores for the current time are the following: The minimum value of GCS_eye is 4.0, GCS_motor is 6.0 and GCS_verbal is 1.0, this produces the sum 11.0 and means the CNS SOFA is 2. Because minimum MAP is 55.333, max Dopamine is 0, max Dobutamine is 0, max Epinephrine is 0 and max Norepinephrine is 0 with a patient weight of 62.8 kg, the cardiovascular SOFA is 1. Given that minimum PO2 is 100.0 and minimum FiO2 is 0.5 the calculated PAO2FIO2 is 200.0, this means the respiratory SOFA is 2. Because the minimum Platelet count is 310.0 the coagulation SOFA is 0. The maximum Bilirubin (Total) is 1 leading to a liver SOFA of 0. Because total Urine output is 1095.0 and maximum creatinine in the blood is 0.4 the renal SOFA is 0. To summarize: the patient has a total SOFA score of 5.

Now we need to calculate the SOFA scores with forecasted values. The SOFA scores in the future based on the forecasted values are the following: The minimum value of GCS_eye will be 4.0, GCS_motor will be 6.0 and GCS_verbal will be 1.0, this produces the sum 11.0 and means the CNS SOFA will be 2. Because future minimum MAP will be 65.333, future max Dopamine will be 0, future max Dobutamine will be 0, future max Epinephrine will be 0 and future max Norepinephrine will be 0 with a patient weight of 62.8 kg, the cardiovascular SOFA will be 1. Given that minimum PO2 will be 100.0 and minimum FiO2 will be 0.5 the forecasted PAO2FIO2 will be 200.0, this means the respiratory SOFA will be 2. Because the Platelet count will be 310.0 the coagulation SOFA is going to be 0. The maximum Bilirubin (Total) will be 1 leading to a liver SOFA of 0. Because Urine output will be 150.0 and maximum creatinine in the blood will be 0.4 the renal SOFA will be 4. **To summarize: the patient will have a future total SOFA score of 5**.

The patient will not develop sepsis in the next 24 hours, because total SOFA increased by 0 and infection is suspected.

Figure 2: Fine-tuning data including a verbalization of Sepsis-3 inference (left column) and inference under an exception due to medical preconditions (right column). Differences are shown in bold blue font. The general prompt is shown above the horizontal line, the gold standard answer below.

Handling Exceptions to the Rules Since consensus rules cannot capture every individual disease progression, it is important to consider the possibility of manual corrections to consensus-based inference by adding exceptions to the inference rules. A possible example are medical preconditions with specific treatments, for example, dialysis in case of chronic kidney disease. While there is clear evidence that incorporating pre-existing medical conditions (comorbidities / chronic organ failure) can improve sepsis prediction (Sarraf et al., 2024; Christensen et al., 2023), there is no consensus on how certain preconditions which will alter clinical measurements and thus the SOFA score. In case of a chronic kidney disease, a conservative exception might be to decide to disregard the kidney SOFA in the calculation of total SOFA, leading to an exception to the Sepsis-3 definition. We test this hypothetical scenario by synthetic data. We synthesized fine-tuning data including exceptions to the rule by adding preconditions for five organ systems in form of ICD-10 codes. During data generation, each datapoint was randomly assigned a precondition, either one of the predefined types or the no-precondition option, with equal probability. An example for a verbalization is shown in Figure 2, right column. The format is the same as for inference according to the definition, except for the inclusion of a precondition in the second text block — in our example, the ICD code N18.9 indicating chronic kidney disease — implying an exception to the rule to disregard the SOFA score of the corresponding organ system in the fourth and fifth text blocks. We prepared the data such

Table 1: Derivation correctness of predicted values. Forced derivation correctness in brackets.

	variable	one-shot	one-shot-70B	me-llama	deepseek	fine-tuned	pipeline	multimodal
current	S_{CNS}	0.693	0.803	0.520	0.288	1.000	1.000	1.000
	S_{cardio}	0.516	0.696	0.258	0.338	0.998	0.998	0.975
	S_{resp}	0.561	0.302	0.404	0.279	0.995	0.996	0.996
	S_{coag}	0.599	0.763	0.511	0.381	0.966	0.966	0.970
	S_{liver}	0.537	0.864	0.601	0.336	0.993	0.992	0.992
	S_{renal}	0.537	0.730	0.396	0.256	1.000	0.998	1.000
	$SOFA_{1:24}$	0.543	0.893	0.777	0.165	1.000	1.000	0.999
future	S_{CNS}	0.504 (0.827)	0.545 (0.895)	0.399 (0.786)	0.261 (0.823)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	S_{cardio}	0.580 (0.694)	0.679 (0.714)	0.353 (0.647)	0.269 (0.671)	0.987 (0.997)	0.975 (0.997)	0.965 (0.997)
	S_{resp}	0.530 (0.756)	0.307 (0.827)	0.384 (0.733)	0.266 (0.722)	0.995 (0.996)	0.996 (0.996)	0.995 (0.996)
	S_{coag}	0.661 (0.786)	0.754 (0.818)	0.519 (0.801)	0.306 (0.747)	0.967 (1.000)	0.970 (0.999)	0.966 (1.000)
	S_{liver}	0.534 (0.835)	0.847 (0.969)	0.599 (0.657)	0.335 (0.776)	0.992 (0.999)	0.992 (0.999)	0.992 (1.000)
	S_{renal}	0.554 (0.708)	0.737 (0.773)	0.414 (0.690)	0.251 (0.634)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	$SOFA_{25:48}$	0.518 (0.875)	0.868 (0.984)	0.768 (0.992)	0.148 (0.550)	1.000 (1.000)	0.999 (1.000)	1.000 (1.000)
	$SOFA_{diff}$	0.845 (0.740)	0.891 (0.477)	0.769 (0.414)	0.503 (0.351)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	SEPSIS	0.859 (0.682)	0.909 (0.727)	0.769 (0.688)	0.738 (0.521)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)

that some preconditions are seen during training and testing, while we also chose some codes from the same ICD class as out-of-distribution data that were only seen at test time. The codes and their distribution are shown in Tables 9 and 10 in Appendix A.6.

4 Experimental Setup

Data and Models In our experiments, we use electronic health records (EHRs) from the MIMIC-III data (Johnson et al., 2016). After filtering for patients with an ICU stay of at least 24 hours with reported gender and age of at least 18 years, our dataset contained 44,858 ICU stays with 56 million data points. We split the data into partitions for fine-tuning (28,708), development (7,270), and testing (8,880). For computational reasons, we further subsampled 15,000 datapoints for fine-tuning, and 3,000 datapoints for development and testing, respectively. The resulting percentage of positive Sepsis cases in the test set was 7.33%. As features, we considered 131 clinical variables and the demographic variables gender and age (see Appendix A.4). This selection comprises all vital signs and laboratory values used in the PhysioNet challenge for early prediction of sepsis (Reyna et al., 2019), together with information on suspected infection as in Nemati et al. (2018). The time series of clinical observations in the data were split into a 24 hour observation window, followed by a 24 hour prediction window. During training and testing, we use a sliding window of 24 hours so that full admission days are given as input and output. Using the consensus definition given by Table 5 and Equation 2, SOFA scores and Sepsis label were calculated deterministically for the given data. Following this, features and SOFA scores were verbalized into texts as seen in Figure 2. In the calculation of the ground truth, missing values are being carried forward from the previous day only.

The basic LLM used in our experiments is a pretrained Llama-3 model with 8B parameters (Grattafiori et al., 2024) [6] used in **one-shot** mode. All one-shot experiments use the prompt shown in Appendix A.5 that includes the explicit Sepsis-3 rules together with an example instantiation to patient data. Our **fine-tuned** model uses LORA adapters (Hu et al., 2022) on verbalizations of inference processes. Furthermore, we use a dedicated medical TSF forecaster, pre-trained on MIMIC-III to improve the inductive inference part of the LLM (see Appendix A.2). The TSF model is used in two ways: First, we extract predictions from the forecaster and augment the prompt with those predictions in a **pipeline** approach. In this approach, the Llama-3 model is still finetuned using LORA, but the forecaster is kept fixed. Second, we use a **multimodal** setup to connect the forecaster with the Llama-3 model. We extract the full decoder prediction from the forecaster. The output vectors are then fed into a connector MLP to produce vectors with Llama-3 embedding dimensionality. These time series "token embeddings" are then prepended to the actual text embeddings. The LORA adapters in the Llama-3 model, the connector MLP, and the TSF model get updated during the finetuning process. A list of hyperparameter settings chosen on the validation set (except the default setting for the Llama-3 8B model) is given in Appendix A.3. For further comparison, we use an 8B parameters LLM pretrained on medical texts, including publications and wikipedia texts on consensus guidelines (**me-llama**

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Table 2: Derivation correctness for in- and out-of-distribution exceptions to inference rules.

variable	% changes ID	% changes OOD	ID score	OOD score
$\text{SOFA}_{1:24}$	0.404	0.424	1.000	1.000
$\text{SOFA}_{25:48}$	0.433	0.452	1.000	1.000

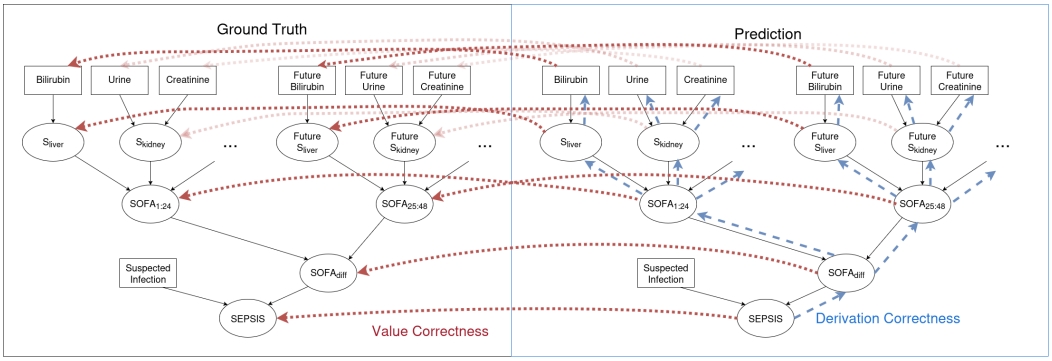

Figure 3: **Derivation correctness** (dashed blue arrows) checks for the predicted inference graph whether each child node (conclusion) follows from the parent node (premise) according to the consensus rule. **Value correctness** (dotted red arrows) maps each node in the predicted inference graph to its corresponding node in the ground truth graph, consisting of real-world clinical measurements for first and second 24 hours, and deterministic calculation of SOFA and SEPSIS on these values.

(Xie et al., 2025)[7]), and a reasoning-style LLM (**deepseek** DeepSeek-AI et al. (2025)[8]), and 70B pretrained Llama-3 model (**one-shot-70B**[9]), all in one-shot mode.

Evaluation Metrics We evaluate the generated output of LLMs in three different ways. First, we evaluate the LLM output according to **derivation correctness** (dashed blue arrows in Figure 3) which checks whether each child node (conclusion) in the predicted inference graph follows from the parent node (premise) according to the consensus rule. This evaluation measures the ability of the LLM to learn the deductive inference rules of the consensus definition. Furthermore, it measures faithfulness of the model's inference process in the sense of checking for an accurate representation of the reasoning process behind the model's prediction. In order to avoid trivial cases of derivation correctness, for example, inference of unchanged SOFA scores from unchanged future measurement predictions, we measure in addition **forced derivation correctness**. This is done by forcing the decoder to use the correct inference chain up to a last measurement token of a SOFA system as context, then continuing the decoding process until the specified target node. Furthermore, we conduct an evaluation according to **value correctness** (dotted red arrows in Figure 3) which compares the numerical value of each partial inference node in the prediction graph with a corresponding real-world output. Since errors earlier in the inference necessarily propagate through the inference chain, this evaluation measures the utility of the LLM output in practical applications. Comparison in all three cases is done by parsing the LLM output for relevant keywords, extracting numbers, and checking if the predicted number lies in a specific interval around the true value. A match is considered positive if the result deviates from the true value by 5% (see Appendix A.7 for a discussion). In case of value correctness, we calculate a contingency table by comparing the application of the Sepsis-3 definition to a model involving forecasted clinical measurements with Sepsis-3 calculated for real-world clinical measurements in the second 24 hours block. From this, metrics such as accuracy, specificity, sensitivity and F1 are computed.

[7] https://huggingface.co/YBXL/Med-LLaMA3-8B
[8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
[9] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

Table 3: Value correctness of predicted clinical variables compared to real-world measurements. All fine-tuning improvements over one-shot learning are statistically significant according to an approximate randomization test.

	variable	one-shot	one-shot-70B	me-llama	deepseek	fine-tuned	pipeline	multimodal
current	GCS-eye	0.855	0.982	0.764	0.608	0.998	0.997	0.998
	GCS-motor	0.914	0.980	0.733	0.616	0.998	0.998	0.998
	GCS-verbal	0.947	0.985	0.750	0.635	0.999	0.998	0.999
	MAP	0.036	0.219	0.046	0.089	0.993	0.976	0.994
	Dopamine	0.875	0.952	0.754	0.078	0.972	0.964	0.968
	Dobutamine	0.905	0.978	0.785	0.083	0.993	0.991	0.992
	Epinephrine	0.891	0.965	0.773	0.082	0.979	0.980	0.980
	Norepinephrine	0.809	0.875	0.692	0.076	0.909	0.895	0.906
	Weight	0.851	0.765	0.722	0.280	0.999	0.999	0.999
	PaO2/FiO2	0.140	0.559	0.090	0.059	0.998	0.997	0.997
	PaO2	0.533	0.563	0.352	0.288	0.998	0.997	0.997
	FiO2	0.881	0.711	0.721	0.479	1.000	1.000	1.000
	Platelet	0.922	0.966	0.676	0.664	0.997	0.996	0.996
	Bilirubin	0.299	0.275	0.262	0.177	0.999	0.998	0.999
	Urine	0.102	0.262	0.095	0.037	0.966	0.761	0.880
	Creatinine	0.887	0.970	0.666	0.408	0.998	0.998	0.998
future	GCS-eye	0.679	0.710	0.525	0.357	0.712	0.743	0.789
	GCS-motor	0.783	0.805	0.620	0.398	0.811	0.832	0.869
	GCS-verbal	0.806	0.840	0.648	0.434	0.848	0.871	0.895
	MAP	0.110	0.197	0.125	0.064	0.274	0.289	0.294
	Dopamine	0.917	0.966	0.766	0.386	0.978	0.987	0.988
	Dobutamine	0.937	0.982	0.787	0.388	0.994	0.993	0.993
	Epinephrine	0.933	0.972	0.779	0.381	0.989	0.989	0.989
	Norepinephrine	0.861	0.887	0.710	0.350	0.913	0.914	0.913
	Weight	0.851	0.765	0.722	0.280	0.915	0.916	0.914
	PaO2/FiO2	0.073	0.156	0.053	0.038	0.565	0.584	0.596
	PaO2	0.181	0.179	0.148	0.098	0.616	0.659	0.715
	FiO2	0.738	0.538	0.613	0.379	0.821	0.840	0.867
	Platelet	0.283	0.300	0.236	0.166	0.364	0.383	0.396
	Bilirubin	0.180	0.166	0.169	0.115	0.811	0.832	0.866
	Urine	0.064	0.077	0.069	0.023	0.133	0.152	0.183
	Creatinine	0.360	0.347	0.295	0.161	0.395	0.424	0.455

5 Results and Discussion

LLMs Shine in Deductive Inference According to Consensus Rules Our first research question asks whether LLMs can learn the deductive inference rules underlying medical consensus definitions by fine-tuning on verbalized instantiations of the inference process to patient data. Table 1 shows that all fine-tuned models, from basic fine-tuning with text encoding to pipeline and multimodal approaches, achieve nearly perfect derivation correctness. This includes mapping clinical variables to SOFA step functions for the first 24 hours (rows 1-7), mapping forecasted clinical variables to future SOFA scores (rows 8-15), and computing a Sepsis label (row 16). In contrast, one-shot learning does not perform nearly as good, even for models that include abstract consensus definitions in their training data (me-llama) or prompt and are significantly larger (one-shot-70B). Forced derivation correctness (in brackets in Table 1) shows that predictions using correct histories are very similar to using predicted histories. Again, all one-shot learners underperform compared to fine-tuning.

LLMs Can Learn Exceptions to Consensus Rules The next question we ask is whether LLMs can learn exceptions to the inference rules that might conflict with the consensus definition. In our hypothetical scenario, we synthesize examples where ICD codes indicate a medical precondition related to an organ system, with the exception to the rule implying that the SOFA score for the respective organ system should be disregarded in the calculation of total SOFA and Sepsis. Table 2 shows that the addition of preconditions to the premises caused changes in 40-45% of the cases, however, the derivation correctness of total SOFA scores was 100% without losing performance on examples without preconditions. This shows that fine-tuning can indeed handle exceptions to the rule, opening the doors to possible extensions of our work to include human feedback in the loop.

Table 4: Value correctness of predicted SOFA and SEPSIS scores compared to derivations from real-world clinical measurements. All fine-tuning improvements over one-shot learning are significant.

	variable	one-shot	one-shot-70B	me-llama	deepseek	fine-tuned	pipeline	multimodal
current	S_{CNS}	0.633	0.794	0.478	0.239	0.998	0.997	0.997
	S_{cardio}	0.405	0.6	0.249	0.273	0.985	0.971	0.987
	S_{resp}	0.223	0.411	0.172	0.128	0.996	0.994	0.996
	S_{coag}	0.614	0.761	0.497	0.428	0.999	0.999	0.999
	S_{liver}	0.492	0.809	0.502	0.208	1.000	0.999	1.000
	S_{renal}	0.540	0.735	0.412	0.303	0.997	0.992	0.996
	$SOFA_{1:24}$	0.111	0.236	0.076	0.044	0.979	0.958	0.981
future	S_{CNS}	0.532	0.625	0.408	0.231	0.701	0.732	0.760
	S_{cardio}	0.426	0.527	0.295	0.195	0.700	0.738	0.745
	S_{resp}	0.201	0.348	0.163	0.082	0.746	0.754	0.780
	S_{coag}	0.618	0.693	0.483	0.278	0.839	0.823	0.834
	S_{liver}	0.492	0.809	0.502	0.208	0.945	0.945	0.945
	S_{renal}	0.482	0.620	0.365	0.211	0.553	0.451	0.473
	$SOFA_{25:48}$	0.117	0.176	0.088	0.059	0.203	0.220	0.243
	$SOFA_{diff}$	0.139	0.167	0.128	0.127	0.151	0.147	0.163
	SEPSIS Accuracy	0.834	0.857	0.763	0.692	0.868	0.873	0.886
	SEPSIS Specificity	0.876	0.915	0.788	0.739	0.936	0.920	0.922
	SEPSIS Sensitivity	0.3318	0.118	0.455	0.091	0.263	0.336	0.386
	SEPSIS F1	0.231	0.108	0.220	0.042	0.254	0.272	0.309

The Bottleneck Lies in Inductive Inference, not Out-of-Distribution Generalization Our final evaluation compares the results of partial inference steps to real-world clinical measurements and to SOFA and Sepsis scores derived from these. As shown in the first half of Table 3, the numerical values that the LLM decoder has to extract from time series encodings are nearly perfect for fine-tuning approaches. The second half of Table 3 reveals the bottleneck of teaching medical inference rules to LLMs: Correctness of predicting values of 131 clinical variables 24 hours into the future drops from the high nineties to values in the tens and twenties, especially for sparsely and irregularly observed lab values. The same is true for value correctness of predicted SOFA scores and Sepsis labels: As shown in Table 4, the numerical values produced by mapping clinical measurements in the first 24 hours to SOFA scores are nearly perfect, however, mapping values of forecasted clinical variables to step functions severely impacts correctness, especially on metrics like F1 that are sensitive to imbalanced data. One-shot learning is impacted even more, while fine-tuning results can be improved by moving from a text-based encoding to coupling dedicated TSF encoders with the LLM.

Discussion Leveraging consensus rules that exist for many medical areas, consistent diagnoses for unseen patients and faithful medical inference even including exceptions to the rules can be achieved by fine-tuning LLMs on instantiations of the respective consensus rule system to patient data. In contrast, training or prompting with abstract rules or on very large general-domain data is insufficient. Since an adaptation of an LLM to consensus rules for a specific medical area yields consistent results and faithful medical inference in that area, the bottleneck in consensus-based medical prediction lies in the orthogonal problem of generalization into the future, i.e., in the inherent complexity of forecasting of sparsely and irregularly sampled clinical variables for early prediction. This problem hinders medical prognosis, and needs to be addressed by improved TSF.

6 Conclusion

LLMs have been shown to excel at various health-related tasks, however, diagnostic AI has not realized the potential to actually reduce the cognitive demand and associated diagnostic errors of clinicians, due to a focus on predictive accuracy of final diagnostic labels (Adler-Milstein et al., 2021). Furthermore, instead of defining diagnostic excellence by perfect labeling based on fully established clinical criteria, Angus & Bindman (2022) advocate to learn patterns of clinical data that are yet unrecognized but still predictive of sepsis. Our approach can be seen as a first step towards these goals by putting a focus on teaching LLMs inference rules that support clinicians, and to learn exceptions to the rules that can capture new clinical patterns. Our experiments show that LLMs perfectly generalize deductive inference to unseen patients, while the bottleneck lies in inductive inference. Future work

shall address the TSF bottleneck of early prediction models, and elevate our work from simulating exceptions to the rule to learning them from real-world clinical data. Furthermore, we will investigate the potential of task association learning (Cai et al., 2025) applied to related consensus definitions.

Reproducibility Statement

This work is based on publicly available data and open-source code. Code to reproduce our experiments is available as supplementary material and will be made public upon acceptance of the paper.

References

Julia Adler-Milstein, Jonathan H. Chen, and Gurpreet Dhaliwal. Next-generation artificial intelligence for diagnosis: From predicting diagnostic labels to "wayfinding". *JAMA*, 326(24):2467–2468, 2021. URL https://doi.org/10.1001/jama.2021.22396.

American Psychiatric Association (ed.). *Diagnostic and Statistical Manual of Mental Disorders*. Fifth edition, 2013. URL https://doi/book/10.1176/appi.books.9780890425787.

Derek C. Angus and Andrew B. Bindman. Achieving diagnostic excellence for sepsis. *JAMA*, 327 (2):117–118, 2022. URL https://doi.org/10.1001/jama.2021.23916.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. In *ICRL Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=L8094Whth0.

Ziyang Cai, Nayoung Lee, Avi Schwarzschild, Samet Oymak, and Dimitris Papailiopoulos. Extrapolation by association: Length generalization transfer in transformers. *arXiv*, abs/2506.09251, 2025. URL https://arxiv.org/abs/2506.09251.

T. Cederholm, G.L. Jensen, M.I.T.D. Correia, M.C. Gonzalez, R. Fukushima, T. Higashiguchi, G. Baptista, R. Barazzoni, R. Blaauw, A. Coats, A. Crivelli, D.C. Evans, L. Gramlich, V. Fuchs-Tarlovsky, H. Keller, L. Llido, A. Malone, K.M. Mogensen, J.E. Morley, M. Muscaritoli, I. Nyulasi, M. Pirlich, V. Pisprasert, M.A.E. de van der Schueren, S. Siltharm, P. Singer, K. Tappenden, N. Velasco, D. Waitzberg, P. Yamwong, J. Yu, A. Van Gossum, C. Compher, Gordon L. Jensen, Compher Charlene, Tommy Cederholm, Andre Van Gossum, Maria Isabel T.D. Correia, M. Cristina Gonzalez, Ryoji Fukushima, Takashi Higashiguchi, G. Baptista, R. Barazzoni, R. Blaauw, A. Coats, A. Crivelli, D.C. Evans, L. Gramlich, V. Fuchs, H. Keller, L. Llido, A. Malone, K.M. Mogensen, J.E. Morley, M. Muscaritoli, I. Nyulasi, M. Pirlich, V. Pisprasert, M.A.E. de van der Schueren, S. Siltharm, P. Singer, K. Tappenden, N. Velasco, D. Waitzberg, P. Yamwong, and J. Yu. GLIM criteria for the diagnosis of malnutrition – a consensus report from the global clinical nutrition community. *Clinical Nutrition*, 38(1):1–9, 2019. URL https://doi.org/10.1016/j.clnu.2018.08.002.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with LLMs. *arXiv*, abs/2412.18925, 2024. URL https://arxiv.org/abs/2412.18925.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv*, abs/2505.05410, 2025. URL https://arxiv.org/abs/2505.05410.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models. *arXiv*, abs/2311.16079, 2023. URL https://arxiv.org/abs/2311.16079.

Jee-Woo Choi, Minuk Yang, Jae-Woo Kim, Yoon Mi Shin, Yong-Goo Shin, and Seung Park. Prognostic prediction of sepsis patient using transformer with skip connected token for tabular data. *Artificial Intelligence in Medicine*, 149:102804, 2024. URL https://doi.org/10.1016/j.artmed.2024.102804.

Erik E. Christensen, Christian H. Prebensen, Anders B. Martinsen, Elisabeth T. Stiff, Rune Hoff, Dag Kvale, and Aleksander R. Holten. Mortality and sequential organ failure assessment score in patients with suspected sepsis: The impact of acute and preexisting organ failures and infection likelihood. *Critical Care Explorations*, 5(2):e0895, 2023. URL https://doi.org/10.1097/cce.0000000000000865.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

Samuel N. Cohen, James Foster, Peter Foster, Hang Lou, Terry Lyons, Sam Morley, James Morrill, Hao Ni, Edward Palmer, Bo Wang, Yue Wu, Lingyi Yang, and Weixin Yang. Subtle variation in sepsis-III definitions markely influences predictive performance within and across methods. *scientific reports*, 14(1920):1–10, 2024. URL https://doi.org/10.1038/s41598-024-51989-6.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, abs/2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

R.P. Dellinger, M.M. Levy, A. Rhodes, D. Annane, H. Gerlach, S.M.Opal, J.E. Sevransky, C.L. Sprung, I.S. Douglas, /R. Jaeschke ant T.M. Osborn, M.E. Nunnally, S.R. Townsend, K. Reinhart, R.M. Kleinpell, D.C. Angus, C.S. Deutschman, F.R. Machado, G.D. Rubenfeld, S.A. Webb, R.J. Beale, J.L. Vincent, and R. Moreno. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Critical Care Medicine*, 41(2):580–637, 2013. URL http://dx.doi.org/10.1097/CCM.0b013e31827e83af.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang

Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=Fkckkr3ya8.

Kevin W. Eva. What every teacher needs to know about clinical reasoning. *Med Educ*, 39(1):98–106, 2005. URL https://doi.org/10.1111/j.1365-2929.2004.01972.x.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban

Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv*, abs/2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30, 2024. URL https://doi.org/10.1038/s41591-024-03097-1.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=90UrTTxp5O.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, virtual, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020. URL https://aclanthology.org/2020.acl-main.386/.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. URL https://doi.org/10.1038/s41597-022-01899-x.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. URL https://doi.org/10.1038/sdata.2016.35.

KDIGO Acute Kidney Injury Work Group. Kidney disease: Improving global outcomes (KDIGO) - Cinical practice guideline for acute kidney injury. *Kidney international supplements*, 2(1):1–138, 2012. URL doi:10.1038/kisup.2012.1.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, abs/2307.13702, 2023. URL https://arxiv.org/abs/2307.13702.

Mitchell M. Levy, Mitchell P. Fink, John C. Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M. Opal, Jean-Louis Vincent, and Graham Ramsay. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Critical Care Medicine*, 31(04):1250–1256, 2003. URL https://doi.org/10.1097/01.ccm.0000050454.01978.3b.

Qiang Li, Hanbo Ma, Dan Song, Yunpeng Bai, Lina Zhao, and Keliang Xie. Early prediction of sepsis using chatgpt-generated summaries and structured data. *Multimedia Tools and Applications*, pp. 1–23, 2024. URL https://doi.org/10.1007/s11042-024-18378-7.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, Song Wu, Ping Zhang, Jian Yang, Chunli Song, and Guangyu Wang. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, 31:932–942, 2025. URL https://doi.org/10.1038/s41591-024-03416-6.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv*, abs/2305.15525, 2023. URL https://doi.org/10.48550/arXiv.2305.15525.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP)*, Nusa Dua, Bali, 2023. URL 10.18653/v1/2023.ijcnlp-main.20.

W. I. McDonald, A. Compston, G. Edan, D. Goodkin, H. P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. van den Noort, B. Y. Weinshenker, and J. S. Wolinsky. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50(1):121–127, 2001. doi: https://doi.org/10.1002/ana.1032.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, 2024. URL 10.18653/v1/2024.findings-emnlp.201.

Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=Lmjgl2n11u.

Michael Moor, Bastian Rieck, Max Horn, Catherine Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine*, 8, 2021. URL https://doi.org/10.3389/fmed.2021.607952.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine*, 46(4):547–553, 2018. URL http://dx.doi.org/10.1097/ccm.0000000000002936.

Geoffrey Norman. Research in clinical reasoning: past history and current trends. *Med Educ*, 39(4): 418–27, 2005. URL https://doi.org/10.1111/j.1365-2929.2005.02127.x.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv*, abs/2112.00114, 2021. URL https://arxiv.org/abs/2112.00114.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(180178), 2018. URL https://doi.org/10.1038/sdata.2018.178.

Justin T. Reese, Daniel Danis, J. Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Christopher J. Mungall, and Peter N. Robinson. On the limitations of large language models in clinical diagnosis. *medRxiv*, 2024. URL https://doi.org/10.1101/2023.07.13.23292613.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P.B. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48 (2):210–217, 2019. URL https://doi.org/10.1097/CCM.0000000000004145.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. URL https://www.bmj.com/content/312/7023/71.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

Elie Sarraf, Alireza Vafaei Sadr, Vida Abedi, and Anthony S. Bonavia. Enhancing sepsis prognosis: Integrating social determinants and demographic variables into a comprehensive model for critically ill patients. *Journal of Critical Care*, 83:154857, 2024. URL https://doi.org/10.1016/j.jcrc.2024.154857.

Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. URL https://doi.org/10.1001/jama.2016.0288.

Mervyn Singer, Clifford S. Deutschman, and Christopher Warren Seymour. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. URL 10.1001/jama.2016.0287.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad

Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 2025. URL https://doi.org/10.1038/s41591-024-03423-7.

Michael Staniek, Marius Fracarolli, Michael Hagmann, and Stefan Riezler. Early prediction of causes (not effects) in healthcare by long-term clinical time series forecasting. *Proceedings of Machine Learning Research: Machine Learning for Healthcare*, 252:1–29, 2024. URL https://arxiv.org/abs/2408.03816.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=DV15UbHCY1.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Disentangling reasoning and knowledge in medical large language models. *arXiv*, abs/2505.11462, 2025. URL https://arxiv.org/abs/2505.11462.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

J.L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. URL https://doi.org/10.1007/BF01709751.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examining the necessity and accuracy of CoT step-by-step. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, 2025. URL https://aclanthology.org/2025.findings-naacl.140/.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv*, 2025. URL https://arxiv.org/abs/2504.00993.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(1), 2025. URL https://doi.org/10.1038/s41746-025-01533-1.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand, 2024. URL 10.18653/v1/2024.acl-long.720.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number cookbook: Number understanding of language models and how to improve it. *arXiv*, abs/2411.03766, 2024. URL https://arxiv.org/abs/2411.03766.

Changchang Yin, Shihan Fu, Bingsheng Yao, Thai-Hoang Pham, Weidan Cao, Dakuo Wang, Jeffrey Caterino, and Ping Zhang. Sepsiscalc: Integrating clinical calculators into early sepsis prediction via dynamic temporal graph construction. *arXiv*, 2025. URL https://arxiv.org/abs/2501.00190.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards, 2025. URL https://arxiv.org/abs/2506.11474.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3365–3373, 2023. URL https://doi.org/10.24963/ijcai.2023/375.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-RLVR: Emerging medical reasoning from a 3B base model via reinforcement learning. *arXiv*, abs/2502.19655, 2025. URL https://arxiv.org/abs/2502.19655.

A APPENDIX

A.1 SOFA SCORE

Table 5: SOFA score for six organ systems, calculated by thresholding corresponding clinical measurements (Vincent et al., 1996). In our setting, we had to recalculate MAP (mean arterial pressure) from SBP and DBP (systolic and diastolic blood pressure), the Horowitz coefficient from PaO2 and FiO2, and had no knowledge about the kind of mechanical ventilation. If no value for calculation in a SOFA subsystem was available, we took a value of 0. Abbreviations: CNS = Central nervous system; GCS = Glasgow Coma Scale; MV = mechanically ventilated including CPAP; MAP = mean arterial pressure, UO = Urine output.

Score	CNS	Cardiovascular	Respiratory	Coagulation	Liver	Renal
	GCS	MAP or vasopressors	PaO2/FiO2 (mmHg)	Platelets ($\times 10^3/\mu$l)	Bilirubin (mg/dl)	Creatinine (mg/dl) or UO
+0	15	MAP $\geq$ 70 mmHg	$\geq$ 400	$\geq$ 150	$<$ 1.2	$<$ 1.2
+1	13-14	MAP $<$ 70 mmHg	$<$ 400	$<$ 150	1.2-1.9	1.2-1.9
+2	10-12	dopamine $\leq$ 5 μg/kg/min OR dobutamine (any dose)	$<$ 300	$<$ 100	2.0-5.9	2.0-3.4
+3	6-9	dopamine $>$ 5 μg/kg/min OR epinephrine $\leq$ 0.1 μg/kg/min OR norepinephrine $\leq$ 0.1 μg/kg/min	$<$ 200 AND MV	$<$ 50	6.0-11.9	3.5-4.9 OR $<$ 500 ml/day
+4	$<$ 6	dopamine $>$ 15 μg/kg/min OR epinephrine $>$ 0.1 μg/kg/min OR norepinephrine $>$ 0.1 μg/kg/min	$<$ 100 AND MV	$<$ 20	$>$ 12.0	$>$ 5.0 OR $<$ 200 ml/day

A.2 Time Series Forecasting

Our TSF model uses the implementation of Staniek et al. (2024) which is based on a Transformer encoder-decoder architecture (Vaswani et al., 2017). First, sparse multivariate input time series are represented as quadruplets $\mathrm{S} = \{(f_i, t_i, v_i, n_i)\}_{i=1}^{n}$, where $f_i \in F$ is a clinical variable identifier, $t_i \in \mathbb{R}_{\geq 0}$ is a time index, $v_i \in \mathbb{R}$ the observed value of f_i at t_i, and n_i the unique stay identifier. Then the quadruplets for a 24 hour time series are encoded into a dense representation x where every timestep is a vector of feature values representing one hour. We construct this vector by choosing the first observed value during the represented hour for each feature. If no value was observed, we impute zero which corresponds to the mean value due to standardization of the data. Additionally, a mask indicating whether a value was imputed is generated and appended to the vector. For TSF, we use a Transformer model with an autoregressive iterative multistep (IMS) decoder that generates an output vector $\hat{y}_t \in \mathbb{R}^{|F|}$. The predicted output $\hat{y}_t$ is a function of the history $\hat{y}_{<t}$ of predicted timesteps until time t, the encoded input x, and the model parameters θ: $\hat{y}_t = f_\theta(\hat{y}_{<t}, x)$. To perform long-term TSF using the autoregressive setup, the outputs $\hat{y}_t$ from each time step $t = 1, \ldots, T$ are concatenated. The complete model is trained with masked mean squared error (MSE).

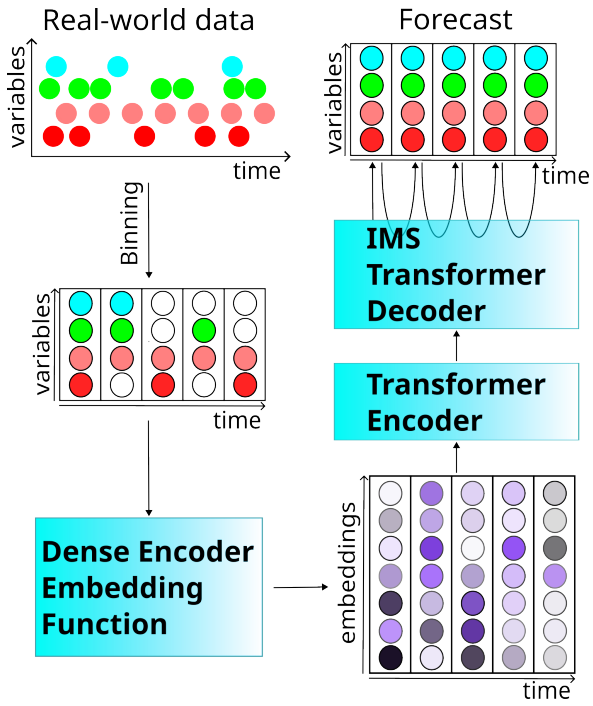

Figure 4: Time Series Forecasting using a dense encoder and iterative multistep decoder architecture.

A.3 HYPERPARAMETERS AND COMPUTE INFRASTRUCTURE

Table 6: Hyperparameter settings of the time series forecaster.

Hyperparameter	value
Embedding Size	512
Hidden size encoder	512
Hidden size decoder	512
# Encoder layers	2
# Decoder layers	1
learning rate	0.0005
attention heads encoder	8
attention heads decoder	1
dropout	0.05
max epochs	100
patience (early stopping)	6
Random Seed	unixtime variation
# GPUs	1
Training all clock time	5 hours
GPU type	Nvidia GTX 1080 Ti

Table 7: Hyperparameter settings of the multi-modal and LLM finetuning architectures.

Hyperparameter	value
optimizer	adam
learning rate	0.000002
max epochs	10
connector first layer	131x2096
connector second layer	2096x4096
lora-r	16
lora-alpha	16
lora-dropout	0.1
lora-target-modules	all-linear
# GPUs	1
Training all clock time	40 hours
GPU type	NVIDIA A100
Starting model	meta-llama/Llama-3.1-8B-Instruct

A.4 CLINICAL FEATURES

Table 8: Feature list for MIMIC-III: Besides the following 131 dynamic variables, only age and gender were extracted. The 15 variables marked with an asterisk are directly used for calculating the SOFA score.

ALP	Epinephrine*	LDH	Packed RBC
ALT	Famotidine	Lactate	Pantoprazole
AST	Fentanyl	Lactated Ringers	Phosphate
Albumin	FiO2*	Levofloxacin	Piggyback
Albumin 25%	Fiber	Lorazepam	Piperacillin
Albumin 5%	Free Water	Lymphocytes	Platelet Count*
Amiodarone	Fresh Frozen Plasma	Lymphocytes (Absolute)	Potassium
Anion Gap	Furosemide	MBP	Pre-admission Intake
BUN	GCS_eye*	MCH	Pre-admission Output
Base Excess	GCS_motor*	MCHC	Propofol
Basophils	GCS_verbal*	MCV	RBC
Bicarbonate	GT Flush	Magnesium	RDW
Bilirubin (Direct)	Gastric	Magnesium Sulfate (Bolus)	RR
Bilirubin (Indirect)	Gastric Meds	Magnesium Sulphate	Residual
Bilirubin (Total)*	Glucose (Blood)	Mechanically ventilated	SBP*
CRR	Glucose (Serum)	Metoprolol	SG Urine
Calcium Free	Glucose (Whole Blood)	Midazolam	Sodium
Calcium Gluconate	HR	Milrinone	Solution
Calcium Total	Half Normal Saline	Monocytes	Sterile Water
Cefazolin	Hct	Morphine Sulfate	Stool
Chest Tube	Heparin	Neosynephrine	TPN
Chloride	Hgb	Neutrophils	Temperature
Colloid	Hydralazine	Nitroglycerine	Total CO2
Creatinine Blood*	Hydromorphone	Nitroprusside	Ultrafiltrate
Creatinine Urine	INR	Norepinephrine*	Urine*
D5W	Insulin Humalog	Normal Saline	Vancomycin
DBP*	Insulin NPH	O2 Saturation	Vasopressin
Dextrose Other	Insulin Regular	OR/PACU Crystalloid	WBC
Dobutamine*	Insulin largine	PCO2	Weight
Dopamine*	Intubated	PO intake	pH Blood
EBL	Jackson-Pratt	PaO2*	pH Urine
Emesis	KCl	PT	
Eoisinophils	KCl (Bolus)	PTT	

A.5 ONE-SHOT PROMPT

For the given 24 hours of measurements, apply the definitions of the SOFA score given by the following table:

```
{| class="wikitable"
!
!Central nervous system
!Cardiovascular system
!Respiratory system
!Coagulation
!Liver
!Renal function
|-
!Score
! [[Glasgow coma scale]]
! Mean arterial pressure OR administration of vasopressors required
! PaO<sub>2</sub>/FiO<sub>2</sub> <nowiki>[mmHg (kPa)]</nowiki>
! Platelets (x10<sup>3</sup>/ul)
! Bilirubin (mg/dl) [umol/L]
! Creatinine (mg/dl) [umol/L] (or urine output)
|-
! +0
| 15 || [[Mean_arterial_pressure|MAP]] >= 70 mmHg || >= 400 (53.3) || >= 150 || < 1.2 [< 20] || < 1.2 [<
↪  110]
|-
! +1
| 13-14 || MAP < 70 mmHg || < 400 (53.3) || < 150 || 1.2-1.9 [20-32] || 1.2-1.9 [110-170]
|-
! +2
| 10-12 || [[dopamine]]  <= 5 ug/kg/min or [[dobutamine]] (any dose) || < 300 (40) || < 100 || 2.0-5.9
↪  [33-101] || 2.0-3.4 [171-299]
|-
! +3
| 6-9 || dopamine > 5 ug/kg/min OR [[epinephrine]] <= 0.1 ug/kg/min OR [[norepinephrine]]  <= 0.1 ug/kg/min
↪  || < 200 (26.7) '''and''' mechanically ventilated including CPAP || < 50 || 6.0-11.9 [102-204] ||
↪  3.5-4.9 [300-440] (or < 500 ml/day)
|-
! +4
| < 6 || dopamine > 15 ug/kg/min OR epinephrine > 0.1 ug/kg/min OR norepinephrine > 0.1 ug/kg/min || < 100
↪  (13.3) '''and''' mechanically ventilated including CPAP || < 20 || > 12.0 [> 204] || > 5.0 [> 440] (or <
↪  200 ml/day)
|}
```

A patient is septic if suspected infection is positive and the future total SOFA score calculated with a best guess on how the values develop is two points (or more) higher than the current SOFA score.
Only answer like in the given example. Here is the example:
Patient is 63.0 years old and is male. Given all the information in this text, answer the question at the end.
Here are the measurements: DBP at time -23.68: 36.0, SBP at time -23.68: 71.0...
Now answer the following question in the given format:
Doctors suspect an infection, based on this information and the other information in this text, will the patient be classified as septic tomorrow?
First we need to calculate the SOFA scores given the extracted values. The SOFA scores for the current time are the following:
The minimum value of GCS_eye is 4.0, GCS_motor is 6.0 and GCS_verbal is 5.0, this produces the sum 15.0 and means the CNS SOFA is 0.
Because minimum MAP is 43.333, max Dopamine is 0, max Dobutamine is 0, max Epinephrine is 0 and max Norepinephrine is 0 with a patient weight of 80 kg, the cardiovascular SOFA is 1.
Given that minimum PO2 is 141.0 and minimum FiO2 is 1 the calculated PAO2FIO2 is 141.0, this means the respiratory SOFA is 3.
Because the minimum Platelet count is 235.0 the coagulation SOFA is 0.
The maximum Bilirubin (Total) is 1.8 leading to a liver SOFA of 1.
Because total Urine output is 1585.0 and maximum creatinine in the blood is 1.4 the renal SOFA is 1.
To summarize: the patient has a total SOFA score of 6.
Now we need to calculate the SOFA scores with forecasted values. The SOFA scores in the future based on the forecasted values are the following:
The minimum value of GCS_eye will be 4.0, GCS_motor will be 6.0 and GCS_verbal will be 5.0, this produces the sum 15.0 and means the CNS SOFA will be 0.
Because future minimum MAP will be 55.667, future max Dopamine will be 0, future max Dobutamine will be 0, future max Epinephrine will be 0 and future max Norepinephrine will be 0 with a patient weight of 80 kg, the cardiovascular SOFA will be 1.
Given that minimum PO2 will be 141.0 and minimum FiO2 will be 1 the forecasted PAO2FIO2 will be 141.0, this means the respiratory SOFA will be 3.
Because the Platelet count will be 295.0 the coagulation SOFA is going to be 0. The maximum Bilirubin (Total) will be 1.8 leading to a liver SOFA of 1.
Because Urine output will be 1635.0 and maximum creatinine in the blood will be 1.1 the renal SOFA will be 0.
To summarize: the patient will have a future total SOFA score of 5.
The patient will not develop sepsis in the next 24 hours, because total SOFA increased only by -1 and infection is suspected.
The example is now finished. Say "The patient will develop sepsis" in the last sentence if the criteria are met (if total SOFA changed by 2 and infection is suspected).
Patient is 76.0 years old and is female. Given all the information in this text, answer the question at the end.
Here are the measurements: DBP at time -22.97: 56.0, GCS_eye at time -22.97: 4.0, GCS_motor at time -22.97: 6.0...
Now answer the following question in the given format:
The doctors don't suspect an infection, based on this information and the other information in this text, will the patient be classified as septic tomorrow?

A.6 SETUP FOR LEARNING FROM PRECONDITIONS

Table 9: Medical preconditions for five organ systems indicated by ICD-10 codes. In-distribution (ID) data were seen, out-of-distribution (OOD) data were unseen during fine-tuning.

	lung	kidney	coagulation	liver	cardiovascular
ID	J40, J41, J42	N18.9, N28	D68.4, D68.5	K70.0, K70.41	I50.0, I50.9
OOD	J44.9	N19	D68.6	K70.3	I50.1

Table 10: ICD-10 Code frequency for in- and out-of-distribution testsets.

ICD-10 Code	# ID Test	# OOD Test
J40	263	194
J41	239	173
J42	253	194
J44.9	-	166
N18.9	266	172
N19	-	204
N28	281	151
D68.4	242	171
D68.5	244	162
D68.6	-	176
I50.0	264	185
I50.1	-	193
I50.9	229	162
K70.0	230	165
K70.3	-	162
K70.41	215	182
No-Preexisting	274	189

A.7 Discussion of 5% Error Margin

The 5% margin for deviations of predictions is justified by the statistics from our experimental results. For example, $SOFA_{25:48}$ score ranges between 0 and 24. The experimental results in Figure 5 plot scores obtained by dividing predicted values by gold standard values for this metric, showing that a 10% margin (dotted vertical lines) would include false positives, while a 5% margin does not. Similar results are obtained for other predictions.

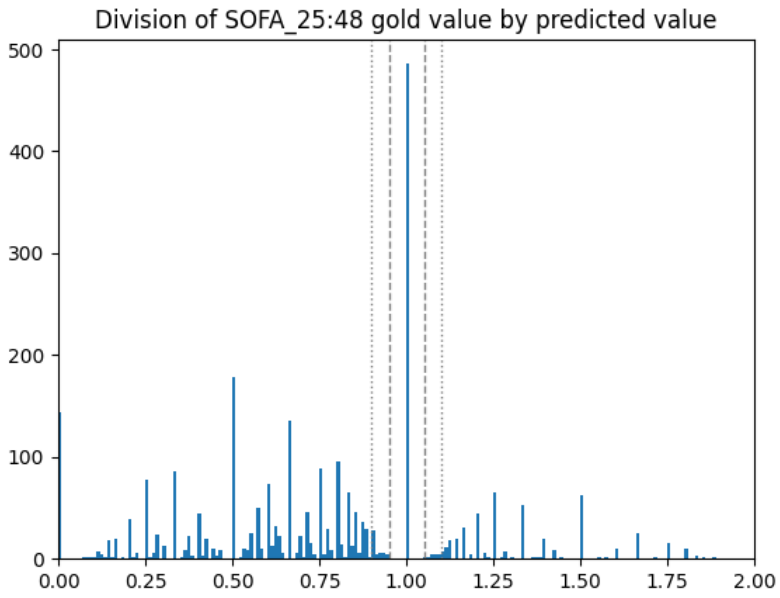

Figure 5: Histogram of the scores obtained by dividing predicted by gold standard values for $SOFA_{25:48}$ score. Dashed vertical lines show the 5% interval, dotted vertical lines show the 10% interval