Pediatric Sepsis Cohort Detection Using In-Context Pointwise $\mathcal{V}$-Usable Information

Yingya Li[1], **Alon Geva**[1], **Steven Bethard**[2], **Timothy Miller**[1], **Kate Madden**[1], **Matthew Eisenberg**[1], **Daniel Kelly**[1], **Guergana Savova**[1]
[1] Harvard Medical School and Boston Children's Hospital
[2] University of Arizona

[1] {firstname.lastname}@childrens.harvard.edu
[2] bethard@arizona.edu

Abstract

Pediatric sepsis diagnosis remains a major clinical challenge due to non-specific symptoms and a lack of reliable diagnostic criteria. Large language models (LLMs) provide a scalable solution for processing and understanding unstructured text in medical records. However, identifying the most suitable model is non-trivial given the rapid growth of available LLMs. In this work, we proposed using in-context pointwise $\mathcal{V}$-usable information (PVI) to estimate task difficulty and guide model selection for pediatric sepsis cohort detection. We applied in-context PVI to estimate task difficulty and inform model selection across 12 state-of-the-art open LLMs on the task, using electronic medical record data from 507 patient encounters at a U.S. children's hospital. We compared the performance of the best-fitting LLM to feature-rich baseline models and a fine-tuned transformer. Our results show that the PVI-selected LLM outperforms the baselines, although the feature-rich bag-of-words model with a support vector machine also achieves competitive performance. We believe our approach demonstrates a promising application of current LLM techniques to high-stakes clinical tasks.

Keywords: large language model, pediatric sepsis detection, model selection and evaluation

1. Introduction

Sepsis is a life-threatening condition resulting from the body's dysregulated immune response to infection, often progressing rapidly to multi-organ failure and death (Singer et al., 2016; Rudd et al., 2020). In high-income countries, sepsis accounts for an estimated 31.5 million cases annually, including 19.4 million severe cases and 5.3 million deaths (Hotchkiss et al., 2016). Diagnosis of pediatric sepsis is particualarly challenging, as children exhibit more subtle and non-specific early symptoms (Watson et al., 2024). Importantly, diagnostic criteria developed for adults often fail to capture the early indicators of pediatric sepsis (de Souza et al., 2024). Although epidemiologic diagnostic criteria based on objective measures of organ dysfunction are emerging (Schlapbach et al., 2024), clinical utility of these criteria remains controversial (Rodriguez and Deep, 2024). Furthermore, sepsis criteria — especially those based on the Electronic Medical Records (EMR) structured data alone such as billing codes, lab results and medication orders — rely on a tradeoff between enough recall to capture most sepsis cases but enough precision to identify a cohort with mortality high enough that it reflects the life-threatening nature of sepsis (Rhee et al., 2019). Thus, nuanced interpretation of longitudinal EMR clinical narratives is critical. This gap highlights an opportunity for natural language processing (NLP) methods to enhance identification of patients with sepsis by leveraging the rich contextual information embedded in EMR clinical narratives.

Large language models (LLMs) have demonstrated strong few-shot learning capabilities and can perform a wide range of NLP tasks via in-context learning (Naveed et al., 2023; Chang et al., 2024). Their adaptability has attracted growing interest in the biomedical domain, leading to a variety of early evaluations and applications across clinical and research settings (Gilson et al., 2023; Zuccon and Koopman, 2023; Chen et al., 2023; Lyu et al., 2023; Singhal et al., 2023; Chen et al., 2024; Liévin et al., 2024; Rider et al., 2025; Shool et al., 2025). For sepsis, recent studies applied GPT4 and open LLMs for early sepsis detection (Shashikumar and Nemati, 2024) or to construct time series analysis for sepsis care (Noroozizadeh and Weiss, 2025). LLMs have also been explored as components of agent-based clinical support systems, aimed at enhancing diagnostic accuracy, guiding treatment decisions, and facilitating coordinated patient care (Cho et al., 2025). However, these efforts raise the question of which LLMs are best suited for the task in a field with frequent introductions of new LLMs or updates of existing ones. Given this growing number of available models, exhaustively evaluating all candidates is both time- and resource-intensive. It is particularly challenging for high-stakes applica-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 336–349

tions such as pediatric sepsis cohort identification, where rapid decision-making is essential in real clinical settings.

In this work, we propose to utilize pointwise $\mathcal{V}$-usable information (PVI) (Ethayarajh et al., 2022) and its in-context learning extension (Lu et al., 2023) to measure the usable information of the pediatric sepsis identification task given a collection of LLMs, thus selecting the most suitable model. PVI (Ethayarajh et al., 2022) estimates the difficulty of data instances for a given model in supervised learning. It builds on the predictive $\mathcal{V}$-information framework (Xu et al., 2020) which incorporates mutual information and the coefficient of determination to quantify data instance difficulty. The metric applies instance-level predictions to quantify how much information a given model can extract from a dataset. The higher the PVI estimate, the easier it is for the model to represent a given data point. In our work, we utilize the recently proposed in-context PVI (Lu et al., 2023) to measure the difficulty of the sepsis identification task given a collection of 12 current state-of-the-art (SOTA) open LLMs with the goal of selecting the best fitting model, i.e., the model with the highest in-context PVI estimates. Our contributions are summarized as follows:

- We applied the recent in-context PVI method as a model selection criterion for pediatric sepsis cohort detection in EMRs, evaluating a range of SOTA open LLMs. We compared performance against baselines including a feature-rich method (Support Vector Machine), fine-tuning a language model, and a code-based sepsis screening algorithm using structured EMR data.
- We developed a novel approach to select the best-matched LLM for pediatric sepsis detection with only a small number of annotated examples. Unlike conventional strategies that rely on large labeled datasets, our method uses few-shot learning with in-context PVI estimates. This contribution is especially important as it addresses a high-stakes clinical application with limited labeled data and demonstrates the broader practical utility of recent advances in language modeling.

2. Related Work

2.1. Sepsis criteria

The recently adopted Phoenix criteria for pediatric sepsis defines sepsis as suspected infection with a Phoenix Sepsis Score of at least 2 points on a scale of respiratory, cardiovascular, coagulation, and/or neurological system organ dysfunction (Schlapbach et al., 2024). Although children meeting these criteria had over an 8-fold increased mortality compared to children with suspected infection but a lower Phoenix Sepsis Score, the primary outcome in the model development and validation paper was mortality (Sanchez-Pinto et al., 2024), and not a clinical impression of sepsis that would fit an "intention to treat" framework. In contrast, the Improving Pediatric Sepsis Outcomes (IPSO) Collaborative uses providers' intention to treat sepsis as the defintion of "suspected sepsis" and a definition paralleling the International Pediatric Sepsis Consensus Conference definition of severe sepsis (Goldstein et al., 2005) to define IPSO critical sepsis (Paul et al., 2023).

Most of the work in the latter domain involves laborious chart review to ascertain clinicians' intent during treatment. While some data, such as medications and other interventions, can be found in the EMR structured data, clinical reasoning is best determined from review of the EMR unstructured clinical narrative. Only the clinical text distinguishes patients treated with antibiotics, evaluated for possible infection, and experiencing organ dysfunction. It indicates whether the dysfunction was present at baseline and whether sepsis-like treatments were given for other indications.

2.2. In-context learning

With the rise of LLMs (Brown et al., 2020; Liu et al., 2023) in-context learning (ICL) has emerged as a prominent learning paradigm. ICL differs from supervised learning in that it does not involve parameter updates through backpropagation. Instead, it relies entirely on pretrained language models, using a few exemplars in the prompt while keeping model parameters unchanged (Brown et al., 2020). This offers an interpretable alternative to traditional supervised training, and significantly reduces the computational overhead (Dong et al., 2022). ICL has demonstrated strong performance across a wide range of NLP tasks (Kojima et al., 2022; Saparov and He, 2022; Srivastava et al., 2022; Wei et al., 2022a,b). A very recent study suggests that ICL with long-context models benefits from large demonstration sets (in the thousands) to reach relatively stable performances (Bertsch et al., 2025).

Many studies have examined the mechanism of ICL. For instance, Akyürek et al. (2022) and von Oswald et al. (2022) analyzed ICL in regression settings and showed that transformer-based models approximate gradient descent implicitly, drawing connections to gradient-based meta-learning. Similarly, Dai et al. (2022) proposed that ICL functions as an implicit form of fine-tuning, interpreting it as meta-optimization where the transformer acts as a meta-optimizer. Exemplars guide the model to generate meta-gradients via forward passes, which are

then integrated into predictions through attention mechanisms. Empirical analyses further support that ICL closely resembles fine-tuning in terms of prediction behavior, representation shifts, and attention dynamics.

2.3. PVI and task dataset difficulty

Task difficulty plays an important role in guiding model design and learning strategies (Torralba and Efros, 2011; Zhao et al., 2022; Cui et al., 2023), and has been extensively studied in NLP (Hahn et al., 2021; Perez et al., 2021; Zhao et al., 2022; Gadre et al., 2023). As an extension of the predictive $\mathcal{V}$-information framework (Xu et al., 2020), pointwise $\mathcal{V}$-usable information (PVI) quantifies instance-level difficulty by measuring the lack of usable information for a given model (Ethayarajh et al., 2022). Higher PVI scores indicate easier examples for the model, while lower scores suggest greater difficulty. PVI has shown utility in dependency quality estimation (Kulmizev and Nivre, 2023), reasoning evaluation (Prasad et al., 2023), intent detection (Lin et al., 2023), and multi-task learning (Li et al., 2025).

Importantly, Lu et al. (2023) extended PVI to ICL with LLMs (in-context PVI). Unlike the original PVI, which relies on fine-tuning a model with input-output and output-only instances, the in-context variant uses two few-shot prompts — one with both input and output, and another with output-only to estimate task difficulty. Empirical results across seven datasets and eight LLMs show that the in-context PVI remains stable across different exemplar selections and shot counts, making it a reliable metric for assessing task difficulty in ICL given an LLM. This is especially valuable in settings where curated training data is scarce and fine-tuning LLMs is constrained by computational resources. In this work, we applied in-context PVI as a proxy for task difficulty to guide LLM selection for the task of pediatric sepsis detection.

3. Method

Our method consists of two primary stages: (1) computing in-context PVI estimates across candidate models, and (2) selecting and applying the model with the highest PVI estimate for the task of sepsis cohort identification.

Figure 1 shows how the original PVI score (Ethayarajh et al., 2022) is calculated. Specifically, it involves fine-tuning a given model $\mathcal{G}$ in two different and separate setups indicated by g' and g. For g', $\mathcal{G}$ is fine-tuned with the input-output pairs $\{(x_i, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For g, $\mathcal{G}$ is fine-tuned only using the outputs, $\{(\varnothing, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For each instance in a dataset $\mathcal{D}$, given a model $\mathcal{G}$, a higher PVI score indicates that the instance x provides more usable information to the model $\mathcal{G}$.

Algorithm 1 The calculation of PVI estimate

1. Input: a dataset D, a model G, a test instance $(x, y) \notin D$
2. $g' \leftarrow$ fine-tune G on $\{(x_i, y_i)|(x_i, y_i) \in D\}$
3. $\emptyset \leftarrow$ empty string
4. $g \leftarrow$ fine-tune G on $\{(\emptyset, y_i)|(x_i, y_i) \in D\}$
5. $\text{PVI}(x \rightarrow y) \leftarrow -\log_2 g[\emptyset](y) + \log_2 g'[x](y)$

Figure 1: The calculation of PVI estimate.

Adapted from the original PVI, in-context PVI (Lu et al., 2023) uses two few-shot prompts, i.e., an *input-output* prompt $p' = (x_1, y_1, x_2, y_2, ..., x_n, y_n, x)$ and a *null-output* prompt $p = (\varnothing, y_1, \varnothing, y_2, ..., \varnothing, y_n, \varnothing)$, to prompt $\mathcal{G}$ rather than fine-tuning it. In-context PVI, denoted as $C(x, y)$, is calculated as Equation (1):

$$C(x, y) = -\log_2 \mathcal{G}[p](y) + \log_2 \mathcal{G}[p'](y). \quad (1)$$

Given the observation that in-context learning resembles fine-tuning (Dai et al., 2022), $\log_2 \mathcal{G}[p](y)$ and $\log_2 \mathcal{G}[p'](y)$ are the ICL approximations of $\log_2 g[\varnothing](y)$ and $\log_2 g'[x](y)$ in Figure 1.

Unlike the original PVI, the calculation of in-context PVI is generation-based, producing an output as a sequence of tokens (e.g., ["not", "sepsis"]) (Lu et al., 2023). Instead of asking the model to produce a label prediction such as "not sepsis", we assign each label a numerical index and require numerical indices as outputs given prompt p' or p (Tables A.1 and A.2 in Appendix A). This ensures that the model's expected output is a single token, which simplifies the calculation of $\log_2 \mathcal{G}[p](y)$ and $\log_2 \mathcal{G}[p'](y)$.

Based on individual PVI estimates for each instance obtained from a given model $\mathcal{G}$, our proposed method further compares the average and total PVI values across all candidate models on a subset of the dataset. The model with the highest overall PVI performance is then selected for the subsequent sepsis inference task and inferencing on the rest of the dataset. See section 4 for more details regarding the experiment setup.

Thus, we take advantage of in-context PVI characterstics as reported in Lu et al. (2023) – its consistency and stability across different selections of exemplars and numbers of shots. This is in contrast to other metrics such as accuracy and F1 which stabilize with more data.

4. Experimental Setup

4.1. Datasets

Our sepsis dataset was collected from screening records for the IPSO Collaborative at a participating

freestanding children's hospital in the United States. The labels thus represent real-world classification for patients seen in the emergency department (ED) or intensive care unit (ICU) with potential sepsis. All patients in the data set met screening criteria for possible sepsis – receipt of systemic antimicrobials and microbiological testing thus presenting a complex decision-making task with multiple potential diagnosis. Patients were labeled as having severe sepsis if they met IPSO Critical Sepsis criteria (Paul et al., 2023). Encounter-level labels were previously assigned by local IPSO Collaborative screeners as part of that quality improvement initiative. If ED and ICU screeners disagreed on the critical sepsis label, the patient encounter was considered to have had critical sepsis, in keeping with our encounter-level labeling (e.g., a patient may not have been considered to have critical sepsis in the ED, but did in the ICU, and thus had critical sepsis for the encounter).

The initial dataset included clinical notes from 796 patient encounters, with a total of 87,190 notes (average word count per note = 709). Notes were aggregated at the encounter level. For patients with more than 6,000 words across notes from the given encounter, we restricted the dataset to those written from three days before to one day after screening for possible sepsis, based on time anchors in the IPSO dataset; patients without anchor information were excluded. The final dataset included 507 patient encounters, where each encounter represents a no/yes critical sepsis (subsequently abbreviated simply "not sepsis" and "sepsis") label and could include multiple clinical notes. We applied stratified sampling based on aggregated note lengths to divide this dataset into train (60%), dev (20%), and test (20%) sets. Table 1 shows the label distribution across the data splits. Table A.3 in Appendix A reports aggregated word counts per encounter.

Label	Input	Train	Dev	Test	Total
Not sepsis	Short	12	4	3	19
	Long	39	16	18	73
Sepsis	Short	33	14	11	58
	Long	220	63	74	357
Total	Short	45	18	14	77
	Long	259	79	92	430

Table 1: Distribution of sepsis labels and input lengths (SHORT: encounters with clinical notes of fewer than 6,000 words; LONG: encounters with clinical notes longer than 6,000+ words).

4.2. Models

Baseline models We compared the performance of LLMs with two baseline models for the sepsis classification task. Our first baseline considers Support Vector Machines (SVM) with different text vectorization approaches as baseline models. SVM with a bag-of-words (BoW) representation is a widely used baseline model in many NLP tasks. The model also fits our task, as it can scale effectively to longer documents without the context length limitations faced by transformer-based models. We used the Scikit-learn Python package (Pedregosa et al., 2011) SVM implementation. The penalty parameter (C) for the LinearSVM classifier was set to 1. A preliminary comparison between various text vectorization methods indicated that term frequency (count vectorization) outperformed tf-idf, and the inclusion of bigram features further enhanced SVM classification performance. The SVM model was trained on the entire training set. Final results were evaluated on the test set.

We chose Longformer [1] (Beltagy et al., 2020), as the second baseline. The Longformer model extends the context window of transformer models like BERT (Devlin et al., 2019) or Roberta (Liu et al., 2019) following work on techniques such as blending local and global attention mechanisms and sliding window attention (Beltagy et al., 2020; Li et al., 2023). We fine-tuned the model on the entire training set and performed hyperparameter tuning on the development set. Specifically, we experimented with training epochs {5, 6, 7, 8, 9, 10}, random seeds {42, 52, 62, 72, 82}, and learning rates {1e-5, 2e-5}, using a fixed batch size of 32 and gradient accumulation steps of 2. Final results were evaluated on the test set.

In addition, we provided a baseline performance comparison to that of EMR structured data in the form of International Classification of Diseases (ICD) codes (World Health Organization, 2024), which are widely available and commonly used for EMR-based cohort identification. Specifically, for all patient encounters included in our study, we extracted all instances of ICD-9 or ICD-10 codes for "severe sepsis" or "septic shock" (R65.2, R65.20, R65.21, 995.92, 785.52). Since diagnostic codes in our EMR were assigned only at the encounter level and not linked to specific notes or dates, the presence of any such code at any point during an encounter was treated as an indicator of critical sepsis.

4.3. Setting and evaluation

We evaluated twelve LLMs for model selection and performance evaluation, including Mistral-7B,

[1] We used longformer-base-4096 (148 million parameters and token length 4,096). As an academic center, we do not have access to computational resources necessary to finetune larger language models with billions of parameters.

Mistral-8×7B, LLaMA-3 (1B, 3B, 8B, 70B), and Qwen-3 (0.6B, 1.7B, 4B, 8B, 14B, 32B). Encounters were categorized by input length: SHORT (<6,000 words) and LONG (>6,000 words). Each encounter text was constructed by concatenating all associated clinical notes. In-context PVI estimates were computed for each LLM using two-shot prompts, aligned with the number of dataset classes and randomly sampled from the SHORT training set, applied to the remaining SHORT examples (N = 43). The PVI estimates follow the in-context PVI procedure outlined in the original study Lu et al. (2023) through the HuggingFace Transformers API (Wolf et al., 2020). Figure 2 shows the sample prompt template used for labeling aggregated encounter-level notes. Model performance was evaluated using accuracy, along with weighted precision, recall, and F1 scores to account for class imbalance and emphasize the clinical importance of correctly identifying sepsis cases.[2]

```
<s>[INST] «SYS»
You are an expert in pediatric sepsis
identification. Below is a clinical document.
Please remember the following clinical context
and answer how likely is the given patient
has sepsis or not?  Keep in mind that
patients without sepsis may still present with
symptoms such as tachycardia, and may have
received fluid boluses (e.g., normal saline)
and antibiotics initially.
Here are some examples: «/SYS»
$FEWSHOT_EXEMPLARS
Here is the clinical document:
<text>
$ALL_ENCOUTNER_CLINICAL_NOTES
</text> [INST] how likely is the given patient
has sepsis or not?
Please only use one word:  0:not_sepsis,
1:sepsis [/INST]
```

Figure 2: Sample prompt template for labeling pediatric sepsis based on aggregated clinical notes at the encounter level.

[2]The weighted scores are computed by first calculating the metric (precision, recall and F1) for each label, and then averaging them using the support (the number of true instances for each label) as weights. This approach adjusts the macro-level average to account for label imbalance. The weighted calculation can also result in an F1-score that is not between precision and recall.

5. Results

5.1. Utilizing the in-context PVI scores to select the best fitting model

Prior studies have shown that PVI estimates, under both fine-tuning and ICL are correlated with prediction accuracy (Ethayarajh et al., 2022; Lu et al., 2023). Note that an in-context PVI estimate is relative, as pointed out in Lu et al. (2023), a value by itself is not as important as when compared to a range. Table 2 presents the in-context PVI estimates alongside performance metrics for the evaluated LLMs on the training set with the SHORT inputs (N = 43). We used only the SHORT instances because as we already pointed out above in-context PVI estimates are sufficiently stable, they show minimal variation across exemplar selections as demonstrated in Lu et al. (2023), while other metrics, e.g. F1, accuracy require larger datasets to get stable estimates. The SHORT inputs from the train set can be used for reliable evaluation with significantly lower inference time than having both SHORT and LONG inputs. We performed a Spearman rank correlation analysis (Zar, 2005) to evaluate the relationship between the average in-context PVI scores and model performance (measured by weighted-F1 scores) among the 12 models. The result revealed a strong and statistically significant positive association (ρ = 0.699, p-value = 0.011), indicating that models with higher PVI scores tend to achieve better prediction performance. For instance, Mixtral has the highest average PVI (0.520) and achieves the highest weighted-F1 score (0.757). Similarly, Qwen 4B (PVI = 0.095), 8B (PVI = 0.426), 14B (PVI = 0.024), and 32B (PVI = 0.214) also show a higher weighted-F1 scores compared to the other models with lower PVI estimates. In general, LLMs with positive in-context PVI estimates outperform LLMs with negative in-context PVI estimates (per weighted F1); note that the ranking of in-context PVI estimates is not exactly the same as the ranking based on the other metrics. This result could indicate more stable estimates for in-context PVI compared to other metrics, which was shown to be the case in Lu et al. (2023) (thus our motivation for using it as a guide for model selection).

5.2. A comparison between the selected LLM and baseline models

Table 3 reports the best performance on the test set of BoW with SVM, the fine-tuned Longformer, and the best-fitting model based on in-context PVI (Mixtral) across different encounter input types (SHORT, LONG, and ALL which combines both SHORT and

Model	Avg PVI	Sum PVI	Prec.	Rec.	W-F1	Acc.
Mistral-7B	0.191	8.225	0.702	0.651	0.669	0.651
Mixtral	0.520	22.389	0.789	0.744	0.757	0.744
LLaMA-1B	-0.054	-2.324	0.694	0.581	0.608	0.581
LLaMA-3B	0.345	14.829	0.681	0.605	0.628	0.605
LLaMA-8B	-0.088	-3.769	0.754	0.581	0.604	0.581
LLaMA-70B	0.343	14.402	0.688	0.721	0.698	0.721
Qwen-0.6B	-0.131	-6.101	0.692	0.512	0.536	0.512
Qwen-1.7B	-0.614	-26.407	0.706	0.465	0.476	0.465
Qwen-4B	0.095	4.091	0.711	0.744	0.716	0.744
Qwen-8B	0.426	18.328	0.786	0.674	0.695	0.674
Qwen-14B	0.024	1.036	0.825	0.698	0.717	0.698
Qwen-32B	0.214	8.346	0.803	0.721	0.738	0.721

Table 2: The average and sum of in-context PVI values and model performance on the train set with the SHORT context input for all the candidate models. Both PVI values and performance metrics are measured on the SHORT input instances (N = 43). Two instances are used as exemplars and the remaining 43 for evaluation.

Input	Model	Prec.	Rec.	W-F1	Acc.
Short	BoW+SVM	0.617	0.786	0.691	0.786
	Longformer	0.617	0.786	0.691	0.786
	Mixtral	0.654	0.571	0.604	0.571
Long	BoW+SVM	0.756	0.804	0.751	0.804
	Longformer	0.647	0.804	0.717	0.804
	Mixtral	0.814	0.826	0.778	0.826
All	BoW+SVM	0.752	0.802	0.744	0.802
	Longformer	0.643	0.802	0.714	0.802
	Mixtral	0.749	0.792	0.758	0.792
	ICD	N/A	N/A	0.632	0.614

Table 3: Comparison of model performance on the test dataset across input types (SHORT, LONG, ALL): accuracy and weighted precision, recall, and F1.

LONG input [3]. Overall, performance was higher on LONG inputs. Mixtral achieved the best results on the combined test set (ALL), with the highest F1-score of 0.758 across all inputs and 0.778 on LONG. In contrast, the fine-tuned Longformer and BoW+SVM baselines yielded lower F1-scores. The ICD code-based algorithm achieved an accuracy of 0.614 and weighted-F1 of 0.632 when evaluated on ALL test set.

Unlike the BoW+SVM and fine-tuned Longformer models, which used the full training dataset (304 examples), our approach utilized only the SHORT subset (45 examples), providing greater efficiency. Because BoW+SVM and Longformer used the entire training dataset (304 examples), the results reported reflect their best possible performance. Table A.4 (in Appendix A) compares BoW+SVM and Longformer trained on SHORT alone versus the combined SHORT+LONG dataset. For BoW+SVM, using only SHORT consistently reduced weighted F1 compared to SHORT+LONG. For Longformer, all cases were assigned to the positive category, leaving weighted F1 unaffected by training input. These findings highlight the efficiency and effectiveness of our approach, especially for pediatric sepsis cohort detection.

To account for the long text in our dataset, we examined how shot number affects the best-performing LLM's performance. Figure 3 presents results for SHORT, LONG, and ALL inputs across few-shot settings ranging from 2 to 10 using Mixtral. Performance was consistently higher for LONG inputs, with the largest improvements in weighted F1 scores (Appendix A Table A.5 includes the detailed performance report). Notably, 2-shot prompting on LONG inputs achieved the best overall test performance (F1 = 0.778), with similarly strong results on the training set (F1 = 0.781). These findings suggest that even minimal in-context supervision can be effective when the input context is sufficiently rich.

[3]We conducted independent runs by randomly sampling three sets of two-shot exemplars for the Mixtral model, applying three random seed values (42, 52, 62) for Longformer, and testing three C parameter values (0.1, 0.5, 1) for the SVM classifier. A one-way ANOVA on weighted F1 scores for long inputs revealed a significant difference.

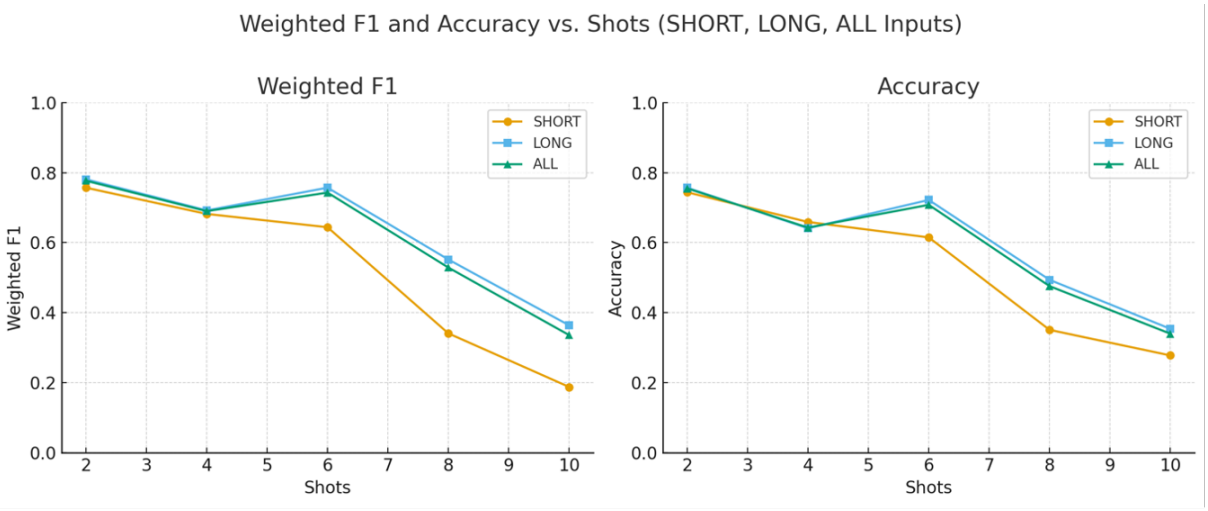

Figure 3: Sample prompt template for labeling pediatric sepsis based on aggregated clinical notes at the encounter level.

As the number of shots increases, performance on LONG inputs generally declines but remains relatively robust through 6-shot prompting. For SHORT inputs, performance drops more sharply. For example, the weighted-F1 score decreases from 0.757 at 2-shot to just 0.188 at 10-shot. This pattern may stem from the smaller number of SHORT instances in our dataset and the difficulty of shorter contexts in handling longer prompts or multiple exemplars.

We further compared in-context PVI and weighted F1 as model selection metrics. Both metrics identified the same best-performing model (Mixtral, Table 2). However, they diverged on the second-ranked model – in-context PVI selects Qwen-8B while weighted F1 selects Qwen-32B. There is also a difference in their results: in-context-PVI-selected Qwen-8B yields 0.685 weighted F1 and 0.651 accuracy, while weighted-F1-selected Qwen-32B yields 0.677 weighted F1 and 0.642 accuracy on the ALL test set. We believe the in-context PVI is a more nuanced metric than F1 as it abstracts away the text contributions from the label distribution. Using a model ranking that jointly considers F1 and in-context PVI could further strengthen the model selection process.

5.3. Utilizing the in-context PVI scores to select exemplars

Hard instances may serve as better exemplars, as models are already adept at handling easier examples, which makes them less informative for demonstration (Lu et al., 2023). To test whether in-context PVI can guide exemplar selection, we used the lowest-PVI training instance from each category as exemplars for the Mixtral model, with scores computed from an initial set of randomly chosen exemplars. Low in-context PVI values indicate instances that are particularly challenging for the model.

Table 4 shows the best overall performance was still achieved with two randomly selected exemplars. However, as the number of shots increased, performance with random exemplars declined, particularly for SHORT and ALL inputs while PVI-selected exemplars maintained more stable results, especially for LONG inputs. These findings underscore the need to balance exemplar difficulty and input context in complex tasks like sepsis detection. While difficult exemplars do not always yield the highest accuracy, in-context PVI offers a useful strategy for stabilizing performance in many-shot settings.

6. Discussion and Conclusion

In this study, we explored a recent task difficulty measurement specifically geared for LLMs – in-context PVI – as a model selection criterion for the high-stakes application of pediatric sepsis cohort identification using real world EMR data of 507 patient encounters. We conducted experiment on a collection of 12 open LLMs for task difficulty estimates and model selection. Our experimental results show that the LLM (Mixtral) selected by in-context PVI has the best model performance on our task. The commonly used evaluation approach normally compares candidate models on the full training set (N = 304) using metrics such as F1, requiring substantial labeled data. In contrast, our approach identifies the best-performing model from a small labeled subset (N = 43), leveraging the stability and consistency of in-context PVI.

The selected best-performing Mixtral model outperformed the two widely used baseline models

Input	Ex.	Shots	Prec.	Rec.	W-F1	Acc.
Short	Rnd.	2	0.654	0.571	0.604	0.571
		4	0.850	0.500	0.518	0.500
		6	0.708	0.500	0.540	0.500
		8	0.563	0.286	0.286	0.286
		10	0.046	0.214	0.076	0.214
	Low	2	0.654	0.571	0.604	0.571
		4	0.676	0.429	0.464	0.429
		6	0.708	0.500	0.540	0.500
		8	0.676	0.429	0.464	0.429
		10	0.850	0.500	0.518	0.500
Long	Rnd.	2	0.814	0.826	0.778	0.826
		4	0.737	0.609	0.648	0.609
		6	0.770	0.707	0.729	0.707
		8	0.742	0.663	0.692	0.663
		10	0.692	0.304	0.295	0.304
	Low	2	0.762	0.804	0.763	0.804
		4	0.736	0.707	0.719	0.707
		6	0.761	0.739	0.749	0.739
		8	0.787	0.750	0.764	0.750
		10	0.750	0.739	0.744	0.739
All	Rnd.	2	0.749	0.792	0.758	0.792
		4	0.745	0.594	0.635	0.594
		6	0.762	0.679	0.707	0.679
		8	0.728	0.613	0.650	0.613
		10	0.691	0.292	0.291	0.292
	Low	2	0.730	0.774	0.744	0.774
		4	0.728	0.670	0.693	0.670
		6	0.752	0.708	0.725	0.708
		8	0.772	0.708	0.730	0.708
		10	0.752	0.708	0.725	0.708

Table 4: Comparison of weighted precision, recall, F1, and accuracy across exemplar types (Random vs. Low) for different shot settings (2–10) and input types using Mixtral on the test dataset. “Low” indicates exemplars with low in-context PVI; “Random” indicates randomly selected exemplars. Exemplars were selected from SHORT instances in the train set, balanced across classes.

BoW+SVM and fine-tuned Longformer particularly for the LONG and combined ALL input using two randomly selected exemplars. It also outperformed the existing ICD-code-based screening algorithm with absolute improvements of 0.178 in accuracy and 0.126 in weighted-F1. Error analysis revealed that the Mixtral model was sometimes misled by confounding cues, such as cancer (a risk factor for sepsis but not evidence of sepsis itself) and asthma (which can produce infection-like symptoms), resulting in false positives. However, the simpler BoW+SVM method still demonstrated competitive performance, especially for SHORT contexts, though in practice it requires larger labeled datasets (in our case, the entire train set) compared to our few-shot prompting approach which utilized only a subset of the train set. Due to the complexity of the sepsis detection task where symptoms often map to multiple differential diagnoses, effective contextualization becomes essential. This may also explain the improved performance on the LONG inputs of the best performing model in our task. Future work could explore agentic methods that combine the strengths of different models via a summarization agent (Celikyilmaz et al., 2018; Noroozizadeh and Weiss, 2025).

Motivated by prior studies suggesting that few-shot ICL may benefit from harder exemplar demonstrations, we conducted preliminary experiments to evaluate the effectiveness of PVI scores in selecting exemplars that enhance sepsis detection performance. Our results show that the best-performing prompt setting still comes from low-shot random exemplar selection for this task, but using exemplars with the lowest PVI scores yields more stable performance across increasing shot numbers. To fully leverage the potential of in-context PVI for exemplar selection, we plan to conduct a more comprehensive study on efficient exemplar sampling and selection, such as combining PVI estimates with larger demonstrations, exemplar ordering, and sparse attention mechanisms (Bertsch et al., 2025). We leave this as a direction for future work.

We believe our study offers a valuable contribution to addressing the challenges of real-world, high-stakes healthcare tasks and demonstrates the potential of applying state-of-the-art language technologies to clinical data. Continued research on datasets such as ours is essential for identifying practical use cases and advancing toward real-world deployment. Important directions include examining model calibration and understanding how model outputs can be effectively integrated into clinical workflows. Additionally, extending evaluation to clinical notes for other types of disorders would help assess the generalizability of our approach across subdomains. Such efforts will be critical for uncovering remaining gaps and enabling broader clinical adoption of state-of-the-art methods.

Limitations

Our study has some limitations. First, we were unable to evaluate proprietary models such as the GPT-4 family of models on the sepsis detection task due to the lack of access to HIPAA-compliant models and restrictions on transmitting protected health information to public APIs. Second, our dataset includes only clinical notes; incorporating additional data sources such as laboratory text reports as well as verbalizing the information stored in the structured part of the may further improve LLM performance, and we plan to include these in future work. Third, our current analysis focused on us-

ing LLMs under the ICL setting for PVI calculation and prompt design. With greater computational resources, future work could explore fine-tuning smaller LLMs (e.g., Mistral-7B), as well as evaluating models trained on clinical text or with longer context windows, to enable a more comprehensive comparison of modeling strategies for long-input settings.

Ethics Statement

The clinical datasets used in this study are identifiable clinical datasets to reflect the real-world scenario of the high-stakes task of sepsis detection. All data were processed locally on a HIPAA-compliant server, and no information was transmitted to any public APIs. This study was conducted under an approved IRB protocol.

Acknowledgements

The work is funded by grants R01GM114355 and R01LM013486 from the US National Institutes of Health.

Bibliographical References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology*, 9(10):1459–1462.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Andrew Cho, Jason M Woo, Brian Shi, Aishwaryaa Udeshi, and Jonathan SH Woo. 2025. The application of matec (multi-AI agent team care) framework in sepsis care. *arXiv preprint arXiv:2503.16433*.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. 2023. Learning sample difficulty from pre-trained models for reliable prediction. *arXiv preprint arXiv:2304.10127*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Daniela C de Souza, Raina Paul, Rebeca Mozun, Jhuma Sankar, Roberto Jabornisky, Emma Lim, Amanda Harley, Samirah Al Amri, Maha Aljuaid, Suyun Qian, et al. 2024. Quality improvement programmes in paediatric sepsis from a global perspective. *The Lancet Child & Adolescent Health*, 8(9):695–706.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $mathcalv$-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

Brahm Goldstein, Brett Giroir, Adrienne Randolph, et al. 2005. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. *Pediatric critical care medicine*, 6(1):2–8.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, 9:891–908.

Richard S Hotchkiss, Lyle L Moldawer, Steven M Opal, Konrad Reinhart, Isaiah R Turnbull, and Jean-Louis Vincent. 2016. Sepsis and septic shock. *Nature reviews Disease primers*, 2(1):1–21.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Artur Kulmizev and Joakim Nivre. 2023. Investigating UD treebanks via dataset difficulty measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1076–1089.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Yingya Li, Timothy Miller, Steven Bethard, and Guergana Savova. 2025. Identifying task groupings for multi-task learning using pointwise v-usable information. *Journal of Biomedical Informatics*, page 104881.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sheng Lu, Shan Chen, Yingya Li, Danielle Bitterman, Guergana Savova, and Iryna Gurevych. 2023. Measuring pointwise $\mathcal{V}$-usable information in-context-ly. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15739–15756, Singapore. Association for Computational Linguistics.

Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian.

2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Shahriar Noroozizadeh and Jeremy C Weiss. 2025. Reconstructing sepsis trajectories from clinical case reports using LLMs: the textual time series corpus for sepsis. *arXiv preprint arXiv:2504.12326*.

Raina Paul, Matthew Niedner, Ruth Riggs, Troy Richardson, Heidi Gruhler DeSouza, Jeffery J Auletta, Frances Balamuth, Deborah Campbell, Holly Depinet, Leslie Hueschen, et al. 2023. Bundled care to reduce sepsis mortality: the improving pediatric sepsis outcomes (ipso) collaborative. *Pediatrics*, 152(2):e2022059938.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pages 8500–8513. PMLR.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*.

Chanu Rhee, Maximilian S Jentzsch, Sameer S Kadri, Christopher W Seymour, Derek C Angus, David J Murphy, Greg S Martin, Raymund B Dantes, Lauren Epstein, Anthony E Fiore, et al. 2019. Variation in identifying sepsis and organ dysfunction using administrative versus electronic clinical data and impact on hospital outcome comparisons. *Critical care medicine*, 47(4):493–500.

Nicholas L Rider, Yingya Li, Aaron T Chin, Daniel V DiGiacomo, Cullen Dutmer, Jocelyn R Farmer, Kirk Roberts, Guergana Savova, and Mei-Sing Ong. 2025. Evaluating large language model performance to support the diagnosis and management of patients with primary immune disorders. *Journal of Allergy and Clinical Immunology*.

Isadora Rodriguez and Akash Deep. 2024. Phoenix criteria for sepsis: are these enough to guide a clinician? *European Journal of Pediatrics*, 183(11):5033–5035.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211.

L Nelson Sanchez-Pinto, Tellen D Bennett, Peter E DeWitt, Seth Russell, Margaret N Rebull, Blake Martin, Samuel Akech, David J Albers, Elizabeth R Alpern, Fran Balamuth, et al. 2024. Development and validation of the phoenix criteria for pediatric sepsis and septic shock. *Jama*, 331(8):675–686.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Luregn J Schlapbach, R Scott Watson, Lauren R Sorce, Andrew C Argent, Kusum Menon, Mark W Hall, Samuel Akech, David J Albers, Elizabeth R Alpern, Fran Balamuth, et al. 2024. International consensus criteria for pediatric sepsis and septic shock. *Jama*, 331(8):665–674.

Supreeth P Shashikumar and Shamim Nemati. 2024. A prospective comparison of large language models for early prediction of sepsis. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 109–120. World Scientific.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*.

R Scott Watson, Enitan D Carrol, Michael J Carter, Niranjan Kissoon, Suchitra Ranjit, and Luregn J Schlapbach. 2024. The burden and contemporary epidemiology of sepsis in children. *The Lancet Child & Adolescent Health*, 8(9):670–681.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization. 2024. International classification of diseases (ICD). Accessed: 2025-05-19.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Xinran Zhao, Shikhar Murty, and Christopher Manning. 2022. On measuring the intrinsic few-shot hardness of datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3955–3963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*.

A. Appendix

Table A.1 and A.2 shows examples of prompts used for in-context PVI calculation.

Context: [the clinical notes of specific patient encounter] **Question:** Is this (0) not sepsis, or (1) sepsis? **Answer:** 0
Context: [the clinical notes of specific patient encounter] **Question:** Is this (0) not sepsis, or (1) sepsis? **Answer:** 1

Table A.1: An example of a 2-shot *input–target* prompt for the sepsis task.

Answer: 0
Answer: 1

Table A.2: An example of a 2-shot *null–target* prompt for the sepsis task.

	Mean	**Median**	**Max**	**Min**
Train	84,647	28,271	1,502,326	1,313
Dev	96,824	24,748	1,305,950	1,454
Test	75,736	21,773	938,151	1,674

Table A.3: Summary statistics (mean, median, max, min) of word counts per encounter across splits.

Table A.4 below shows that training SVM on only the SHORT input type led to lower performance of weighted F1 compared to the SHORT+LONG training dataset (BoW+SVM 0.615 v 0.691, 0.652 v 0.751, 0.648 v 0.744); the results for the Longformer showed that all cases were assigned to the "Sepsis" category, therefore weighted F1 was not affected.

Test Set	**Model (Training)**	**Precision**	**Recall**	**Weighted F1**	**Accuracy**
SHORT	BoW+SVM (all train)	0.617	0.786	0.691	0.786
	Longformer (all train)	0.617	0.786	0.691	0.786
	BoW+SVM (short train)	0.589	0.643	0.615	0.643
	Longformer (short train)	0.617	0.786	0.691	0.786
LONG	BoW+SVM (all train)	0.756	0.804	0.751	0.804
	Longformer (all train)	0.647	0.804	0.717	0.804
	BoW+SVM (short train)	0.652	0.652	0.652	0.652
	Longformer (short train)	0.647	0.804	0.717	0.804
ALL	BoW+SVM (all train)	0.752	0.802	0.744	0.802
	Longformer (all train)	0.643	0.802	0.714	0.802
	BoW+SVM (short train)	0.645	0.651	0.648	0.651
	Longformer (short train)	0.643	0.802	0.714	0.802

Table A.4: Comparison of model performance on the test dataset across input types (SHORT, LONG, ALL) using only SHORT training data (short train) versus combined SHORT+LONG training data (all train).

Input	Shots	Precision	Recall.	Weighted F1	Accuracy.
Short	2	0.789	0.744	0.757	0.744
	4	0.816	0.659	0.682	0.659
	6	0.708	0.615	0.644	0.615
	8	0.722	0.351	0.341	0.351
	10	0.830	0.278	0.188	0.278
Long	2	0.823	0.757	0.781	0.757
	4	0.842	0.641	0.692	0.641
	6	0.832	0.722	0.757	0.722
	8	0.839	0.494	0.552	0.494
	10	0.708	0.354	0.364	0.354
All	2	0.817	0.755	0.777	0.755
	4	0.838	0.643	0.690	0.643
	6	0.816	0.708	0.743	0.708
	8	0.828	0.476	0.529	0.476
	10	0.713	0.340	0.336	0.340

Table A.5: Weighted precision, recall, F1, and accuracy across different shot settings (2–10) for SHORT, LONG, and ALL input types using Mixtral on the train dataset. Few-shot exemplars were randomly selected from SHORT instances (N = 45), balanced across classes. Performance was evaluated on the remaining SHORT instances and all LONG instances (N = 259).