

MedDistill: Improving Clinical LLM Performance Through Natural Language Tabular Insights

Stanford CS224N Custom

Joshua Shunk

Department of Computer Science
Stanford University
jshunk@stanford.edu

Patrick Li

Department of Computer Science
Stanford University
prli@stanford.edu

Abstract

Large language models (LLMs) are increasingly used to support clinical reasoning, yet much reliable medical knowledge resides in structured tabular formats—vitals, lab results, and scoring systems—that LLMs cannot directly leverage. We introduce **MedDistill**, a system that extracts interpretable tabular patterns using an Explainable Boosting Machine (EBM), converts them into natural-language knowledge cards, and distills their utility into a student model that requires no knowledge cards at inference time. We train EBMs on MIMIC-IV ICU admissions to predict in-hospital mortality and sepsis, then compress the learned shape functions, feature importances, and interaction terms into a structured insight bank. For each patient note, a personalized knowledge card is assembled by matching mentioned clinical variables against the insight bank’s top features. A teacher LLM is fine-tuned with these cards in context, and a student is trained to match the teacher’s output distribution via a combined cross-entropy and Kullback–Leibler divergence loss. We evaluate on two tasks: a preliminary 22-calculator MedCalc-Bench subset (100 held-out samples) and a custom MIMIC-IV discharge note classification task (3,946 test examples) from which we derive our main results. On discharge notes, the card-augmented teacher reaches 84.7% accuracy and 0.847 F1, significantly outperforming all card-free models ($\sim 79\%$). However, a student distilled from a no-card teacher performs comparably to one distilled from the card-augmented teacher on both tasks, indicating that EBM knowledge did not transfer. We compare our distilled models with a simple knowledge card fine-tuning approach as an ablation, and again find no evidence of EBM insights being helpful. While we show promise for EBM-based knowledge cards to improve LLM performance on medical NLP tasks, we leave an open question on how to train this performance boost into the model.

1 Key Information to include

- Mentor: Arpandeeep Khatua
- Late Days (if you are using any):

Team member	Late days remaining
<Joshua Shunk>	<5>
<Patrick Li>	<4>
Total pooled	<9>
Team size	<2>
Late days applied for the report	<1>

2 Introduction

Large language models (LLMs) have demonstrated impressive performance across a wide range of medical reasoning tasks, from answering board-style multiple choice questions to summarizing clinical notes. However, a fundamental asymmetry persists in how medical knowledge is stored versus how LLMs are trained: a substantial portion of reliable, evidence-based clinical knowledge is encoded in structured tabular formats—electronic health records, ICU vitals, laboratory panels, and validated scoring systems—while LLMs are predominantly trained on unstructured natural language. Bridging this gap is a central challenge in the deployment of LLMs in real clinical settings.

Existing approaches to this problem fall into two broad categories. The first is retrieval augmentation, where structured data is converted into text at inference time and prepended to the LLM’s context window. Although flexible, this approach requires the structured source to be available at deployment time and places the burden of interpreting the numerical context on the model’s in-context reasoning capabilities. The second is fine-tuning on synthetic question-answer pairs derived from tabular data, which bakes structured knowledge into model parameters, but typically requires large volumes of labeled examples and offers limited insight into what the model has actually learned.

We propose **MedDistill**, a system that occupies a third position: extracting structured knowledge from tabular data using an interpretable model, verbalizing that knowledge into natural-language *knowledge cards*, and then distilling the cards’ utility into a compact student model that requires no structured input at inference time. Our interpretable backbone is an Explainable Boosting Machine (EBM) [1], a glass-box generalized additive model that exposes learned univariate shape functions and pairwise interaction terms. We train EBMs on MIMIC-IV ICU admissions [2] to predict in-hospital mortality and sepsis, then export feature importances, risk direction trends, and high-risk thresholds into a structured *insight bank*. At training time, a patient-specific knowledge card is assembled from the insight bank by matching clinical features mentioned in a given patient note against the EBM’s top features, and this card is prepended to the instruction prompt of a teacher LLM. A student model is then trained on the same inputs *without* any card, optimizing a combined cross-entropy and KL divergence loss against the teacher’s output distribution—following the rule distillation framework of Yang et al. [3].

We evaluated on two tasks in the course of our experiments. We conducted preliminary experiments on a few (100) examples from a 22-calculator subset of MedCalc-Bench [4]. These results indicated a poor fit of insight bank (feature effects predicting mortality/sepsis) and benchmark (medical calculation), with no discernible benefit from presence of the EBM knowledge card. Thus, we constructed a second, more appropriate MIMIC-IV discharge note classification task (3,946 test examples, predicting sepsis/mortality or the lack thereof). Here, the card-augmented teacher reaches 84.7% accuracy and 0.847 F1, significantly outperforming all card-free models (~79%), confirming that the knowledge cards carry meaningful clinical signal. However, a key finding is that a student distilled from a *no-card* teacher achieves accuracy comparable to a student distilled from a card-augmented teacher, suggesting that the KL distillation objective itself—rather than EBM card content transfer—is the primary driver of student performance currently. We analyze this result and identify concrete directions, including refining our pipeline, reinforcement learning, and soft prompting where more robust improvements may be possible.

Our main contributions are: (1) an EBM-to-insight-bank pipeline that converts interpretable tabular model outputs into a retrievable knowledge store; (2) a dynamic knowledge card construction mechanism that personalizes cards to individual patient notes; (3) an empirical study of KL-based distillation as a training objective for clinical reasoning tasks across both numerical calculation and binary classification settings; and (4) diagnostic findings that identify current limitations of card-teacher alignment and motivate concrete future directions.

3 Related Work

LLMs have been applied to a wide range of clinical tasks, including medical question answering, clinical note summarization, and structured report generation. Benchmarks such as MedQA [5] and MedMCQA evaluate general medical knowledge, while MedCalc-Bench [4] targets the harder subproblem of numerical clinical calculation, requiring models to extract variables from free-text patient notes and apply validated scoring formulas. Even GPT-4 achieves only around 77.5%

accuracy on select MedCalc-Bench tasks at benchmark release time, confirming that numerical clinical reasoning remains an open challenge. A natural question is whether the structured clinical knowledge embedded in tabular health records can be used to close this gap.

One principled way to extract knowledge from tabular clinical data is through interpretable models. Explainable Boosting Machines (EBMs) [1], developed from the intelligible boosting framework of Caruana et al. [6], are generalized additive models with pairwise interaction terms that achieve near-black-box predictive accuracy while remaining fully inspectable. Each feature contribution is represented as a shape function, enabling identification of risk trends, high-risk thresholds, and feature interaction strengths. EBMs have demonstrated strong performance on clinical tabular tasks including ICU mortality prediction [1]. However, the outputs of such models are numerical and structured, and how to translate them into a form useful for LLM training is an open problem that MedDistill directly addresses.

The most closely related work to ours is Tables2Traces [7], which also attempts to unlock medical tabular data for LLM training by converting individual rows into trace-style instruction examples and fine-tuning directly on the resulting pairs. Rather than converting rows, MedDistill uses an EBM to extract dataset-level patterns that aggregate statistical knowledge across all patients, which generalizes beyond individual records and avoids encoding patient-specific noise. This distinction also has implications for the training regime: direct fine-tuning as used in Tables2Traces requires a large volume of examples to be effective, whereas the insights produced by an interpretable model are moderate in quantity but high in quality, motivating a different training objective.

Knowledge distillation [8] is well suited to exactly this regime, transferring the soft output distribution of a teacher model into a student via a KL divergence objective over temperature-scaled logits rather than relying on large labeled datasets. Subsequent work extended this to intermediate hidden representations, showing that aligning internal layer activations further improves student performance [9]. Nagarajan et al. [10] further showed that students can sometimes exceed teacher performance on held-out tasks, consistent with our own results. These findings motivate using distillation as the core training mechanism in MedDistill.

The specific distillation framework we build on is that of Yang et al. [3], who propose encoding explicit textual rules into LLM parameters by aligning a student’s output distribution, without the rule in context, to a teacher that receives the rule as input. Their ablation study shows that hidden state alignment is critical for rule internalization beyond what logit-level KL alone provides. MedDistill adopts this framework but departs from it in one key respect: rather than encoding hand-authored rules, we derive rules automatically from EBM shape functions and verbalize them into dynamic, patient-specific knowledge cards, allowing the approach to scale to large clinical datasets without manual rule writing.

4 Approach

Our approach consists of four stages: (1) training an interpretable model on structured clinical data, (2) compressing its outputs into a queryable insight bank and generating patient-specific knowledge cards, (3) fine-tuning a teacher LLM that conditions on these cards, and (4) distilling the teacher’s behavior into a student model that requires no cards at inference time.

Figure 1 provides an overview.

4.1 EBM Training on Structured ICU Data

We train an Explainable Boosting Machine (EBM) [1] on MIMIC-IV ICU admissions [2] to predict in-hospital mortality/sepsis. The EBM is a generalized additive model with pairwise interactions:

$$g(\mathbb{E}[y]) = \beta_0 + \sum_{j=1}^p f_j(x_j) + \sum_{(j,k) \in \mathcal{I}} f_{jk}(x_j, x_k) \quad (1)$$

where g is the logit link, each f_j is a univariate shape function learned via cyclic gradient boosting over fine-grained bins, and $\mathcal{I}$ contains the top 50 pairwise interactions selected by FAST interaction detection. We use high-resolution binning (max_bins = 4096), 25 outer bagging rounds for stability, a learning rate of 0.005, and 200 smoothing rounds to regularize the shape functions. The model

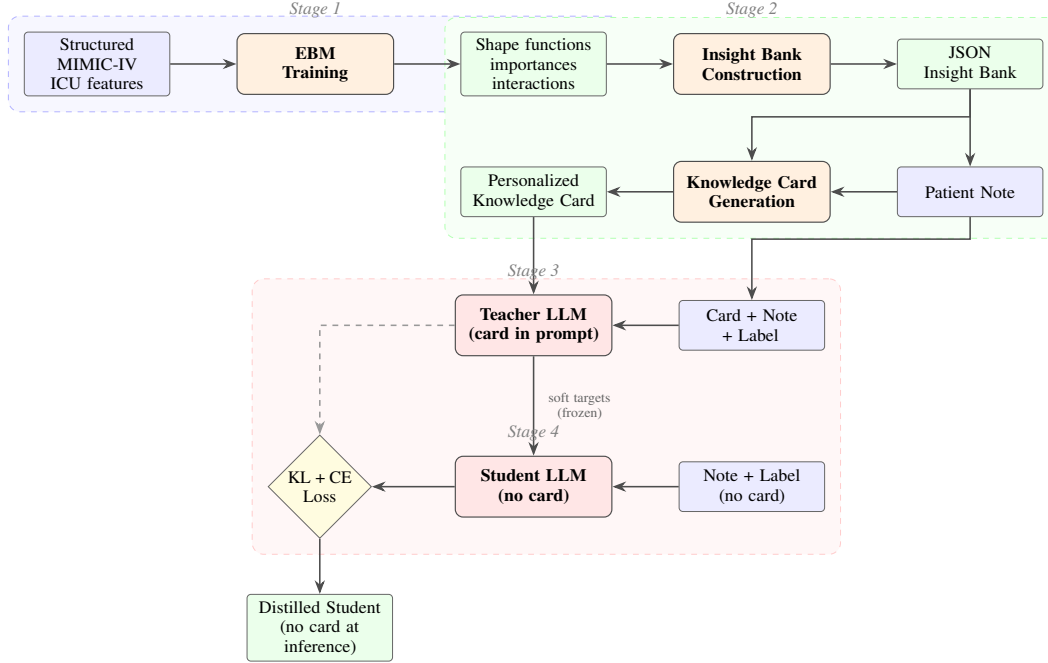


Figure 1: Overview of the MedDistill pipeline. **Stage 1:** An EBM is trained on structured MIMIC-IV ICU data, producing interpretable shape functions and feature importances. **Stage 2:** EBM outputs are compressed into a JSON insight bank; for each patient note, clinical variables are matched against the bank to produce a personalized knowledge card. **Stage 3:** A teacher LLM is fine-tuned with the knowledge card prepended to the prompt. **Stage 4:** A student LLM is trained on card-free inputs to match the frozen teacher’s output distribution via combined KL divergence and cross-entropy loss. At inference, only the student is deployed—no card or EBM is required.

trains on approximately 20–45 clinical features spanning vitals (heart rate, blood pressure, SpO₂, temperature, respiratory rate), laboratory values (lactate, creatinine, bilirubin, albumin, WBC), and demographics (age, admission type). We train separate EBMs for mortality and sepsis prediction tasks.

Because each f_j is independently inspectable, the EBM exposes, for every feature: (a) a complete shape function mapping feature values to log-odds contributions, (b) a global importance score proportional to the mean absolute contribution across patients, and (c) pairwise interaction strengths with other features. These structured outputs form the basis of our knowledge extraction pipeline.

4.2 Insight Bank Construction and Knowledge Card Generation

Most of our novel code implements these steps.

Insight bank. We compress the EBM’s per-feature outputs into a structured JSON *insight bank* via several post-processing steps. For each continuous feature, we downsample its shape function to at most 100 representative points, apply rolling-window smoothing, and then: (i) *classify the risk pattern* as monotone-up, monotone-down, U-shaped, N-shaped, or complex, using the relative positions of the smoothed global minimum and maximum (interior vs. edge zones at the 15% threshold) and a monotone diff-ratio test (>80% of differences in one direction); (ii) *extract key thresholds* by detecting zero crossings on the smoothed curve (risk-direction transitions) and locating the global peak and trough risk values; and (iii) *generate a natural-language risk summary* of 1–3 sentences describing the feature’s association pattern, calibrated to its classified risk shape. Binary and categorical features receive simpler presence-based summaries. The resulting insight bank stores, for each feature, its importance rank, unit, display name, text-matching aliases, binned risk scores with confidence intervals, key thresholds, classified risk pattern, natural-language summary, and top interaction partners.

Dynamic knowledge cards. At training time, we generate a personalized knowledge card for each patient note by matching clinical variable mentions against the insight bank. For each feature in the bank, we maintain a curated set of regular-expression aliases (e.g., HR, heart rate, pulse for heart rate) and scan the patient note for matches. Detected features are ranked by their EBM importance score, and the top- K (we use $K=10$) features above a minimum importance threshold of 0.02 are selected. For each selected feature, the card includes its risk-pattern descriptor (e.g., “higher values increase mortality risk”), key thresholds (e.g., “risk increases above ~ 95 bpm”), and notable interaction partners. The card is formatted as a bulleted list, bracketed by a task-specific header and a footer instructing the model to use the EBM-derived evidence alongside the clinical vignette. Importantly, the card describes *general population-level trends* learned by the EBM, not per-patient risk scores—the LLM is expected to integrate these trends with the patient-specific values it reads from the note.

For the discharge note evaluation, the appropriate insight bank is selected based on the patient’s condition: the sepsis-trained EBM bank for patients with sepsis labels, and the mortality-trained bank for patients with mortality labels, ensuring the card content aligns with the most relevant clinical risk model.

4.3 Teacher Fine-Tuning

We fine-tune a teacher LLM on clinical prediction tasks using the Alpaca instruction-following format. Each training example consists of a task instruction, a patient note as input, and a target output (“Yes”/“No” for the discharge task; a numeric answer for MedCalc-Bench). For the teacher model, the knowledge card is prepended to the instruction field, providing EBM-derived clinical context that the model can leverage during training. A parallel baseline model is trained on identical data but *without* knowledge cards in the instruction field. Task-specific hyperparameters are detailed in Section 5.3.

We use DeepSeek-R1-Distill-Qwen-7B as the base model, fine-tuned with LoRA [11] under 4-bit NormalFloat quantization (NF4 with double quantization). LoRA adapters with rank $r=16$ and scaling factor $\alpha=16$ are applied to the query, key, value, and output projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`), with dropout 0.05. In all configurations, the loss is computed only over response tokens (`train_on_inputs=False`).

4.4 Knowledge Distillation

Following the rule distillation framework of Yang et al. [3], we train a student model to internalize the teacher’s card-augmented behavior without ever receiving the knowledge card as input. The student and teacher share the same base architecture (DeepSeek-R1-Distill-Qwen-7B). During distillation, the teacher processes the *card-augmented* version of each training example (from the `train_full` split) while the student processes the *card-free* version (from the `train_no` split). The teacher’s weights are frozen.

The training objective combines a Kullback–Leibler divergence loss on output logits with the standard cross-entropy loss:

$$\mathcal{L} = \text{KL}\left(\sigma\left(\frac{\mathbf{z}^T}{\tau}\right) \parallel \sigma\left(\frac{\mathbf{z}^S}{\tau}\right)\right) \cdot \tau^2 + \lambda \cdot \text{CE}(\mathbf{z}^S, y) \quad (2)$$

where $\mathbf{z}^T$ and $\mathbf{z}^S$ are the teacher and student logits, σ is the softmax function, τ is the temperature (set to 1), λ is the cross-entropy weight (set to 1.0), and y is the ground-truth label. Both losses are computed only over response tokens, masked by a label indicator that excludes instruction and input tokens (positions with label = -100). The τ^2 scaling ensures gradient magnitudes remain comparable across temperature settings [8].

5 Experiments

We evaluate MedDistill on two tasks: (1) a preliminary **MedCalc-Bench** clinical calculator evaluation requiring numerical reasoning over patient notes, and (2) a **MIMIC-IV discharge note** binary classification task predicting whether a patient has sepsis/in-house mortality or not. Both tasks test

whether EBM-derived knowledge, distilled through a teacher, can improve a student model that never sees the knowledge card.

5.1 Data

MedCalc-Bench. MedCalc-Bench [4] is a benchmark of clinical calculator tasks grounded in patient notes from medical examinations. Details are left in the appendix for space considerations, as we do not rely on this benchmark for any of our main results. Crucially, our teacher model with knowledge card unexpectedly underperformed the baseline finetuning approach on our test subset, indicating a poor fit between EBM knowledge card insights and the MedCalc-Bench evaluation.

MIMIC-IV discharge notes. For a more directly relevant evaluation, we construct a binary classification dataset from MIMIC-IV [2] discharge summaries. The positive class ($y=1$) includes patients diagnosed with sepsis or who died in-hospital; the negative class includes other ICU patients (hard negatives, not healthy controls). In this dataset, MIMIC-IV EBM insights found in our knowledge bank should be directly helpful for the classification task. We downsample negatives to achieve a balanced 1:1 ratio, yielding approximately 26,000 examples. We split by `subject_id` (ensuring no patient leakage) into 18,352 training, 3,980 validation, and 3,946 test examples, with $\sim 50\%$ positive rate in each split.

To prevent label leakage, each discharge note is truncated at the first occurrence of outcome-revealing sections (e.g., “Brief Hospital Course,” “Discharge Diagnoses,” “Expired”), and always-anonymized administrative headers (name, dates, unit number) are stripped. A hard token budget of 4,076 tokens (4,096 minus 20 tokens of headroom), measured using the `tiktoken cl100k_base` tokenizer as a proxy for the Qwen2 tokenizer, is enforced by trimming only the note text from the tail. Approximately 11.5% of notes reach this cap.

For each example, a dynamic knowledge card is generated using the appropriate EBM insight bank: the sepsis-trained bank for patients with sepsis labels, and the mortality-trained bank for mortality-only patients. Cards select the top $K=10$ features (minimum importance 0.02) detected in the note text. Two parallel JSON files are produced per split: `*_full.json` with the knowledge card prepended to the instruction, and `*_no.json` without.

5.2 Evaluation method

MedCalc-Bench. Our preliminary results were evaluated following the MedCalc-Bench evaluation protocol [4] (more details in appendix).

MIMIC-IV discharge notes. For the binary classification task, we extract “Yes” or “No” from the model’s generated text (case-insensitive, first occurrence after stripping any chain-of-thought `<think>` blocks). We report accuracy, macro-averaged F1, and the confusion matrix (false positives and false negatives). We compute 95% bootstrap confidence intervals using 10,000 resamples of the test set.

5.3 Experimental details

Models. All experiments use DeepSeek-R1-Distill-Qwen-7B as the base model. We fine-tune with LoRA [11] under 4-bit NormalFloat (NF4) quantization with double quantization. LoRA adapters use rank $r=16$, scaling factor $\alpha=16$, dropout 0.05, applied to the query, key, value, and output projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`). All experiments use a random seed of 1234.

Training configurations. We train four primary model variants per task:

1. **Baseline:** Standard fine-tuning on card-free data (`train_no.json`). Serves as the lower bound.
2. **Teacher:** Fine-tuning on card-augmented data (`train_full.json`). Evaluated with the card present at inference, serving as an upper bound.
3. **Student-EBM:** KL distillation from the teacher. The teacher processes card-augmented inputs; the student processes card-free inputs and matches the teacher’s output distribution. Evaluated without a card.

4. **Student-Base:** KL distillation from the baseline (control). Same distillation procedure, but the teacher has no EBM knowledge. Isolates the effect of the distillation objective itself from EBM knowledge transfer.

For MedCalc-Bench, teacher and baseline train for 3 epochs with batch size 16, micro-batch size 1, learning rate 10^{-4} , 100 warmup steps, maximum sequence length of 4,096 tokens, and a validation set of 953 examples (80/20 stratified split of the ICU-relevant training data). Distillation runs for 3 epochs with batch size 32, learning rate 3×10^{-4} , 50 warmup steps, cosine schedule, and the same KL + CE loss configuration.

For the discharge task, teacher and baseline models train for 3 epochs with batch size 32 (micro-batch 4, gradient accumulation 8), learning rate 10^{-4} , 100 warmup steps, a maximum sequence length of 4,096 tokens, and a validation set of 3,980 examples. Distillation runs for 1 epoch with learning rate 5×10^{-4} , 30 warmup steps, micro-batch size 1, cosine learning rate schedule, temperature $\tau=1$, cross-entropy weight $\lambda=1.0$, and KL divergence loss. Hidden-state distillation is disabled. We only conduct 1 epoch of distillation due to time constraints, with each run taking 15 hours.

We additionally evaluate two sequential fine-tuning variants on the discharge task as an ablation for our distillation methodology. These variants seek to assess whether our knowledge cards are directly useful for fine-tuning: **Baseline-Plus**, which first fine-tunes on card-augmented data for 1 epoch (lr 10^{-4}) then continues on card-free data for 1 epoch (lr 5×10^{-5}), and its control (**Baseline-Plus Control**) which uses card-free data in both stages. Both Baseline-Plus and Baseline-Plus Control are evaluated without the knowledge card, so the second epoch on card-free data ensures performance is not degraded by formatting differences between train and test time.

Inference. All inference uses greedy decoding (temperature 0, no sampling, beam width 1). Maximum generation length is 16 tokens for the discharge task (where the answer is a single “Yes”/“No”) and 256 tokens for MedCalc-Bench (which requires step-by-step numerical reasoning).

EBM training. The Explainable Boosting Machines are trained on structured MIMIC-IV ICU features with high-resolution settings: 4,096 maximum bins, 50 pairwise interactions, 25 outer bagging rounds, learning rate 0.005, 200 smoothing rounds, early stopping after 200 rounds without improvement, and a validation split of 15%. Separate EBMs are trained for mortality prediction (~ 38 features) and sepsis prediction (~ 25 features).

5.4 Results

MedCalc-Bench. Table 1 (in appendix) reports overall accuracy on the 100-example MedCalc-Bench ICU-relevant test set (DeepSeek-R1-Distill-Qwen-7B, 32-shot training). Both distilled students substantially outperform the directly fine-tuned models: Student-EBM achieves 46% and Student-Base achieves 47%, compared to 33% for the baseline and 30% for the teacher evaluated with the knowledge card present. When the teacher is evaluated *without* the card at inference (a setting it was not trained for), accuracy drops to 22%, confirming that the teacher relies on the card during inference and cannot generalize without it.

Notably, the Student-Base control—distilled from the baseline teacher that never saw the knowledge card—performs comparably to Student-EBM (47% vs. 46%), suggesting that the gains stem primarily from the KL distillation training objective itself rather than from transfer of EBM-derived knowledge through the teacher.

MIMIC-IV discharge notes. Figure 2 reports results on the full 3,946-example discharge note test set with 95% bootstrap confidence intervals (10,000 resamples).

The teacher model, which receives the EBM knowledge card at both training and inference time, significantly outperforms all other models (non-overlapping 95% CIs). Its advantage is concentrated in false negative reduction: 278 FNs compared to 434 for the baseline, a 36% decrease in missed positive cases. This confirms that the knowledge card provides meaningful clinical signal.

However, neither distillation variant nor sequential fine-tuning successfully transfers the teacher’s advantage to a card-free model. Student-EBM (79.3%) and Student-Base (79.4%) both perform comparably to the baseline (79.9%), with overlapping confidence intervals across all non-teacher

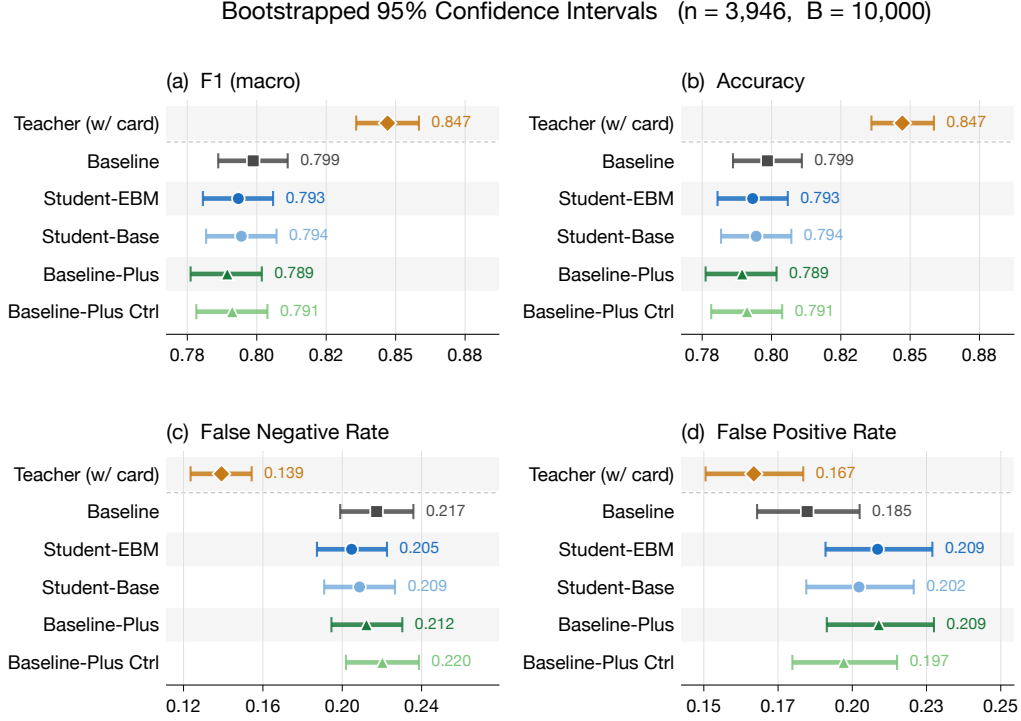


Figure 2: Bootstrapped 95% confidence intervals across all models trained on MIMIC-IV discharge notes

models. The distilled students do shift the decision boundary: the FP/FN ratio moves from 0.83 (baseline) toward 0.99 (Student-EBM), indicating that the students adopt the teacher’s bias toward predicting “Yes” but do not acquire its discriminative accuracy. Of the 319 test examples that the teacher alone answers correctly, Student-EBM recovers 101 (31.7%) while Student-Base recovers 102 (32.0%)—essentially indistinguishable.

The Baseline-Plus model (sequential card exposure followed by card-free fine-tuning) also fails to retain the teacher’s advantage: card exposure in stage 1 improves the Yes class slightly but hurts the No class, yielding a net accuracy below the standard baseline. Across all six models, 70% of test examples are answered unanimously correctly, and 9.4% are answered unanimously incorrectly, suggesting that much of the difficulty lies in inherently ambiguous cases where the truncated discharge note provides insufficient clinical evidence.

6 Analysis

Our results on MIMIC-IV discharge notes improve on those for MedCalc-Bench but are still mixed. As seen in figure 2, our teacher model with knowledge card significantly improves on the fine-tuned baseline on the discharge notes. Unlike for MedCalc-Bench, where our knowledge card had no effect or was even slightly harmful, our results clearly indicate that EBM-based tabular insights can improve LLM performance on medical NLP tasks. However, our student performance is indistinguishable whether the teacher was EBM assisted or not. To rule out the possibility of distillation especially hindering knowledge transfer, we ran the previously described Baseline-Plus fine-tuned ablations as a “simple” way of using the knowledge card during training. Performance was again indistinguishable between using EBM insights and the control.

Indeed, all models that are not the teacher (fine-tuned and provided with the knowledge card at inference) are worse and statistically indistinguishable in terms of performance, with the main difference being that the teacher sees custom-chosen EBM knowledge at inference while every other model does not. With relatively small models (≈ 7 billion parameters) and our limited training time,

Discharge Note (abridged)	EBM Knowledge Card (teacher input only)
CC: LLE Erythema / Hypotension Patient with HTN, atrial fib/flutter, multiple DVTs on coumadin presents with R leg erythema, swelling, pain. Intense cold feeling, trembling. ED: BP 63/38, afebrile, tachycardia to 100. Bandemia 19%, lactate 2.9, Cr 1.6, INR 3.6. CTA: no PE. Resuscitated with 4L NS, vancomycin 1g. Norepinephrine started; BP to mid-90s. MICU: T 96.3, BP 99/43, P 78, O2 99% RA. Labs: WBC 8.2, Hgb 13.4, Neuts 74, Bands 19, Glucose 218, BUN 32, Cr 1.6, HCO₃ 23 UA: WBC 112, RBC >182, few bacteria Imaging: TTE normal. CTA no PE. US no DVT.	ICU Sepsis Risk Factors (personalized): • Creatinine: risk increases above ≥ 1.35 mg/dL • INR: risk increases above ≥ 1.38 • Lactate: risk increases above ≥ 16.9 mmol/L (interacts: Hemoglobin, Hematocrit) • WBC: risk increases above ≥ 12.8 k/μL (interacts: Resp. Rate, GCS) • Temperature: risk increases above $\geq 37.0^{\circ}\text{C}$ • Bicarbonate: extremes increase risk • Hemoglobin: peak risk near 8.30 g/dL • Platelet Count: extremes increase risk

Model	Prediction
Teacher (w/ card)	Yes ✓
Baseline / Student-EBM / Student-Base	No ✗
Baseline-Plus / Baseline-Plus Ctrl	No ✗

Figure 3: The note describes cellulitis with hemodynamic collapse but never mentions “sepsis.” The patient’s creatinine and INR exceed the EBM-derived thresholds and only the teacher, with access to the card and knowledge of relevant features, correctly identifies this pattern. Of 156 such teacher-unique cases, 75% are true positives requiring inference from indirect clinical signals.

it may be too much to ask for a LLM to internalize all EBM insights and learn how to apply them. While we have proof that having the right EBM insights boosts performance, we seem to have run into a training knowledge transfer limitation. Figure 3 provides an example of a discharge note where our teacher model succeeds at diagnosing sepsis and every other model does not.

7 Conclusion

Overall, we showed the potential for tabular understanding through EBM insights to improve LLM performance on medical NLP tasks. However, we were unable to train LLMs to by themselves match the performance of being provided with EBM knowledge in context, leaving an opening for future work. Due to time and resource constraints, we were unable to experiment with different hyper-parameters, larger models, or structural training variations for our distillation or fine-tuning approaches. Recent work has also suggested reinforcement learning as a viable approach for embedding medical knowledge into LLMs [12]. Additionally, soft prompting could be a way of efficiently injecting a large array of EBM insights into LLM context, building on the knowledge card approach that we already show works in the teacher model. We recognize the strong ethical standards needed when conducting research on patient health data and have adhered to the guidelines in PhysioNet’s Data or Specimens Only Research course throughout the research process. We believe a continuation of our work has potential to advance medical NLP and improve quality of life for patients.

Team contributions (Required for multi-person team)

Patrick conducted the bulk of distillation experiments and designed the custom discharge note dataset. Josh proposed and conducted the ablation model training post-distillation as well as the majority of the writeup. Research to develop our methodology and pipeline was shared equally between the two of us.

References

- [1] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.
- [2] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.
- [3] Wenkai Yang, Yankai Lin, Jie Zhou, and Ji-Rong Wen. Distilling rule-based knowledge into large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 913–932, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.
- [4] Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. MedCalc-Bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745, 2024.
- [5] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.
- [7] Mikkel Werling, Nabeel Seedat, Jiashuo Liu, Lars Grønlykke, Carsten Utoft Niemann, Mihaela van der Schaar, and Rudi Agius. Tables2Traces: Distilling tabular data to improve LLM reasoning in healthcare. In *EurIPS 2025 Workshop: AI for Tabular Data*, 2025.
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NeurIPS 2014 Deep Learning Workshop.
- [9] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.
- [10] Vaishnavh Nagarajan, Aditya Krishna Menon, Srinadh Bhojanapalli, Hossein Mobahi, and Sanjiv Kumar. On student-teacher deviations in distillation: Does it pay to disobey? *arXiv preprint arXiv:2301.12923*, 2024.
- [11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.
- [12] Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*, 2025.

A Appendix

A.1 Details of MedCalc-Bench.

MedCalc-Bench [4] is a benchmark of clinical calculator tasks grounded in patient notes from medical examinations. We filter to 22 ICU-relevant calculators whose input variables overlap with our EBM feature set, including APACHE II, SOFA, SIRS Criteria, Mean Arterial Pressure (MAP), Anion Gap, CKD-EPI GFR, MDRD GFR, Creatinine Clearance (Cockcroft-Gault), Serum Osmolality, MELD Na, Sodium Correction for Hyperglycemia, Calcium Correction for Hypoalbuminemia, and others. We construct balanced few-shot training sets of 8, 16, and 32 examples per calculator. Dynamic

knowledge cards are generated for each example using the MIMIC-IV mortality EBM insight bank with $K=8$ and a minimum importance threshold of 0.03. We sampled a fixed 100-example test subset sampled across all 22 calculators to calculate preliminary results.

Our MedCalc-Bench evaluation was conducted to exactly match the specified protocol [4]. For integer-output calculators (e.g., SIRS, CURB-65, GCS), correctness requires an exact match after rounding. For decimal-output calculators (e.g., GFR, Serum Osmolality, FENa), a prediction is correct if it falls within a 5% tolerance window around the ground truth (i.e., $0.95 \cdot y \leq \hat{y} \leq 1.05 \cdot y$). Answers are extracted from model output by attempting JSON parsing, then regex matching for a numeric “answer” field, then extracting the last number found in the text. We report overall accuracy (fraction of 100 test examples answered correctly) and per-calculator accuracy.

Table 1: MedCalc-Bench ICU-relevant test accuracy (DeepSeek-7B, 32-shot, $n=100$).

Model	Accuracy (%)
Teacher (with card)	30
Teacher (no card at inference)	22
Baseline (no card)	33
Student-EBM (distilled from teacher)	46
Student-Base (distilled from baseline)	47

Table 2 presents per-calculator accuracy for the five models. Distillation yields broad improvements: Student-EBM achieves 18/22 on MAP (vs. 17 baseline), 5/6 on Delta Gap (vs. 1), 6/6 on Sodium Correction (vs. 5), and 3/5 on Calcium Correction (vs. 1). Student-Base shows a similar pattern, achieving 19/22 on MAP, 5/6 on Delta Gap, and 6/6 on Sodium Correction. Several complex calculators (CKD-EPI, Creatinine Clearance, MELD Na, FIB-4) remain at or near zero across all models, indicating that these multi-step formulas are beyond what our training regime can teach at 32 examples per calculator.

Table 2: Per-calculator accuracy on MedCalc-Bench (DeepSeek-7B, 32-shot). Only calculators with ≥ 3 test examples are shown.

Calculator	n	Baseline	Teacher	Stu-EBM	Stu-Base
MAP	22	17	10	18	19
CKD-EPI GFR	11	0	0	0	0
Serum Osmolality	7	4	3	5	5
Delta Gap	6	1	3	5	5
Na Correction	6	5	5	6	6
SIRS Criteria	6	0	4	1	1
Anion Gap	5	4	4	4	4
Ca Correction	5	1	0	3	0
MDRD GFR	4	0	0	1	1
Creatinine Cl.	4	0	0	0	0
Delta Ratio	4	0	1	1	2
Alb-Corr Δ Ratio	3	0	0	0	0
Alb-Corr Δ Gap	3	0	0	1	0
MELD Na	3	0	0	0	1