# WHEN LLM CONFIDENCE FAILS: EXTERNAL CALIBRATION FOR TABULAR PREDICTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large language models (LLMs) are increasingly applied to structured clinical predictions, making reliable uncertainty quantification essential for safe deployment. Evaluating across three model families (both open and closed weights), three prompting conditions and nine binary clinical prediction tasks with over $16,000$ combined patient records, we demonstrate that verbalized confidence from LLMs is fundamentally uninformative regarding instance-level correctness on tabular data. By formally disentangling two commonly conflated evaluation metrics, we show that an LLM carrying zero correctness information can deceptively attain a classifier AUROC of $0.806$, while its true self-knowledge inverts to below random chance on complex etiologies like acute kidney injury. To address this, we introduce a scalable, two-call external calibrator that predicts per-instance correctness using task-level descriptors, rank-aggregation statistics, and the divergence between the LLM's self-reported feature attribution and a data-derived reference. Because our calibrator requires no model internals, it is directly applicable to proprietary APIs. While cross-task transfer without target adaptation reveals that external calibration fails on certain targets due to concept shift, adapting the calibrator with a small few-shot target set resolves this degradation. Crucially, source-task pretraining reduces the required target labeling budget by roughly an order of magnitude, achieving correctness AUROC gains (up to $+0.363$) and providing a robust, lightweight guardrail for LLM deployment.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly applied to structured clinical prediction from electronic health records (Brown et al., 2025; Hegselmann et al., 2023; Yildiz et al., 2025). While their performance does not yet match traditional machine learning on these tasks, they do not require task-specific training, making them attractive for environments where a trained model does not yet exist. In that setting, average accuracy is not the quantity that determines whether the model can be safely deployed. In healthcare, a clinician deciding whether to act on a prediction needs to know whether *this specific* prediction is likely correct, and in most deployments, the only available uncertainty signal is what the model says about itself. Prior work has shown that on clinical tabular data, this verbalized signal is uninformative and largely determined by the prompt template rather than the patient instance (Dasula et al., 2026), with models remaining confident regardless of question difficulty or answer correctness (Boie et al., 2026; Xiong et al., 2024). We ask a different question: if verbalized confidence is uninformative, why does it evaluate so well? The answer lies in how it is evaluated. Reported confidence is signed against the predicted class and scored on how well it ranks patients by outcome. Qwen2.5-7B makes this consequence visible, it emits only three distinct confidence values (Appendix E). Its self-knowledge AUROC is exactly $0.500$ on most tasks, and yet the same signed score attains a classifier AUROC of $0.806$. Across three model families and nine tasks, the classifier metric exceeds the self-knowledge metric in $25$ of $27$ combinations.

To address this, we propose an external calibrator that predicts per-instance correctness based on the divergence between the features the model reports as driving its prediction and a reference ordering estimated from the data, supplemented by rank-aggregation statistics and task-level descriptors. Unlike prior cross-model approaches, this reference requires no secondary trained model, it is computed directly from the task data. The calibrator consumes only parsed model outputs, requires two inference calls per record, and needs no access to internal representations, making it directly



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

applicable to the proprietary APIs dominating LLM deployments. In a leave-one-task-out evaluation, cross-task transfer without target-task adaptation succeeds on some targets but fails on others (e.g., dropping to $0.432$ on in-hospital mortality). These failures are consistent across models, illustrating task-level concept shift. Adapting the calibrator with a small labeled target set resolves this. At $k = 200$, the calibrator improves over verbalized confidence on all nine tasks (Gemini, zero-shot), achieving the largest gains precisely where confidence fails ($+0.363$ on AKI, 95% CI $[+0.342, +0.375]$). Furthermore, source-task pretraining is what makes these small budgets viable, contributing $+0.101$ AUROC at 10 labeled examples compared to just $+0.032$ at 200 examples.

**Contributions.** First, we formally disentangle two commonly conflated evaluations of verbalized confidence, showing mathematically and empirically that a model carrying zero correctness information still attains a classifier AUROC of $0.806$. Second, we characterize verbalized confidence across 3 model families and 9 clinical tasks, median confidence remains uniformly high ($0.68 - 0.92$) regardless of task difficulty, and self-knowledge inverts below chance on specific complex etiologies (Appendix E,F). Third, we introduce a scalable, two-inference call external calibrator that requires no model internals or task-specific trained models, and quantify the labeling requirements it needs to mitigate concept shift, source-task pretraining reduces target labels by roughly an order of magnitude (Appendix J.3).

## 2 Related Works

**Language models on clinical tabular data and verbalized confidence.** Applying LLMs to structured EHR data has drawn attention as hospitals consider integrating language models into existing workflows. Brown et al. (2025) find LLMs substantially weaker than gradient-boosted trees across discrimination, calibration, and fairness. Hegselmann et al. (2023) show that serialization format and example selection materially affect few-shot tabular performance and Yildiz et al. (2025) identify poor calibration, limited external validation, and infrastructure cost as persistent barriers. This line of work compares predictive performance between LLMs and structured models. It leaves open the question that matters for deployment, when an LLM is used, can a clinician tell which of its predictions to trust? Previous works attempt to elicit reliability directly via stated probabilities like $P(\text{true})$ (Kadavath et al., 2022). Tian et al. (2023); Xiong et al. (2024) study verbalized numerical confidence and document systematic overconfidence across model families. Heo et al. (2024); Boie et al. (2026) show that verbalized confidence responds to task formatting rather than to correctness, and is largely independent of question difficulty. Lin et al. (2022) instead train models to express uncertainty in words directly. These studies establish that verbalized confidence is poorly calibrated, but they evaluate it against the signed task label, we show that this is responsible for the overstatement. We borrow the term *self-knowledge* from Yin et al. (2023), who ask whether models recognize unanswerable questions, and use it for a graded version, whether its confidence tracks answer correctness.

**External calibration.** Post-hoc calibration adjusts a model's scores using held-out labeled data via Platt scaling, isotonic regression, or temperature scaling (Guo et al., 2017). Conformal prediction (Vovk et al., 2005) provides distribution-free coverage without modifying the confidence itself. Applied to verbalized confidence, these methods are limited by the granularity of the signal, a few distinct values cannot recover ranking information that the variable lacks. Closer to our setting, Shen et al. (2024) train an auxiliary model to predict an optimal temperature scaling parameter using the LLM's internal feature representations. While this shares our structure of an external correctness predictor, it operates on free-form text and requires internal model representations alongside multiple inferences. We instead exploit the structured nature of tabular prediction. By measuring how the LLM's self-reported feature attribution deviates from a data-derived reference, we achieve scalable calibration using only external features and two inference calls. Because the mapping between calibration features and correctness can invert across domains, unsupervised test-time adaptation (Wang et al., 2020a) does not apply. This necessitates the use of a small, labeled few-shot support set to anchor shifting decision boundaries.

**Attribution and rank aggregation.** Feature attribution methods (Lundberg & Lee, 2017) produce per-instance importance rankings for tabular models. While differing methods often produce conflicting rankings within a single model (Krishna et al., 2022), recent cross-model comparisons

propose using attribution divergence as an uncertainty signal on clinical data (Dasula et al., 2026). Comparing these rankings is a classical rank aggregation problem. Dwork et al. (2001) established scalable Markov-chain approximations for rank aggregation, while Sarkar et al. (2014) aggregated rankings from different methods to obtain robust feature orderings by exploiting complementary statistical biases. We apply this mathematical intuition differently, rather than combining rankers for a consensus, we treat the LLM's stated attribution and an empirical data-derived reference as two independent rankers. We use their disagreement, measured by deviations from consensus, inversion counts, and intraclass correlation (Shrout & Fleiss, 1979), as per-instance features predicting whether the model was correct.

# 3 METHODS

## 3.1 PROBLEM SETUP

We consider a set of binary clinical prediction tasks $\mathcal{T} = \{t_1, \ldots, t_T\}$. Each task $t$ is associated with a dataset $\mathcal{D}_t = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_t}$, where $\boldsymbol{x}_i \in \mathbb{R}^{d_t}$ is a structured patient record and $y_i \in \{0, 1\}$ is the clinical outcome. Feature spaces differ across tasks (i.e., $d_t$ varies). An LLM is queried once per row with a serialized rendering of $\boldsymbol{x}_i$ and a task-specific prompt. The response consists of a binary prediction $\hat{y}_i \in \{0, 1\}$, verbalized confidence $c_i$ expressing model's own belief that its own prediction is correct and a per-instance feature attribution $\pi_i$, an ordered list of $K$ record features the model reports as the most influential. A separate query elicits $p_i^{\text{true}} \in [0, 1]$, the model's stated probability that its prediction was correct (Kadavath et al., 2022). We write $\mathcal{L}_t$ for the collection of LLM outputs on task $t$. We set $K = 5$ throughout. Larger $K$ gives a finer comparison between the model's attribution and the reference, but lengthens every response and increases the burden of eliciting a reference ordering of comparable depth from a domain expert (Sec. 6). The question of interest is whether the model was right, $z_i = \mathbb{I}[\hat{y}_i = y_i]$. We refer to $z_i$ as ***LLM correctness***. Our goal is to predict $z_i$ for individual instances *without* additional inferences and access to its internals, since both are unavailable for the closed-weight models in deployment.

## 3.2 TWO METRICS

Verbalized confidence is routinely evaluated by treating it as a classifier score. Because $c_i$ is expressed relative to the predicted class rather than to class 1, this requires a change of reference which we call the *signing operation*:

$$s_i = \begin{cases} c_i & \text{if } \hat{y}_i = 1, \\ 1 - c_i & \text{if } \hat{y}_i = 0 \end{cases} \tag{1}$$

The signed score $s_i$ is then interpreted as an estimate of $P(y_i = 1)$ and evaluated against the clinical outcome. This yields what we call the *classifier* metric, $A_{clf} = \text{AUROC}(s, y)$, which asks whether the model's confidence, once signed, ranks patients by clinical outcome. The quantity relevant to reliability is different, it asks whether the magnitude of $c_i$ is informative about $z_i$, whether the model is more often right when it reports higher confidence. This requires no signing, since $c_i$ is already expressed as a belief about its own correctness, $A_{sk} = \text{AUROC}(c, z)$. We call $A_{\text{sk}}$ the *self-knowledge* metric. The same definitions apply to $p^{\text{true}}$ in place of $c$.

These two quantities are frequently conflated, but they are not interchangeable. Consider a patient with $y_i = 1$ for whom the model correctly predicts $\hat{y}_i = 1$ with $c_i = 0.55$. Under Eq. 1, this contributes $s_i = 0.55$ to a positive case ranking above any predicted negatives, thereby helping the classifier metric $A_{clf}$. However, under $A_{sk}$, this same response assigns a low confidence score to a correct prediction ($z_i = 1$). If the model typically assigns higher confidence (e.g., $c = 0.85$) to its incorrect predictions, this instance ranks below those mistakes, severely penalizing the self-knowledge metric. A single response can thus move the two metrics in opposite directions. The mechanism is simple, signing maps the binary prediction $\hat{y}_i$ into the score, so $s_i$ inherits whatever information $\hat{y}_i$ carries about $y_i$ regardless of how $c_i$ is distributed. In the case where verbalized confidence is globally constant across all instances, $AUROC(s, y)$ reduces exactly to the balanced accuracy of the underlying prediction $\hat{y}$, while $AUROC(c, z)$ collapses entirely to chance (0.500). Qwen2.5-7B approximates this limit directly, emitting only three discrete confidence values, its $A_{sk}$ is 0.500 across seven tasks, yet its $A_{clf}$ deceptively reaches up to 0.806. We take $A_{\text{sk}}$ as the target our calibrator must improve upon. A method that predicts $z_i$ better than $c_i$ does is useful for deployment.

### 3.3 ATTRIBUTION DIVERGENCE FEATURES

Our calibrator represents each patient by how far the LLM's stated reasoning deviates from a data-derived reference, together with descriptors of the task itself. No component requires re-querying the model. Definitions of these features appear in Appendix C.

**Reference Ordering.** For each task $t$, we estimate the mutual information $\mathrm{MI}(x^{(j)}, y)$ between every feature and the binary clinical outcome to induce a reference ranking $\rho_t$. To handle a mixture of continuous and discrete measurements, we utilize a $k$-nearest neighbor entropy estimator designed for mixed discrete-continuous data (Kraskov et al., 2004; Ross, 2014). Categorical and binary features are explicitly masked as discrete during estimation. Stability checks across stratified subsamples ($m = 200$) confirm this ordering is robust under low-data regimes (Appendix G).

**Divergence and aggregation features.** Given the model's per-instance attribution $\pi_i$ and the reference $\rho_t$, we compute 15 features capturing agreement between the two rankings through set overlap, rank correlation, and the position of the model's selections within the task's mutual information distribution (Appendix C.3). Treating $\pi_i$ and $\rho_t$ as two rankers over a shared item set, we add 6 consensus-based statistics drawn from the rank aggregation literature (Sarkar et al., 2014; Dwork et al., 2001), intraclass correlation between the two rankings (Shrout & Fleiss, 1979), deviation from the Borda (de Borda, 1781) and Kemeny (Kemeny, 1959) consensus orderings, Kemeny inversion count, and concentration and flatness of the model's attribution weights (Appendix C.2). These summarize *how* the two orderings disagree rather than *whether* they do.

**Task descriptors.** 14 features describe the reference itself, the top-$k$ mutual information values, gaps between them, their share of total mass, and counts of features exceeding fixed thresholds (Appendix C.4). These are constant within a task and allow the calibrator to learn the task's signal.

**Model Outputs and Feature Space.** Finally, we include the model's prediction $\hat{y}_i$, verbalized confidence $c_i$, and self-assessment probability $p_i^{\text{true}}$. Combined with the 15 divergence features, 6 rank aggregation metrics, and 14 task descriptors, this forms our base feature vector $\boldsymbol{\phi}_i \in \mathbb{R}^{38}$. The full dimensionality $d_\phi$ adapts to model access and prompting style. The base representation yields $d_\phi = 38$ for closed-weight models under zero- or few-shot prompting, while evaluating open-weight models allows the extraction of raw token-level logits for prediction and self-assessment, adding two features ($d_\phi = 40$). Under Chain-of-Thought prompting (Wei et al., 2022), we capture the semantics of the reasoning by appending 384-dimensional sentence embeddings using a local `all-MiniLM-L6-v2` encoder (Wang et al., 2020b), expanding the space to $d_\phi = 422$ for closed-weight and $d_\phi = 424$ for open-weight models.

### 3.4 CALIBRATOR

The calibrator $g_\theta : \mathbb{R}^{d_\phi} \to [0, 1]$ is an XGBoost model (Chen & Guestrin, 2016) trained to predict the correctness target $z_i$. Hyperparameters are held fixed across all experiments (Appendix D). We avoided per-task hyperparameter tuning to ensure the calibrator remains a lightweight and domain-agnostic, keeping the AUROC comparisons interpretable.

**Leave-one-task-out (LOTO).** To evaluate cross-task transfer without target-task adaptation, we hold out task $t$, train $g_\theta$ on $\bigcup_{t' \neq t} \mathcal{L}_{t'}$, and evaluate $\mathrm{AUROC}(g_\theta(\boldsymbol{\phi}), z)$ on $\mathcal{L}_t$. Rather than training on the new task, this evaluates whether generalized LLM correctness behaviors transfer to an unseen domain without tuning the calibrator for this task.

**Few-shot adaptation.** LOTO assumes the mapping from $\boldsymbol{\phi}$ to $z$ transfers across tasks. Where it does not (Sec. 5.4), a small amount of target supervision is required. We draw a class-stratified support set $S \subset \mathcal{L}_t$ with $|S| = k$, train on $\bigcup_{t' \neq t} \mathcal{L}_{t'} \cup S$, and evaluate on $\mathcal{L}_t \setminus S$. Excluding $S$ from evaluation prevents leakage. We set $k \in \{0, 10, 50, 100, 200\}$ and repeat each setting over 5 independent draws.

**Baselines.** To isolate what the source tasks contribute, we train on $S$ alone holding the feature set, model class, and evaluation split fixed (Appendix J.3). The difference between few-shot and target-

Table 1: Clinical prediction tasks. Prevalence is the positive-class rate in the sampled cohort. Accuracy is thresholded at $0.5$

| Task | Source | Feat. | Prev. | XGBoost AUROC | XGBoost Acc. |
|---|---|---|---|---|---|
| Acute kidney injury | MIMIC-IV | 50 | 51% | 0.88 | 0.83 |
| In-hospital mortality | MIMIC-IV | 26 | 36% | 0.77 | 0.75 |
| Hospitalization | MIMIC-IV-ED | 42 | 39% | 0.89 | 0.80 |
| Critical outcome | MIMIC-IV-ED | 42 | 39% | 0.92 | 0.84 |
| ED in-hospital mortality | MIMIC-IV-ED | 42 | 39% | 0.92 | 0.85 |
| ICU transfer (12 h) | MIMIC-IV-ED | 42 | 39% | 0.93 | 0.85 |
| Revisit (3 d) | MIMIC-IV-ED | 42 | 39% | 0.72 | 0.68 |
| Pulmonary embolism | MC-MED | 64 | 60% | 0.95 | 0.89 |
| ACS / myocardial infarction | MC-MED | 63 | 61% | 0.96 | 0.93 |

only performance at a given $k$ measures the value of source-task pretraining at that labeling budget. We also compare against $c$ and $p^{\text{true}}$ directly as predictors of $z$. Maximum softmax probability (Hendrycks & Gimpel, 2016) reduces to $c$ here, since the model reports confidence in its selected class and $c \geq 0.5$ throughout. AUROC differences are accompanied by $95\%$ bootstrap confidence intervals over support draws ($1000$ resamples). Further discussion on baselines Appendix M.

# 4 Experiments

## 4.1 Data and Tasks

We evaluate on $9$ binary clinical prediction tasks drawn from the MIMIC-IV (Johnson et al., 2024), MIMIC-IV-ED (Johnson et al., 2023), and MC-MED (Kansal et al., 2025) databases. As summarized in Table 1, these tasks span intensive care, emergency department triage, and diagnostic classification. They differ significantly in base rate, cohort size, and the type of clinical judgment required. The MIMIC-IV-ED tasks follow the benchmark workflow of Xie et al. (2022), while the MC-MED tasks utilize the standard MC-MED pipeline in (fengx13, 2025). Acute Kidney Injury (AKI) evaluation follows Wu et al. (2020), and in-hospital mortality adapts the pipeline of Harutyunyan et al. (2019) to MIMIC-IV. Full task-definitions appear in Appendix A. For each task, we sample $N \approx 2000$ patients stratified by outcome, except in-hospital mortality (IHM), restricted to $N = 659$ due to cohort limitations. Evaluated cohorts fall slightly below these figures where LLM responses could not be parsed. Records comprise demographics, vital signs, laboratory values, and comorbidity indicators derived from diagnosis codes, serialized as one feature per line using original clinical terms rather than encoded column identifiers (Appendix B).

## 4.2 Models

We evaluate three language models spanning open and closed weights `gemini-3-flash-preview` and `gpt-5.5`, accessed through OpenRouter with zero-data retention for clinical safety, and a local `qwen2.5-7B-instruct` (Qwen et al., 2025). Decoding uses temperature $0.0$ with a maximum of $1000 - 1500$ new tokens. For the local model, we additionally record the logits of the tokens of interest like task-label and confidence, these are unavailable for the two proprietary models (Appendix C.5). Because traditional machine learning remains the gold standard for purely tabular clinical prediction (Brown et al., 2025), we train an XGBoost classifier per task on the exact same feature sets supplied to the LLMs, to maintain fairness. We report this performance in Table 1 not to propose LLMs as tabular replacements, but to establish the task-level difficulty against which reliability must be evaluated.

## 4.3 Prompting conditions

Each model is queried under three conditions applied identically across tasks and models, **Zero-shot** (serialized record, task description, and a request for a prediction, a confidence value, and the $K = 5$ features that most influenced it with the direction of their influence). **Few-shot** (the

Table 2: Prediction accuracy by task and model, zero-shot prompting.

| Task | Gemini | GPT-5.5 | Qwen2.5-7B |
|---|---|---|---|
| Revisit (3 d) | 0.544 | 0.641 | 0.410 |
| In-hospital mortality (IHM) | 0.696 | 0.704 | 0.382 |
| Pulmonary embolism (PE) | 0.633 | 0.507 | 0.429 |
| Acute kidney injury (AKI) | 0.624 | 0.626 | 0.494 |
| Hospitalization | 0.722 | 0.761 | 0.635 |
| Critical outcome | 0.803 | 0.771 | 0.642 |
| ICU transfer (12 h) | 0.838 | 0.775 | 0.683 |
| ED mortality | 0.805 | 0.715 | 0.706 |
| ACS / MI | 0.849 | 0.820 | 0.829 |

same prompt preceded by four labeled exemplars, two per class, drawn once per task, held fixed, and excluded from evaluation), and **Chain-of-thought** (an instruction to reason step by step, with the reasoning retained alongside the structured output). Responses are requested in a fixed JSON schema and parsed into $\hat{y}$, $c$, and $\pi$; invalid responses are retried up to 3 times and unparseable calls are excluded. Full templates appear in Appendix B. Self-assessment is a separate call, identical across the three conditions, supplying the record, the model's own prediction and stated confidence, and a request for the probability that the prediction is correct. For CoT, the generated reasoning is included. This gives $p^{\text{true}}$ and does not revise $\hat{y}$.

# 5 Results

## 5.1 Task performance

Table 2 reports zero-shot accuracy across the nine tasks, spanning from $0.382$ to $0.849$. Across all architectures, 3-day revisit and in-hospital mortality (IHM) are hardest, whereas acute coronary syndrome (ACS/MI) is easiest. All three models underperform the task-dedicated XGBoost reference (Table 1). We observed that models performed consistently better on Emergency Department triage tasks than on intensive care tasks. On the most difficult tasks, accuracy drops to or below chance, GPT-5.5 achieves $0.507$ on pulmonary embolism (PE), while Qwen2.5-7B falls below chance on four tasks, reaching $0.382$ on IHM. These failures establish the baseline against which model self-knowledge must be evaluated.

## 5.2 How confidence is verbalized

Before evaluating if confidence is informative, we characterize what the models actually produce.

**Expressiveness with model capability.** The number of distinct confidence values differs by an order of magnitude across models. Under zero-shot prompting, GPT-5.5 produces $42$ distinct values within $[0.56, 0.99]$, Gemini emits $14$ within $[0.65, 0.99]$, and Qwen2.5-7B collapses to just 3 discrete values within $[0.75, 0.95]$ (Appendix E). While prompting styles have minimal impact on closed models, CoT expands Qwen's output to $10$ distinct values lowering its minimum confidence to $0.35$.

**Confidence is uniformly high and insensitive to difficulty.** Median confidence lies between $0.68$ and $0.92$ on every task-model combination, and never falls below $0.68$, while accuracy over the same combinations ranges from $0.38$ to $0.85$. Mean $p^{\text{true}}$ shows the same pattern, spanning $0.71$ to $0.85$. Qwen's $p^{\text{true}}$ has mean and median both equal to $0.75$ under all three prompting styles, indicating essentially all mass on a single value. The three models organize confidence differently with respect to the predicted class. Gemini separates into two class-conditional centers on most tasks, predicted negatives receive median confidence $0.88$–$0.92$ against $0.72$–$0.82$ for predicted positives, with within-class standard deviation of $0.03$–$0.08$. GPT-5.5 shows a similar pattern with median confidence $0.68$–$0.88$ and larger within-class spread ($0.05$–$0.12$) potentially due to more distinct values. Qwen's confidence is always near-constant (Appendix F.3). Because confidence is close to constant within predicted class, we report task performance using accuracy rather than AUROC.

Table 3: Classifier ($A(s, y)$) and self-knowledge ($A(c, z)$) AUROC under zero-shot prompting. Gap is $A(s, y) - A(c, z)$.

| Task | Gemini | | | GPT-5.5 | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|---|---|---|
| | $A(c, z)$ | $A(s, y)$ | Gap | $A(c, z)$ | $A(s, y)$ | Gap | $A(c, z)$ | $A(s, y)$ | Gap |
| Revisit | 0.671 | 0.616 | −0.055 | 0.562 | 0.608 | +0.046 | 0.499 | 0.502 | +0.003 |
| AKI | 0.447 | 0.602 | +0.155 | 0.462 | 0.598 | +0.135 | 0.500 | 0.500 | +0.000 |
| PE | 0.550 | 0.840 | +0.289 | 0.683 | 0.830 | +0.147 | 0.500 | 0.523 | +0.023 |
| IHM | 0.588 | 0.770 | +0.182 | 0.620 | 0.767 | +0.147 | 0.498 | 0.511 | +0.013 |
| Hosp. | 0.711 | 0.820 | +0.109 | 0.712 | 0.839 | +0.127 | 0.500 | 0.672 | +0.172 |
| Critical | 0.748 | 0.877 | +0.130 | 0.791 | 0.893 | +0.103 | 0.500 | 0.671 | +0.171 |
| ED mort. | 0.742 | 0.897 | +0.154 | 0.837 | 0.906 | +0.069 | 0.500 | 0.741 | +0.241 |
| ICU tr. | 0.727 | 0.897 | +0.170 | 0.798 | 0.912 | +0.113 | 0.500 | 0.702 | +0.202 |
| ACS/MI | 0.745 | 0.903 | +0.158 | 0.824 | 0.917 | +0.094 | 0.500 | 0.806 | +0.305 |

### 5.3 Classifier and Self-Knowledge Metrics Diverge

Table 3 compares the classifier metric $A(s, y)$ against the self-knowledge metric $A(c, z)$. Across 25 of 27 task-model combinations, the classifier metric overstates reliability, by up to 0.305 AUROC. The finding is that the two metrics support opposite conclusions on the same outputs. On PE, for GPT-5.5 $A(s, y) = 0.830$, which indicates a well-discriminating classifier, while predicting at chance (0.507 accuracy) and achieving $A(c, z) = 0.683$. Gemini on the same task attains $A(s, y) = 0.840$ against $A(c, z) = 0.550$, a gap of 0.289. On AKI, self-knowledge AUROC falls below 0.5 on both larger models, 0.447 for Gemini and 0.462 for GPT-5.5. Confidence is not merely uninformative about correctness on this task, it is inversely related to it. Qwen's self-knowledge AUROC is 0.500 on seven of nine tasks and within 0.002 of it on the remaining two. Its $p^{\text{true}}$ AUROC is exactly 0.500 on all nine. Yet Qwen's classifier AUROC reaches 0.806 on ACS/MI and exceeds 0.67 on four tasks. A score containing no correctness information whatsoever produces reliable classifier AUROC values. This is the clearest demonstration available that $A(s, y)$ cannot distinguish a well-calibrated confidence from a constant (Appendix H).

### 5.4 External calibration

Having established that verbalized confidence fails as a reliability indicator, we evaluate whether our model-external calibrator can accurately predict correctness $z_i$.

**Target-reference transfer and Concept Shift.** In the LOTO setting without calibrator-label adaptation ($k = 0$), the calibrator is trained on 8 source tasks and evaluated on the held-out ninth, removing the need for retraining. For Gemini (ZS), the calibrator outperforms raw confidence on 4 of 9 tasks, with the largest gains where the baseline is most miscalibrated, AKI ($\Delta = +0.064$, AUROC 0.511 vs. 0.447) and PE ($\Delta = +0.035$, AUROC 0.585 vs. 0.550), and incurs negative transfer on IHM ($\Delta = -0.156$, AUROC 0.432 vs. 0.588) and hospitalization ($\Delta = -0.117$, AUROC 0.594 vs. 0.711). GPT-5.5 succeeds on 5 of 9 tasks with the same two failures ($\Delta = -0.139$ on IHM, $\Delta = -0.131$ on hospitalization). Qwen improves on 8 of 9 tasks because its raw confidence collapses to a degenerate, near-chance baseline ($A(c, z) \approx 0.500$). AKI remains its sole failure ($\Delta = -0.136$, AUROC 0.364). Full per-task results across all models appear in Appendix I.1. Across both proprietary models, the failure pattern is identical, **IHM** and **hospitalization** consistently degrade, while **ACS/MI** and **3-day revisit** exhibit mild negative transfer. In this setup, the source pool is predominantly composed of acute emergency triage and diagnostic tasks, transferring feature-divergence decision boundaries trained on these to inpatient trajectories inverts the mapping, demonstrating that cross-task transfer is insufficient without task-specific adaptation in such cases.

**Few-Shot Adaptation Overcomes Concept Shift.** Incorporating a minimal support set of $k$ labeled target examples directly resolves task-level concept shift. As illustrated in the improvement heatmap (Figure 1), target supervision systematically restores calibrator performance across all tasks as $k$ increases from 0 to 200. Gains are largest exactly where the base LLM suffered severe confidence inversion, on Gemini (ZS), $k = 200$ yields a +0.363 AUROC improvement over raw con-

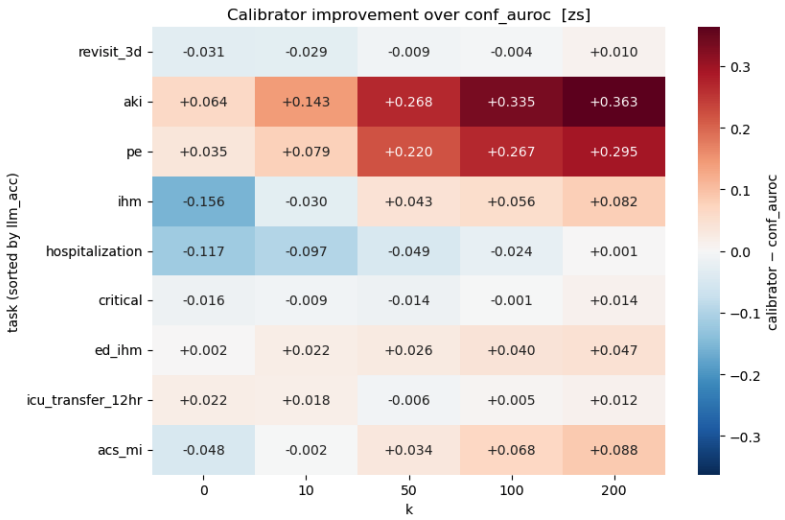


Figure 1: Gemini ZS improvement over the confidence baseline per task and $k$, $k = 0$ is LOTO

fidence on AKI and $+0.295$ on PE. Where zero-shot LOTO transfer initially failed, supervision rapidly reverses the deficit, the $-0.156$ drop on IHM in LOTO is eliminated by $k = 50$ and reaches $+0.082$ at $k = 200$, and the $-0.117$ penalty on hospitalization is eliminated by $k = 200$. On acute triage tasks the calibrator mostly maintains positive gains across all values of $k$ ($+0.012$ to $+0.047$ at $k = 200$). Most improvements are statistically significant with $95\%$ bootstrap confidence intervals in Appendix I.2.1. These adaptation dynamics hold across all evaluated model families and prompting conditions (Appendix I.2). GPT-5.5 mirrors Gemini's trajectory, achieving its largest $k = 200$ gains on the hard tasks ($+0.330$ on AKI, $+0.141$ on PE, and $+0.112$ on IHM). For Qwen, where raw confidence is uninformative ($A(c, z) \approx 0.500$), few-shot adaptation steadily elevates correctness AUROC to $0.61$–$0.82$ across all tasks. While Few-Shot and Chain-of-Thought prompting slightly increase base prediction accuracy, they do not alleviate the underlying confidence miscalibration, our calibrator achieves comparable few-shot gains across all three prompting conditions. Source-task pretraining is what makes small $k$ values viable. Its contribution shrinks by a factor of three to five as target labels accumulate: $+0.101$ mean AUROC at $k = 10$ against $+0.032$ at $k = 200$ for Gemini, $+0.164$ against $+0.036$ for GPT-5.5, and $+0.098$ against $+0.026$ for Qwen2.5-7B. On Gemini (ZS), 10 pretrained examples exceed 200 target-only examples on four of nine tasks. Full sweeps across $k \in \{10, 50, 100, 200\}$ for all three models appear in Appendix J.3.

## 6 Discussion and Limitations

Verbalized confidence is the signal a practitioner has immediate access to, and it does not work. Across three models and nine tasks, self-knowledge AUROC never exceeds $0.837$ (in ZS) and falls below chance on AKI for both larger models ($0.447$, $0.462$). Qwen2.5-7B illustrates the limiting case, its confidence values reorder nothing relative to the binary prediction, yet its classifier AUROC reaches $0.806$. The gap between what the commonly reported metric shows and what the model actually knows about its own correctness is the gap this work addresses. A calibrator trained on eight tasks and applied to a ninth works on few tasks but reaches $0.432$ on in-hospital mortality, below chance. The mapping from calibration features to correctness differs across tasks and inverts on at least one, and such a shift is not identifiable from unlabeled target data (Ben-David et al., 2010). Target labels are necessary in this setting as it motivates few-shot adaptation to identify the shift. We work with binary outcomes, which makes the signing of Eq. 1 well-defined and the two metrics directly comparable on the same score. The calibrator's target $z = \mathbf{1}[\hat{y} = y]$ is binary however many classes a task has, so its intuition and construction extends without modification but the classifier metric does not, since with $K > 2$ there is no single signed score (Appendix L).

The mutual information reference is estimated on labeled task data, so the features are not strictly label-free even under LOTO. Under stratified subsamples of 200, the top-ranked feature is reli-

ably recovered (on at least $4/5$ seeds) and the calibrator AUROC fluctuates by at most $0.007$ (Appendix G). The same $200$ labels can be used to construct the reference and adapt the calibrator. For LOTO, the reference ordering need not come from the data. Discussions with clinicians noted that trained models usually do not agree with the clinician's feature ordering, so an expert specified reference ranking would remove the labeling requirement and the calibrator would be aligned on how a domain expert would want. This changes what divergence measures, from disagreement with the empirical feature-outcome relationship to disagreement with clinical reasoning. These are not the same, calibrators built on each would plausibly flag different cases as unreliable. Which is more useful in deployment is a question we have not tested. Furthermore, because dedicated classifiers are better (Table 1), training a fully supervised tabular model is more accurate. Our calibrator serves a different purpose and its utility is therefore explicitly restricted to settings where LLMs are utilized for their generative ability (see Appendix K for a detailed discussion). Finally, we did not test varying the number of LLM-elicited features, prompt perturbations and the transfer of calibrator's intuition to other data modalities remains as future work. Sampled cohorts should not be read as estimates of deployed performance, though the comparison between metrics stays unaffected.

We have worked with only three models, one per family, each emitting confidence differently, whether these findings transfer across or within families is untested. The tasks are all clinical, few in number (9), and drawn from emergency department or intensive care settings. Nothing in the calibrator's construction is domain-specific, the features compare a reported attribution ranking against a mutual-information reference, and both are defined for any tabular prediction task where the model reports which features drove its answer. Two tasks, hospitalization and in-hospital mortality, show no significant improvement from source-task pretraining and negative gain at low $k$, and what distinguishes them from the seven where the method works is not explored. From signals of token-level distribution in open models, SHAP places `logit_p1` as the calibrator's most influential feature for Qwen, above every attribution and task descriptor (Appendix J.1), suggesting the distribution the model computes may carry correctness information not its verbalized confidence. Whether this holds for the hosted models is not tested and our findings make it worth exploring.

## 7 Conclusion and Future Work

In this work, we show that verbalized confidence, commonly used for evaluation, carries little to no information about actual correctness. Illustrating this degenerate limit, a model whose confidence values do not reorder instances relative to its prediction can still attain a classifier AUROC of $0.806$, while against the quantity that matters for deployment this signal never exceeds $0.837$ (ZS) across three models and nine clinical tasks, and it falls below chance on complex tasks like AKI for both frontier models. We address this problem with an external calibrator built from attribution divergence, rank aggregation, and task descriptor features, requiring only two inference calls and no access to model internals. While zero-shot transfer across tasks fails under concept shift, providing ten to two hundred labeled target examples allows the calibrator to recover substantial signal, improving over verbalized confidence on all nine tasks (Gemini, zero-shot) and gaining most where confidence fails most severely. Source-task pretraining is what makes these small budgets viable, contributing $3$ to $5$ times more AUROC at $10$ labeled examples than at $200$ across all three models.

**Future work.** Including addressing some of the limitations mentioned in Sec 6, most direct extension is more of both language models and tasks, to test whether these patterns hold within and across families, and non-clinical tabular tasks, which would separate whether the pattern reflects clinical prediction specifically or how models verbalize confidence on structured inputs generally. On the calibrator itself, regularized linear discriminants may prove more label-efficient at very small $k$ values, while neural calibrators would enable architectures that model multiple tasks explicitly, hard or soft parameter sharing with per-task heads, or mixture-of-experts routing, would allow shared representations to be regularized by all tasks while task-specific components learn idiosyncratic mappings. Further, active selection of optimal $k$ target examples would intelligently reduce the labeling budget further. Finally, our calibrator estimates reliability without altering it. Trained across a wider range of tasks, its output could correct confidence at inference, for open-weight models, calibrated correctness estimates serve as a reinforcement signal for self-knowledge requiring no annotation beyond observed outcomes. For APIs, only prompt-level intervention is viable, and whether that reliably grounds verbalized confidence remains untested.

## ETHICS STATEMENT

This study utilizes de-identified EHR data (MIMIC-IV, MIMIC-IV-ED, MC-MED) under strict PhysioNet Data Use Agreements requiring credentialed access and human subjects research training. We conducted no patient contact or re-identification and do not redistribute data, reproducibility steps for credentialed researchers are detailed in Reproducibility Statement. Proprietary API queries utilized zero-data-retention policies, though real-world deployment of such APIs necessitates institutional privacy review. Our calibrator is designed as alerts to trigger human review, not a final decision-maker. It still makes errors and performs poorly on complex tasks unless it is adapted using labeled examples from that specific task. Finally, while serialized records include demographic attributes (e.g., age, race, sex, insurance), we did not evaluate calibration disparities across these groups, given documented biases in clinical LLMs (Brown et al., 2025), rigorous fairness auditing remains a strict prerequisite for deployment.

## REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our findings, we detail the software, compute infrastructure, and data access procedures utilized in this work. Exact prompt templates, output schemas, and fixed XGBoost hyperparameters are fully documented in Appendices B and D, respectively.

**Code Availability.** To strictly maintain double-blind anonymity during the review process, the complete, anonymized source code (including all Jupyter notebooks for inference and calibrator training) required to execute all experiments is included as a `.zip` file in the supplementary materials. Due to the 100MB file-size limit for supplementary materials, the individual task-level raw model output folders have been excluded and the files calibrator data files have been heavily compressed. To directly reproduce the calibrator experiments, researchers can use the master datasets provided in the `model_results/calibrator_data/` directory. Upon acceptance, the full codebase and all raw inference outputs will be open-sourced via a public repository.

**Data Access and Reproducibility.** The underlying datasets utilized in this study are derived from MIMIC-IV, MIMIC-IV-ED and MC-MED. Due to strict credentialing requirements (e.g., CITI training) and PhysioNet Data Use Agreements, we are strictly prohibited from redistributing the preprocessed, task-specific data. Therefore, to reproduce our findings, researchers must follow a two-step protocol: (1) independently obtain PhysioNet access and run the standard cohort construction scripts from the referenced benchmark to extract the task data, and (2) execute our provided codebase on top of those resulting datasets to run the LLM inference and calibrator training pipelines.

**Compute Infrastructure.** Closed-weight proprietary models (Gemini 3 Flash Preview and GPT-5.5) were queried via the OpenRouter API with zero-data retention. Because exact token logits were inaccessible for these endpoints, our pipeline relies exclusively on parsed verbalized JSON outputs. Open-weight local models (e.g., Qwen2.5-7B-Instruct) were hosted and evaluated using cloud compute, utilizing AWS SageMaker with A40 GPUs and Google Colab Pro nodes equipped with NVIDIA A100 GPUs. The external XGBoost calibrator ($g_\theta$) is computationally lightweight and requires no heavy GPU acceleration. Calibrator training, feature extraction, and few-shot adaptation experiments were executed on standard CPU environments.

**Software Versions.** The external calibrator and feature extraction pipelines were implemented in Python 3.10. Core dependencies for reproducing the results include `numpy`, `pandas`, `scipy`, `pingouin`, `scikit-learn`, `xgboost`, `shap`, and `sentence-transformers`.

## AI USE STATEMENT

Because this work empirically investigates large language models (LLMs), the specific models, inference procedures, and prompting conditions evaluated as subjects of this study are described separately in Sections 4 and 5, and Appendix B. The following disclosure applies exclusively to the use of generative AI as a supplementary tool during the research and writing process.

The research problem, metric decomposition (Section 3.2), calibrator design, and overall experimental program were conceived, developed, and executed entirely by the authors. During the research process, generative AI tools were used to generate pseudocode for portions of the feature-extraction pipeline. All AI-generated code was thoroughly reviewed, rewritten, and corrected by the authors, and its correctness was verified. Furthermore, every finding reported in this manuscript was independently derived and verified by the authors. AI was strictly **not** used to generate synthetic data, formulate mathematical proofs, assist with translation, or conduct cohort construction and data cleaning (which followed the published benchmark pipelines cited in Section 4.1).

During manuscript preparation, AI served as a guide, suggesting structural reorganizations, and refining prose for readability. All AI-drafted text was heavily edited or entirely rewritten by the authors. No references were included on the basis of a model output alone. The authors have rigorously reviewed all AI-assisted content and assume full responsibility for the final content, claims, and artifacts presented in this work.

## REFERENCES


Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. URL `http://www.springerlink.com/content/q6qk230685577n52/`.

Sebastian Daniel Boie, Florian Reis, Nicolas Frey, Elias Grünewald, and Felix Balzer. Calibration of self-reported confidence and accuracy of large language models in medical question answering. *Journal of Medical Systems*, 50(1):103, 2026.

Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5):811–822, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Akshat Dasula, Prasanna Desikan, and Jaideep Srivastava. Llm doesn't know what it doesn't know: Detecting epistemic blind spots via cross-model attribution divergence on clinical tabular data. *arXiv preprint arXiv:2606.19509*, 2026.

Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pp. 613–622, 2001.

fengx13. MC-MED benchmark: Python workflow for generating benchmark datasets and machine learning models from the MC-MED database. `https://github.com/fengx13/mc-med-benchmark`, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pp. 5549–5581. PMLR, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do llms estimate uncertainty well in instruction-following? *arXiv preprint arXiv:2410.14582*, 2024.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, January 2023. doi: 10.13026/5ntk-km72. URL https://doi.org/10.13026/5ntk-km72. Version 2.2.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Aman Kansal, Emma Chen, Tom Jin, Pranav Rajpurkar, and David Kim. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). *PhysioNet*, September 2025. doi: 10.13026/wvyw-g663. URL https://doi.org/10.13026/wvyw-g663. Version 1.0.1.

John G. Kemeny. Mathematics without numbers. 88(4):577–591, Fall 1959. ISSN 0011-5266 (print), 1548-6192 (electronic). doi: https://doi.org/10.2307/20026529. URL http://www.jstor.org/stable/20026529.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ernesto Quevedo, Jorge Yero Salazar, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *World Congress in Computer Science, Computer Engineering & Applied Computing*, pp. 154–173. Springer, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Brian C. Ross. Mutual information between discrete and continuous data sets. *PLoS ONE*, 9, 2014. URL https://api.semanticscholar.org/CorpusID:8950163.

Chandrima Sarkar, Sarah Cooley, and Jaideep Srivastava. Robust feature selection technique using rank aggregation. *Applied Artificial Intelligence*, 28(3):243–257, 2014.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. *arXiv preprint arXiv:2403.08819*, 2024.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86 2:420–8, 1979. URL https://api.semanticscholar.org/CorpusID:13168820.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020b.

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

Lijuan Wu, Yong Hu, Borong Yuan, Xiangzhou Zhang, Weiqi Chen, Kang Liu, and Mei Liu. Which risk predictors are more likely to indicate severe aki in hospitalized patients? *International journal of medical informatics*, 143:104270, 2020. URL `https://api.semanticscholar.org/CorpusID:221863262`.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, 2022.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Yusuf Yildiz, Goran Nenadic, Meghna Jani, and David A Jenkins. Will large language models transform clinical prediction? *Diagnostic and Prognostic Research*, 9(1):1–7, 2025.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

# APPENDIX

## A TASK DEFINITIONS AND COHORT CONSTRUCTION

The clinical prediction tasks and corresponding patient cohorts evaluated in this study are directly adopted from established electronic health record (EHR) benchmarks utilizing the MIMIC-IV, MIMIC-IV-ED and MC-MED databases mentioned in Sec 4.1.

The benchmark encompasses 9 distinct clinical prediction tasks:

**In-Hospital Mortality (IHM):** A task predicting whether a patient, admitted to the ICU with acute heart failure, will die during their current hospital admission, utilizing data available within the first 24 hours.

**ED Mortality:** Predicting whether a patient will expire while in the Emergency Department or during the immediate subsequent hospitalization.

**ICU Transfer (12h):** Predicting whether a patient currently in the ED or general ward will be transferred to ICU within the next 12 hours.

**Hospitalization (Admission):** Predicting whether a patient presenting to the Emergency Department will require inpatient admission versus being discharged.

**Revisit (3d):** Predicting whether a patient discharged from the Emergency Department will have an unplanned return to the ED within 3 days (72 hours).

**Critical Outcome:** Predicting severe patient deterioration, typically defined as either in-hospital mortality or the need for direct ICU transfer.

**Acute Kidney Injury (AKI):** Predicting the onset of Acute Kidney Injury during the patient's encounter, formulated according to the Kidney Disease: Improving Global Outcomes (KDIGO) criteria utilizing sequential serum creatinine, labs, vitals, comorbidities, etc. from the previous 24hr measurements.

**Pulmonary Embolism (PE):** Predicting the clinical diagnosis of a pulmonary embolism based on patient presentation, vitals, and structured EHR data.

**ACS / MI:** Predicting the presence or subsequent diagnosis of Acute Coronary Syndrome (ACS) or Myocardial Infarction (MI) during the clinical encounter.

**Data Processing and Structuring.** We utilized the benchmark's established data extraction and preprocessing pipelines to ensure standardized cohort construction. Prior to LLM inference, we applied a reverse-mapping step to decode the preprocessed, encoded columns back into clinically valid attributes which substantially reduced the feature dimensionality. This allowed for a more token-efficient prompt construction without any loss of the original patient data.

## B Prompt Templates and Output Schema

To avoid redundancy, we present the parameterized master templates used across our experiments. The templates share a common skeleton and are instantiated for each specific task by substituting the clinical setting, task description, and outcome labels.

Templates are identical across all three evaluated language models.

**Output Schema.** To ensure parsed predictions, confidence values, and feature attributions could be systematically extracted without relying on regex heuristics, all models were strictly instructed to return their responses conforming to the following JSON schema:

```
{
    "prediction": 1/0,
    "confidence": 0.85,
    "feature_ranking": [
        {"rank": 1, "feature": "exact_feature_name", "direction": "[POS_LABEL]"},
        {"rank": 2, "feature": "exact_feature_name", "direction": "[NEG_LABEL]"},
        {"rank": 3, "feature": "exact_feature_name", "direction": "[POS_LABEL]"},
        {"rank": 4, "feature": "exact_feature_name", "direction": "[POS_LABEL]"},
        {"rank": 5, "feature": "exact_feature_name", "direction": "[NEG_LABEL]"}
    ]
}
```

**Zero-Shot (ZS) Template.**

```
You are a clinical decision support system analyzing [SETTING] data.

[TASK_DESCRIPTION]

Each feature is a clinical measurement, demographic characteristic, or
clinical history variable. Binary features (0 or 1) indicate absence or presence.

PATIENT DATA:
{patient_text}

INSTRUCTIONS:
1. Predict outcome: 1 ([POS_OUTCOME_PHRASE]) or 0 ([NEG_OUTCOME_PHRASE])
2. Confidence: a score from 0.0 to 1.0
3. Rank the TOP {top_ft} features from the patient data above that MOST
```

```
   influenced your prediction, in order from most to least important
4. For each feature, state which outcome it supports | either "[POS_LABEL]"
   or "[NEG_LABEL]"
5. Use the EXACT feature names as they appear in the patient data above

Respond ONLY in this JSON format, no other text:
[JSON_SCHEMA]
```

**Few-Shot (FS) Template.** The Few-Shot template modifies the Zero-Shot prompt by prepending four class-stratified exemplars immediately before the target patient data. These exemplars were drawn once per task and held fixed.

```
You are a clinical decision support system analyzing [SETTING] data.

[TASK_DESCRIPTION]

Each feature is a clinical measurement, demographic characteristic, or
clinical history variable. Binary features (0 or 1) indicate absence or presence.

Few examples for this task:
{example_section}

PATIENT DATA:
{patient_text}

[INSTRUCTIONS AND JSON SCHEMA IDENTICAL TO ZERO-SHOT]
```

**Chain-of-Thought (CoT) Template.** The CoT condition replaces the standard instruction block with a structured reasoning framework, forcing the model to articulate its clinical logic prior to emitting the final prediction.

```
You are a clinical decision support system analyzing [SETTING] data.

[TASK_DESCRIPTION]

Each feature is a clinical measurement, demographic characteristic, or
clinical history variable. Binary features (0 or 1) indicate absence or presence.

PATIENT DATA:
{patient_text}

Reason through this case using the following structured reasoning approach:

Step 1: EVIDENCE IDENTIFICATION
Identify the {top_ft} most clinically relevant features from the patient data above.

Step 2: PER-FEATURE REASONING
For each feature, briefly explain what it suggests about the likelihood of
[POS_LABEL] versus [NEG_LABEL]. Be concise | one sentence per feature.

Step 3: WEIGHING
Compare the evidence supporting [POS_LABEL] versus alternative explanations.
Which side is stronger and why?

Step 4: FINAL PREDICTION
Based on the reasoning above, provide your final prediction, confidence,
and feature ranking.

Respond ONLY in this JSON format, no other text (keep the reasoning field
concise | 3-5 sentences maximum):
{
    "reasoning": "Step 1: Key features are... Step 2: ... Step 3: ... Step 4: ...",
    "prediction": 1/0,
    "confidence": 0.85,
```

```
    "feature_ranking": [ ... ]
}
```

**Self-Assessment $P(true)$ Template.** Issued as a distinct, secondary inference call after the primary prediction, conditioned on the model's own output. The original prediction is not revised based on this.

```
You previously reviewed this patient's data:

PATIENT DATA:
{patient_text}

You predicted: {prediction_text} with confidence {confidence}.

{reasoning_section}

Now consider whether that prediction was correct. Given the patient's features,
what is the probability from 0.0 to 1.0 that your prediction is correct?

Respond ONLY in this JSON format, no other text:
{"p_true": 0.75}
```

**Serialized Record.** Below is an example of how a structured EHR row is serialized into plain text ({patient_text}) prior to prompt injection. For all tasks original clinical terms are preserved rather than utilizing encoded column identifiers, enabling accurate alignment between the model's textual attribution and the data-derived empirical reference.

```
- gender: M
- race: WHITE
- arrival_transport: AMBULANCE
- ethnicity: WHITE
- insurance: Medicare
- age: 51.0
- triage_temperature: 36.66666666666667
- triage_heartrate: 84.0
- triage_resprate: 18.0
- triage_o2sat: 99.0
- triage_sbp: 133.0
- triage_dbp: 77.0
- triage_pain: 0
- triage_acuity: 3.0
- ed_temperature_last: 36.5
- ed_heartrate_last: 83.0
...
- score_MEWS: 1
- charlson_conditions: None
- elixhauser_conditions: None
- chief_complaint: FOR EVAL
```

# C Feature Definitions

The external calibrator $g_\theta$ utilizes a base feature vector $\phi_i \in \mathbb{R}^{38}$. These 38 features are extracted purely from the model's textual outputs and the task's empirical Mutual Information (MI) distribution, requiring no model internals. Let $K$ denote the number of task-features returned by the LLM ($K = 5$ in our experiments), $\pi_i$ denote the LLM's reported feature ranking for instance $i$, and $\rho_t$ denote the empirical MI reference ranking for task $t$.

The features are grouped into four distinct families:

## C.1 Model Outputs (3 Features)

These features capture the direct numerical outputs generated by the language model:

- **llm_pred** ($\hat{y}_i$): The prediction emitted by the model.
- **llm_confidence** ($c_i$): The verbalized confidence probability.
- **ptrue** ($p_i^{\text{true}}$): The elicited correctness probability obtained via the secondary self-assessment prompt.

### C.2 Rank aggregation features (6)

Both $\pi_i$ and the reference top-$K$ of $\rho_t$ are treated as rankings over a fixed domain: the reference top-$K$ feature set. Features in that domain that the model did not select receive a penalty rank of $K+1$. Model selections outside the reference top-$K$ do not enter, so the reference defines the comparison domain.

- **icc_llm_mi**: ICC(A,1) between the two rank vectors over the reference top-$K$ domain, treating features as targets and the two rankings as raters.
- **borda_dev_llm**: one minus the Jaccard similarity between the model's top-$K$ set and the top-$K$ set of the Borda consensus of the two rankings.
- **kemeny_dev_llm**: the same set-level distance computed against the Kemeny consensus, which is computationally tractable and equivalent to the median for $M = 2$ rankers.
- **kemeny_inv_llm**: the fraction of feature pairs on which the model's ranking is discordant with the Kemeny consensus,

$$\frac{1}{\binom{|F|}{2}} \sum_{u \neq v} \mathbb{I}[r_{\text{llm}}(u) < r_{\text{llm}}(v) \ \wedge \ r_{\text{kem}}(u) > r_{\text{kem}}(v)] \,,$$

  where $F$ is the shared feature domain.
- **concentration_llm**: Fraction of reciprocal rank mass captured by the top-$K$ features within the comparison domain. It is computed as the sum of the reciprocal ranks ($1/r$) of the highest-ranked features divided by the total reciprocal rank mass of the domain.
- **flatness_llm**: Normalized Shannon entropy of the model's reciprocal rank distribution over the comparison domain. It evaluates $-\sum(p \ln p)/\ln(|F|)$, where the probability mass $p_u$ for a feature $u$ is proportional to its reciprocal rank $1/r_{\text{llm}}(u)$. This measures how uniformly the LLM's assigned importance is distributed across the target domain.

### C.3 Attribution divergence features (15)

These compare the model's selected features $\pi_i$ against the reference ordering $\rho_t$ by set overlap, rank correlation, and position.

- **mi_kendall_tau**: Kendall's $\tau$ between the two rankings over the *union* $\pi_i \cup \text{Top}_K(\rho_t)$, with features absent from either ranking assigned rank $K+1$.
- **mi_ads**: $1 - \tau$, the attribution divergence score. A monotone transform of `mi_kendall_tau`; both are retained for interpretability.
- **mi_jaccard**: $|\pi_i \cap \text{Top}_K(\rho_t)| \,/\, |\pi_i \cup \text{Top}_K(\rho_t)|$.
- **mi_top1_match**: $\mathbb{I}[\pi_{i,1} = \rho_{t,1}]$.
- **mi_coverage**: $|\pi_i \cap \text{Top}_K(\rho_t)| \,/K$, the fraction of the reference top-$K$ recovered.
- **mi_avg_shift**: mean absolute displacement between each selected feature's position in $\pi_i$ and its position in the global ordering $\rho_t$, with features absent from $\rho_t$ assigned the maximum possible shift.
- **mi_mean_score**, **mi_top1_score**: mean and top-ranked mutual information value among the model's selections.
- **mi_mean_global_rank**, **mi_top1_global_rank**: mean and top-ranked position of the model's selections within $\rho_t$.
- **mi_rho_zone**: Spearman correlation between zone indices, where each feature's zone is its rank divided by the zone width $\max(K, |F|/5)$. Measures whether the model preserves coarse ordering when fine ordering is noisy.

- **mi_rho_soft**: Spearman correlation between the model's rank positions and the negated mutual information values of those same features, rewarding orderings that track MI magnitude rather than MI rank.
- **mi_zscore**: $(\bar{m}_i - \mu)/\sigma$, where $\bar{m}_i$ is the mean MI value of the model's selections and $\mu, \sigma$ are the mean and standard deviation of the task's full MI distribution.
- **mi_percentile**: mean percentile of the model's selections within the task's MI distribution; $0.5$ corresponds to random selection.
- **n_llm_features**: number of valid parseable features returned, at most $K$.

### C.4 Task Descriptors (14 Features)

Because the concentration of predictive signal varies heavily across clinical tasks, the calibrator is conditioned on 14 task-level constants derived from $\rho_t$. These remain static across all patients within a given task $t$:

- **task_mi_top1_value** ... **task_mi_top5_value**: The raw MI scores of the top 5 most predictive features in the dataset.
- **task_mi_gap_1_2** & **task_mi_gap_1_5**: The signal drop-off gaps between the 1st/2nd and 1st/5th ranked features.
- **task_mi_top5_sum** & **task_mi_top5_share**: The total predictive mass of the top 5 features and their fractional share of the dataset's total mutual information.
- **task_mi_max_value**, **task_mi_mean_value**, **task_mi_std_value**: The maximum, mean, and standard deviation of the task's MI distribution.
- **task_mi_n_above_0.05** & **task_mi_n_above_0.1**: The discrete count of features possessing an MI score exceeding fixed thresholds of $0.05$ and $0.1$, serving as proxies for task complexity and feature sparsity.

### C.5 Logit-derived features (open-weight model only)

For Qwen2.5-7B, run locally, we record the next-token distribution at two positions during generation. Neither is available for the two hosted models, so these features are used exclusively in the Qwen2.5-7B calibrator configurations, expanding its base feature space from $d_\phi = 38$ to $d_\phi = 40$.

- **logit_p1**: the model's probability of the positive class, read at the token position where the `prediction` field is emitted. We restrict the vocabulary distribution to the single-token representations of `0` and `1` and renormalize:

$$P(\hat{y} = 1) = \frac{\exp(\ell_1)}{\exp(\ell_0) + \exp(\ell_1)},$$

  where $\ell_0, \ell_1$ are the logits at that position. This is the conditional probability given that the model emits one of the two label tokens; the residual mass on other tokens is negligible under the JSON schema, which constrains the output at that position.
- **logit_ptrue**: the analogous quantity for self-assessment. The self-assessment prompt is appended with `"Answer:"` and the next-token distribution is restricted to the single-token forms of affirmation and negation, taking the first candidate from {`" True"`, `" Yes"`, `"True"`, `"Yes"`} and {`" False"`, `" No"`, `"False"`, `"No"`} that the tokenizer encodes as one token.

These are scalar quantities computed from the same forward pass as the prediction, requiring no additional inference. They let us ask whether the failures we document are specific to what the model verbalizes or are present in the distribution it computes.

### C.6 Chain-of-thought reasoning embeddings

Under chain-of-thought prompting the model returns a `reasoning` field alongside the structured output. We encode this text with the local `all-MiniLM-L6-v2` encoder (Wang et al., 2020b),

yielding 384 dimensions per patient, and append them to the calibrator's feature set for the chain-of-thought condition only (expanding the space to $d_\phi = 422$ for closed-weight and $d_\phi = 424$ for open-weight models).

The inclusion of these semantic embeddings did not significantly alter the calibrator's AUROC compared to the structured Zero-Shot and Few-Shot representations. Because they increased feature dimensionality by an order of magnitude, required an additional local encoding step, and did not alleviate the underlying confidence miscalibration, our primary external calibration results rely strictly on the lightweight structured feature space.

## D Calibrator Configuration

**XGBoost Hyperparameters.** The external calibrator $g_\theta$ utilizes an XGBoost classifier with hyperparameters held constant across all tasks, models, and prompting conditions to ensure it remains domain-agnostic. The configuration is as follows:

- `objective`: `binary:logistic`
- `eval_metric`: `logloss`
- `tree_method`: `hist`
- `n_estimators`: 300
- `max_depth`: 5
- `learning_rate`: 0.05
- `subsample`: 0.8
- `colsample_bytree`: 0.8
- `reg_lambda`: 1.0

**Seed Handling and Support-Set Sampling.** For all few-shot adaptation experiments, we evaluate across $k \in \{0, 10, 50, 100, 200\}$ labeled target examples. At each $k > 0$, we perform 5 independent random draws for the support set ($N_{\text{draws}} = 5$). To ensure reproducibility, these draws are initialized with a global random seed (`SEED = 42`). The subset is drawn using class-stratified random sampling to ensure the target label distribution in the few-shot support set accurately reflects the broader task prevalence.

## E Confidence Emission Statistics

This section details the global distribution of verbalized confidence ($c$) and elicited self-assessment probability ($p^{\text{true}}$) across all nine clinical tasks.

As discussed in Section 5.2, confidence remains uniformly high and granularity remains restricted regardless of the underlying task difficulty. Tables 4, 5, and 6 present the distinct values, observed ranges, and statistical averages for each model under Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT) prompting conditions.

**Gemini 3 Flash Preview.** Gemini remains restricted to exactly 14 distinct values across all three prompting conditions, with median confidence remaining at or above 0.82 regardless of the prompt.

Table 4: Global confidence statistics for Gemini across prompting styles.

| Prompt | Distinct Values | Observed Range | Mean ($c, p^{\text{true}}$) | Median ($c, p^{\text{true}}$) |
|---|---|---|---|---|
| ZS | 14 | 0.65–0.99 | 0.82, 0.84 | 0.82, 0.88 |
| FS | 14 | 0.65–0.99 | 0.84, 0.78 | 0.85, 0.85 |
| CoT | 14 | 0.65–0.99 | 0.86, 0.85 | 0.85, 0.85 |

**GPT-5.5.** GPT-5.5 produces the highest granularity among the tested models, utilizing over 40 distinct values. However, its median verbalized confidence remains consistently high (0.74–0.78), heavily overstating its reliability on complex tasks.

Table 5: Global confidence statistics for GPT-5.5 across prompting styles.

| Prompt | Distinct Values | Observed Range | Mean $(c, p^{\text{true}})$ | Median $(c, p^{\text{true}})$ |
|---|---|---|---|---|
| ZS | 42 | 0.56–0.99 | 0.78, 0.78 | 0.78, 0.82 |
| FS | 43 | 0.56–0.99 | 0.76, 0.71 | 0.74, 0.78 |
| CoT | 44 | 0.56–0.99 | 0.77, 0.76 | 0.78, 0.78 |

**Qwen2.5-7B.** Qwen exhibits severe mode collapse, particularly under Zero-Shot and Few-Shot conditions, where it emits only 3 distinct confidence values. While Chain-of-Thought prompting slightly expands its expressiveness to 10 distinct values and lowers the minimum bound to 0.35, the median confidence and self-assessment remain statically anchored at 0.85 and 0.75, respectively.

Table 6: Global confidence statistics for Qwen2.5-7B across prompting styles.

| Prompt | Distinct Values | Observed Range | Mean $(c, p^{\text{true}})$ | Median $(c, p^{\text{true}})$ |
|---|---|---|---|---|
| ZS | 3 | 0.75–0.95 | 0.85, 0.75 | 0.85, 0.75 |
| FS | 3 | 0.75–0.92 | 0.85, 0.75 | 0.85, 0.75 |
| CoT | 10 | 0.35–0.95 | 0.84, 0.75 | 0.85, 0.75 |

# F Class-Conditioned Confidence Emission Statistics

This section provides the complete breakdown of verbalized confidence ($c_i$) partitioned by the predicted class ($\hat{y}_i = 1$ vs. $\hat{y}_i = 0$) across all nine clinical prediction tasks and three prompting conditions (Zero-Shot, Few-Shot, and Chain-of-Thought). Values are reported as **Median (Mean)**.

## F.1 Gemini 3 Flash Preview

Table 7 details the confidence distributions for Gemini across prompting conditions. Across most emergency and diagnostic tasks, predicted negative instances ($\hat{y} = 0$) systematically receive higher baseline confidence (median $0.85$–$0.92$) than predicted positive instances ($\hat{y} = 1$, median $0.72$–$0.85$), demonstrating persistent class-conditional anchoring across prompting styles.

Table 7: Gemini verbalized confidence partitioned by predicted class across prompting conditions.

| | Zero-Shot (ZS) | | Few-Shot (FS) | | Chain-of-Thought (CoT) | |
|---|---|---|---|---|---|---|
| **Task** | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ |
| Acute Kidney Injury | 0.78 (0.81) | 0.82 (0.81) | 0.85 (0.84) | 0.85 (0.84) | 0.78 (0.81) | 0.82 (0.81) |
| In-Hospital Mortality | 0.82 (0.84) | 0.82 (0.82) | 0.85 (0.83) | 0.85 (0.82) | 0.82 (0.84) | 0.82 (0.82) |
| Hospitalization | 0.82 (0.84) | 0.88 (0.87) | 0.88 (0.86) | 0.85 (0.83) | 0.85 (0.86) | 0.92 (0.87) |
| Critical Outcome | 0.78 (0.81) | 0.92 (0.91) | 0.82 (0.83) | 0.92 (0.88) | 0.82 (0.82) | 0.92 (0.91) |
| ED Mortality | 0.78 (0.78) | 0.92 (0.92) | 0.75 (0.78) | 0.92 (0.90) | 0.78 (0.79) | 0.92 (0.92) |
| ICU Transfer (12h) | 0.78 (0.81) | 0.92 (0.91) | 0.82 (0.82) | 0.92 (0.89) | 0.82 (0.83) | 0.92 (0.91) |
| Revisit (3d) | 0.72 (0.77) | 0.88 (0.87) | 0.75 (0.78) | 0.75 (0.76) | 0.78 (0.80) | 0.85 (0.86) |
| Pulmonary Embolism | 0.78 (0.77) | 0.88 (0.87) | 0.78 (0.80) | 0.85 (0.83) | 0.82 (0.80) | 0.92 (0.87) |
| ACS/MI | 0.92 (0.91) | 0.92 (0.89) | 0.88 (0.87) | 0.85 (0.85) | 0.95 (0.91) | 0.92 (0.88) |

## F.2 GPT-5.5

Table 8 details the confidence distributions for GPT-5.5. Similar to Gemini, GPT-5.5 assigns higher confidence to predicted negatives in acute triage tasks (e.g., ED Mortality: $0.88$ vs. $0.72$), but displays wider within-class dispersion.

Table 8: GPT-5.5 verbalized confidence partitioned by predicted class across prompting conditions.

| | Zero-Shot (ZS) | | Few-Shot (FS) | | Chain-of-Thought (CoT) | |
|---|---|---|---|---|---|---|
| **Task** | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ |
| Acute Kidney Injury | 0.74 (0.78) | 0.72 (0.72) | 0.74 (0.78) | 0.72 (0.72) | 0.68 (0.74) | 0.68 (0.67) |
| In-Hospital Mortality | 0.74 (0.76) | 0.70 (0.71) | 0.78 (0.77) | 0.68 (0.71) | 0.78 (0.77) | 0.70 (0.72) |
| Hospitalization | 0.78 (0.77) | 0.78 (0.79) | 0.82 (0.79) | 0.78 (0.76) | 0.74 (0.75) | 0.78 (0.79) |
| Critical Outcome | 0.76 (0.76) | 0.84 (0.82) | 0.78 (0.78) | 0.78 (0.80) | 0.74 (0.75) | 0.86 (0.83) |
| ED Mortality | 0.72 (0.73) | 0.88 (0.84) | 0.72 (0.72) | 0.88 (0.84) | 0.72 (0.73) | 0.88 (0.85) |
| ICU Transfer (12h) | 0.78 (0.77) | 0.88 (0.85) | 0.74 (0.75) | 0.82 (0.82) | 0.74 (0.76) | 0.88 (0.85) |
| Revisit (3d) | 0.68 (0.70) | 0.72 (0.72) | 0.68 (0.69) | 0.62 (0.64) | 0.68 (0.69) | 0.72 (0.73) |
| Pulmonary Embolism | 0.68 (0.68) | 0.78 (0.78) | 0.72 (0.70) | 0.72 (0.72) | 0.68 (0.68) | 0.78 (0.79) |
| ACS/MI | 0.86 (0.83) | 0.78 (0.80) | 0.90 (0.86) | 0.76 (0.77) | 0.82 (0.82) | 0.78 (0.79) |

### F.3 QWEN2.5-7B

Table 9 details the confidence distributions for Qwen2.5-7B. Qwen displays near-total mode collapse across Zero-Shot and Few-Shot conditions, returning a static $0.85$ confidence across virtually all tasks and classes.

Table 9: Qwen2.5-7B verbalized confidence partitioned by predicted class across prompting conditions.

| | Zero-Shot (ZS) | | Few-Shot (FS) | | Chain-of-Thought (CoT) | |
|---|---|---|---|---|---|---|
| **Task** | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ | $\hat{y}=1$ | $\hat{y}=0$ |
| Acute Kidney Injury | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.84) | 0.70 (0.66) |
| In-Hospital Mortality | 0.85 (0.85) | 0.85 (0.84) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) |
| Hospitalization | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) |
| Critical Outcome | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.83) |
| ED Mortality | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) |
| ICU Transfer (12h) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) |
| Revisit (3d) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.84) |
| Pulmonary Embolism | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.82) |
| ACS/MI | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.85) | 0.85 (0.80) | 0.85 (0.85) | 0.85 (0.82) |

## G MUTUAL INFORMATION REFERENCE STABILITY

In this section, we empirically validate that the reference feature ordering and downstream calibrator performance are stable under low-sample estimation ($m = 200$).

### G.1 EXPERIMENTAL SETUP AND EVALUATION METRICS

We evaluated three clinical tasks spanning distinct etiologies and feature dimensions: Acute Kidney Injury (ICU), Pulmonary Embolism (ED diagnostic), and Hospitalization.

For each task:

1. We computed the benchmark reference ranking $\rho_t^{\text{full}}$ using all available patient records ($N \approx 2000$).
2. We drew 5 independent, class-stratified subsamples of $m = 200$ patients using random seed 42 and computed subsampled reference rankings $\rho_t^{(s)}$ for $s \in \{1, \ldots, 5\}$.
3. We extracted attribution divergence features using both $\rho_t^{\text{full}}$ and $\rho_t^{(s)}$, subsequently training and evaluating the external calibrator under few-shot adaptation ($k = 200$).

To quantify ranking stability and downstream calibration fidelity, we measured three metrics across the 5 seeds:

- **Top-1 Feature Match Rate**: The proportion of seeds where $\arg\max_j \text{MI}^{(s)}(x^{(j)}, y) = \arg\max_j \text{MI}^{\text{full}}(x^{(j)}, y)$.
- **Top-10 Overlap:** The mean fraction of the full-cohort top-10 feature set recovered by the subsampled ordering:

$$\text{Overlap@10} = \frac{|\text{Top10}(\rho_t^{\text{full}}) \cap \text{Top10}(\rho_t^{(s)})|}{10}. \tag{2}$$

- **Calibrator AUROC**: The comparison of the 200-sample reference (AUROC($\rho_t^{(s)}$)) versus the full-cohort reference (AUROC($\rho_t^{\text{full}}$)).

### G.2 Stability Results

As summarized in Table 10, in all three tested tasks for GPT-5.5, the primary driver (Top-1 match) was recovered on all five seeds for acute kidney injury and hospitalization, and on four of five for pulmonary embolism. Top-10 overlap ranges from $0.64$ (PE) to $0.90$ (AKI), with PE showing the least stable ordering of the three.

Table 10: Stability of mutual information reference orderings and calibrator AUROC under 200-sample subsampling versus full-cohort estimation (mean $\pm$ std across 5 independent seeds).

| Task | Top-1 Match | Top-10 Overlap | Full Ref AUROC | 200-Sample Ref AUROC |
|---|---|---|---|---|
| Acute Kidney Injury (AKI) | 5/5 (100%) | $0.90 \pm 0.00$ | 0.828 | $0.828 \pm 0.001$ |
| Pulmonary Embolism (PE) | 4/5 (80%) | $0.64 \pm 0.09$ | 0.832 | $0.839 \pm 0.008$ |
| Hospitalization | 5/5 (100%) | $0.76 \pm 0.09$ | 0.729 | $0.724 \pm 0.008$ |

Calibrator AUROC evaluated on the 200-sample reference remains within $0.007$ AUROC of the full-cohort reference on all three tasks. This stability holds for the two other models tested in this work as well. This indicates that large datasets are not required to construct a usable reference ordering, though the ordering is not perfectly recovered on every task.

## H Cross-Model Metric Comparison

In Section 3.2, we established that verbalized confidence is routinely evaluated using two fundamentally conflated formulations:

1. **Classifier Metric ($A_{\text{clf}} = \textbf{AUROC}(s, y)$)**: Signed confidence $s_i$ evaluated against the true clinical outcome $y_i$:

$$s_i = \begin{cases} c_i & \text{if } \hat{y}_i = 1, \\ 1 - c_i & \text{if } \hat{y}_i = 0. \end{cases} \tag{3}$$

2. **Self-Knowledge Metric ($A_{\text{sk}} = \textbf{AUROC}(c, z)$)**: Raw verbalized confidence $c_i$ evaluated against instance-level model correctness $z_i = \mathbb{I}[\hat{y}_i = y_i]$.

This appendix provides the complete cross-model metric comparison tables across all 9 clinical tasks under Few-Shot (FS), and Chain-of-Thought (CoT) prompting conditions, as Zero-shot (ZS) was already discussed in Table 3. In each table, the evaluation gap is defined as Gap $= A(s, y) - A(c, z)$.

### H.1 Few-Shot Prompting

Table 11 displays metric divergence under Few-Shot (FS) in-context prompting (4 labeled exemplars). On complex tasks like Acute Kidney Injury and Pulmonary Embolism, few-shot prompting shifts the conditional distribution of confidence scores such that $A(c, z)$ inverts or changes direction relative to the Zero-Shot baseline without resolving the underlying miscalibration.

Table 11: Few-Shot (FS) metric comparison across models: Self-Knowledge $A(c, z)$ vs. Classifier Metric $A(s, y)$.

| | Gemini | | | GPT-5.5 | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | $A(c,z)$ | $A(s,y)$ | **Gap** | $A(c,z)$ | $A(s,y)$ | **Gap** | $A(c,z)$ | $A(s,y)$ | **Gap** |
| Revisit (3 d) | 0.603 | 0.584 | $-0.019$ | 0.553 | 0.576 | $+0.023$ | 0.500 | 0.500 | $+0.000$ |
| AKI | 0.507 | 0.718 | $+0.211$ | 0.462 | 0.635 | $+0.173$ | 0.500 | 0.500 | $+0.000$ |
| PE | 0.691 | 0.824 | $+0.133$ | 0.694 | 0.785 | $+0.091$ | 0.500 | 0.612 | $+0.112$ |
| IHM | 0.620 | 0.745 | $+0.125$ | 0.610 | 0.752 | $+0.142$ | 0.500 | 0.504 | $+0.004$ |
| Hospitalization | 0.646 | 0.781 | $+0.135$ | 0.671 | 0.812 | $+0.141$ | 0.500 | 0.584 | $+0.084$ |
| Critical Outcome | 0.768 | 0.852 | $+0.084$ | 0.768 | 0.871 | $+0.103$ | 0.500 | 0.563 | $+0.063$ |
| ED Mortality | 0.769 | 0.884 | $+0.115$ | 0.801 | 0.892 | $+0.091$ | 0.500 | 0.711 | $+0.211$ |
| ICU Transfer (12h) | 0.734 | 0.872 | $+0.138$ | 0.783 | 0.899 | $+0.116$ | 0.500 | 0.684 | $+0.184$ |
| ACS/MI | 0.812 | 0.910 | $+0.098$ | 0.806 | 0.908 | $+0.102$ | 0.438 | 0.762 | $+0.324$ |

### H.2 Chain-of-Thought (CoT) Prompting

Table 12 presents the evaluation gap under Chain-of-Thought prompting. Although this expands output expressiveness (e.g., Qwen produces up to 10 distinct values rather than 3), self-knowledge remains severely impaired, with $A(c, z)$ falling below chance on AKI for both frontier models ($0.447$ for Gemini, $0.453$ for GPT-5.5).

Table 12: Chain-of-Thought (CoT) metric comparison across models: Self-Knowledge $A(c, z)$ vs. Classifier Metric $A(s, y)$.

| | Gemini | | | GPT-5.5 | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | $A(c,z)$ | $A(s,y)$ | **Gap** | $A(c,z)$ | $A(s,y)$ | **Gap** | $A(c,z)$ | $A(s,y)$ | **Gap** |
| Revisit (3 d) | 0.608 | 0.621 | $+0.013$ | 0.566 | 0.612 | $+0.046$ | 0.504 | 0.518 | $+0.014$ |
| AKI | 0.447 | 0.615 | $+0.168$ | 0.453 | 0.604 | $+0.151$ | 0.538 | 0.524 | $-0.014$ |
| PE | 0.609 | 0.812 | $+0.203$ | 0.701 | 0.819 | $+0.118$ | 0.499 | 0.621 | $+0.122$ |
| IHM | 0.588 | 0.761 | $+0.173$ | 0.635 | 0.772 | $+0.137$ | 0.500 | 0.501 | $+0.001$ |
| Hospitalization | 0.690 | 0.804 | $+0.114$ | 0.714 | 0.822 | $+0.108$ | 0.502 | 0.641 | $+0.139$ |
| Critical Outcome | 0.767 | 0.865 | $+0.098$ | 0.780 | 0.879 | $+0.099$ | 0.490 | 0.638 | $+0.148$ |
| ED Mortality | 0.741 | 0.889 | $+0.148$ | 0.823 | 0.898 | $+0.075$ | 0.503 | 0.712 | $+0.209$ |
| ICU Transfer (12h) | 0.732 | 0.881 | $+0.149$ | 0.797 | 0.902 | $+0.105$ | 0.494 | 0.672 | $+0.178$ |
| ACS/MI | 0.718 | 0.892 | $+0.174$ | 0.846 | 0.915 | $+0.069$ | 0.501 | 0.751 | $+0.250$ |

## I Cross-Model Calibrator Results

This section provides the comprehensive performance results for the external XGBoost calibrator ($g_\theta$) across all evaluated language models, clinical tasks, and prompting conditions.

### I.1 Leave-One-Task-Out (LOTO) Generalization

We first evaluate the calibrator's ability to generalize to unseen tasks without any target-task supervision ($k = 0$). Tables 13, 14, and 15 report the LOTO AUROC and the absolute improvement over the model's raw verbalized confidence ($\Delta$ vs. Conf) for Gemini 3 Flash Preview, GPT-5.5, and Qwen2.5-7B across Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT) conditions. A positive $\Delta$ indicates the external calibrator successfully improved reliability estimation over the uncalibrated LLM baseline.

### I.2 Few-Shot Target Adaptation

To quantify the reliability of the calibrator when adapted to a specific clinical environment, we report the performance after incorporating $k \in \{10, 50, 100, 200\}$ labeled target-task examples into the LOTO training set.

Table 13: Zero-Shot (ZS) LOTO Generalization ($k = 0$). Values represent Calibrator AUROC and $\Delta$ vs. LLM Confidence.

| | Gemini | | GPT-5.5 | | Qwen2.5-7B | |
|---|---|---|---|---|---|---|
| **Task** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** |
| Acute Kidney Injury (AKI) | 0.511 | $+0.064$ | 0.475 | $+0.013$ | 0.364 | $-0.136$ |
| Pulmonary Embolism (PE) | 0.585 | $+0.035$ | 0.780 | $+0.097$ | 0.576 | $+0.076$ |
| In-Hospital Mortality (IHM) | 0.432 | $-0.156$ | 0.481 | $-0.139$ | 0.632 | $+0.134$ |
| Revisit (3d) | 0.640 | $-0.031$ | 0.548 | $-0.014$ | 0.524 | $+0.025$ |
| Hospitalization | 0.594 | $-0.117$ | 0.581 | $-0.131$ | 0.630 | $+0.130$ |
| Critical Outcome | 0.732 | $-0.016$ | 0.824 | $+0.033$ | 0.698 | $+0.198$ |
| ED Mortality | 0.744 | $+0.002$ | 0.882 | $+0.045$ | 0.747 | $+0.247$ |
| ICU Transfer (12h) | 0.749 | $+0.022$ | 0.851 | $+0.053$ | 0.724 | $+0.224$ |
| ACS/MI | 0.697 | $-0.048$ | 0.776 | $-0.048$ | 0.603 | $+0.103$ |

Table 14: Few-Shot (FS) LOTO Generalization ($k = 0$). Values represent Calibrator AUROC and $\Delta$ vs. LLM Confidence.

| | Gemini | | GPT-5.5 | | Qwen2.5-7B | |
|---|---|---|---|---|---|---|
| **Task** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** |
| Acute Kidney Injury (AKI) | 0.429 | $-0.078$ | 0.488 | $+0.026$ | 0.243 | $-0.257$ |
| Pulmonary Embolism (PE) | 0.622 | $-0.069$ | 0.585 | $-0.109$ | 0.519 | $+0.019$ |
| In-Hospital Mortality (IHM) | 0.659 | $+0.039$ | 0.646 | $+0.036$ | 0.626 | $+0.126$ |
| Revisit (3d) | 0.659 | $+0.056$ | 0.643 | $+0.090$ | 0.626 | $+0.126$ |
| Hospitalization | 0.586 | $-0.060$ | 0.545 | $-0.126$ | 0.667 | $+0.167$ |
| Critical Outcome | 0.755 | $-0.013$ | 0.723 | $-0.045$ | 0.662 | $+0.162$ |
| ED Mortality | 0.711 | $-0.058$ | 0.812 | $+0.011$ | 0.717 | $+0.217$ |
| ICU Transfer (12h) | 0.731 | $-0.003$ | 0.814 | $+0.031$ | 0.627 | $+0.127$ |
| ACS/MI | 0.742 | $-0.070$ | 0.805 | $-0.001$ | 0.539 | $+0.101$ |

### I.2.1 Bootstrap Confidence Intervals

To rigorously quantify the reliability of the calibrator when adapted to a specific clinical environment, we report calibrator performance after incorporating $k = 200$ labeled target-task examples into the LOTO training set. Tables 16, 17, and 18 present the mean AUROC and 95% bootstrap confidence intervals (derived across 5 independent random support-set draws) for Gemini 3 Flash Preview, GPT-5.5, and Qwen2.5-7B, respectively. $^*$ denotes a statistically significant improvement over best label-free baseline, $\max\{A(c, z),\ A(p^{\text{true}}, z)\}$, selected per task using the evaluation labels and therefore favorable to the baseline ($\Delta > 0$ with the 95% bootstrap interval of $\Delta$ lying strictly above 0). The baseline is computed on the full task cohort while the calibrator is evaluated on $\mathcal{L}_t \setminus S$, so the two differ by the 200 held-out support examples.

**Adaptation Insights.** Figures in this section illustrate the adaptation trajectories and corresponding improvement heatmaps for each model. Analyzing these across all prompting conditions yields several key insights:

- Across all models and prompting conditions, the learning curves demonstrate a sharp increase in calibrator AUROC as target-domain examples ($k$) are injected. The calibrator rapidly aligns with the target distribution even in very low-data regimes ($k = 50$).
- The heatmaps reveal that the most significant calibrator improvements (yielding $+0.20$ to $+0.30$ AUROC gains) consistently occur on structurally difficult tasks like AKI and PE.
- For diagnostic tasks where the LLM is already highly accurate (e.g., ACS/MI, ICU Transfer), the calibrator maintains or slightly improves upon the raw confidence baseline.
- For Qwen2.5-7B, which suffers from severe confidence collapse (defaulting to a static $0.50$ AUROC), the calibrator completely reconstructs the reliability estimate, showing massive $\Delta$ improvements across the entire heatmap grid.

Table 15: Chain-of-Thought (CoT) LOTO Generalization ($k = 0$). Values represent Calibrator AUROC and $\Delta$ vs. LLM Confidence.

| | Gemini | | GPT-5.5 | | Qwen2.5-7B | |
|---|---|---|---|---|---|---|
| **Task** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** | **AUROC** | **$\Delta$ vs Conf** |
| Acute Kidney Injury (AKI) | 0.445 | −0.002 | 0.527 | +0.074 | 0.435 | −0.103 |
| Pulmonary Embolism (PE) | 0.551 | −0.058 | 0.770 | +0.069 | 0.598 | +0.099 |
| In-Hospital Mortality (IHM) | 0.417 | −0.171 | 0.483 | −0.152 | 0.482 | −0.018 |
| Revisit (3d) | 0.595 | −0.013 | 0.591 | +0.025 | 0.579 | +0.075 |
| Hospitalization | 0.650 | −0.040 | 0.574 | −0.140 | 0.633 | +0.131 |
| Critical Outcome | 0.743 | −0.024 | 0.845 | +0.065 | 0.706 | +0.216 |
| ED Mortality | 0.738 | −0.003 | 0.877 | +0.054 | 0.738 | +0.235 |
| ICU Transfer (12h) | 0.752 | +0.020 | 0.859 | +0.062 | 0.715 | +0.221 |
| ACS/MI | 0.694 | −0.024 | 0.791 | −0.055 | 0.707 | +0.206 |

Table 16: **Gemini 3 Flash Preview**: Mean Calibrator AUROC and 95% bootstrap confidence intervals at $k = 200$ target examples.

| **Task** | **Zero-Shot** | **Few-Shot** | **Chain-of-Thought** |
|---|---|---|---|
| Acute Kidney Injury (AKI) | 0.810* [0.789 - 0.822] | 0.778* [0.759 - 0.797] | 0.743* [0.727 - 0.760] |
| Pulmonary Embolism (PE) | 0.845* [0.839 - 0.851] | 0.800* [0.792 - 0.807] | 0.818* [0.805 - 0.828] |
| In-Hospital Mortality (IHM) | 0.670 [0.646 - 0.690] | 0.637 [0.621 - 0.654] | 0.615 [0.599 - 0.628] |
| Revisit (3d) | 0.681* [0.675 - 0.688] | 0.659 [0.649 - 0.668] | 0.612 [0.596 - 0.628] |
| Hospitalization | 0.712 [0.695 - 0.727] | 0.746* [0.741 - 0.752] | 0.702 [0.690 - 0.711] |
| Critical Outcome | 0.762* [0.756 - 0.767] | 0.798* [0.793 - 0.803] | 0.756 [0.754 - 0.759] |
| ED Mortality | 0.789* [0.787 - 0.791] | 0.790* [0.780 - 0.796] | 0.758* [0.746 - 0.770] |
| ICU Transfer (12h) | 0.739* [0.735 - 0.742] | 0.740* [0.735 - 0.745] | 0.759* [0.745 - 0.772] |
| ACS/MI | 0.833* [0.822 - 0.845] | 0.857* [0.850 - 0.864] | 0.854* [0.846 - 0.862] |

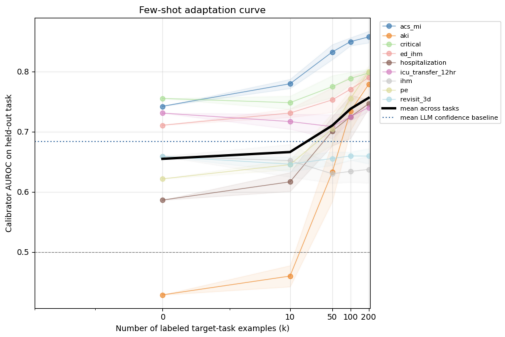


Figure 2: Gemini FS Adaptation Curve

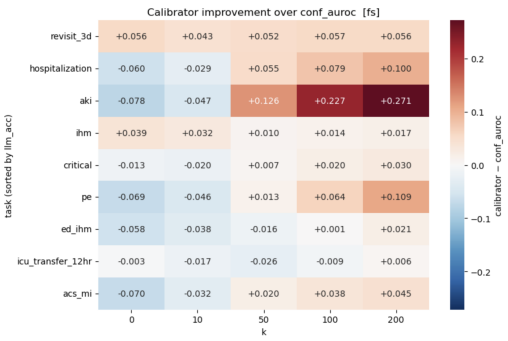


Figure 3: Gemini FS Improvement Heatmap

Table 17: **GPT-5.5**: Mean Calibrator AUROC and 95% bootstrap confidence intervals at $k = 200$ target examples.

| **Task** | **Zero-Shot** | **Few-Shot** | **Chain-of-Thought** |
|---|---|---|---|
| Acute Kidney Injury (AKI) | 0.792* [0.785 - 0.798] | 0.790* [0.780 - 0.801] | 0.773* [0.754 - 0.788] |
| Pulmonary Embolism (PE) | 0.824* [0.814 - 0.831] | 0.807* [0.802 - 0.815] | 0.833* [0.829 - 0.836] |
| In-Hospital Mortality (IHM) | 0.732* [0.717 - 0.745] | 0.689 [0.677 - 0.698] | 0.645 [0.605 - 0.686] |
| Revisit (3d) | 0.572 [0.561 - 0.585] | 0.665* [0.662 - 0.667] | 0.604 [0.589 - 0.617] |
| Hospitalization | 0.716 [0.701 - 0.726] | 0.739 [0.737 - 0.742] | 0.687 [0.683 - 0.691] |
| Critical Outcome | 0.830* [0.825 - 0.832] | 0.772 [0.766 - 0.777] | 0.846* [0.843 - 0.849] |
| ED Mortality | 0.885* [0.882 - 0.887] | 0.854* [0.852 - 0.857] | 0.884* [0.878 - 0.891] |
| ICU Transfer (12h) | 0.841* [0.835 - 0.846] | 0.825* [0.821 - 0.830] | 0.861* [0.857 - 0.865] |
| ACS/MI | 0.896* [0.895 - 0.899] | 0.891* [0.889 - 0.894] | 0.854* [0.848 - 0.859] |

Table 18: **Qwen2.5-7B**: Mean Calibrator AUROC and 95% bootstrap confidence intervals at $k = 200$ target examples.

| Task | Zero-Shot | Few-Shot | Chain-of-Thought |
|---|---|---|---|
| Acute Kidney Injury (AKI) | 0.818* [0.809 - 0.828] | 0.821* [0.814 - 0.829] | 0.770* [0.760 - 0.782] |
| Pulmonary Embolism (PE) | 0.747* [0.733 - 0.762] | 0.698* [0.691 - 0.708] | 0.751* [0.745 - 0.756] |
| In-Hospital Mortality (IHM) | 0.689* [0.681 - 0.700] | 0.752* [0.741 - 0.763] | 0.551* [0.543 - 0.563] |
| Revisit (3d) | 0.613* [0.603 - 0.623] | 0.639* [0.632 - 0.645] | 0.630* [0.620 - 0.637] |
| Hospitalization | 0.762* [0.756 - 0.768] | 0.766* [0.763 - 0.770] | 0.684* [0.680 - 0.687] |
| Critical Outcome | 0.737* [0.731 - 0.743] | 0.745* [0.732 - 0.755] | 0.715* [0.710 - 0.720] |
| ED Mortality | 0.804* [0.797 - 0.809] | 0.753* [0.744 - 0.762] | 0.744* [0.738 - 0.750] |
| ICU Transfer (12h) | 0.738* [0.734 - 0.742] | 0.701* [0.689 - 0.714] | 0.717* [0.713 - 0.721] |
| ACS/MI | 0.785* [0.774 - 0.795] | 0.823* [0.815 - 0.831] | 0.823* [0.815 - 0.830] |

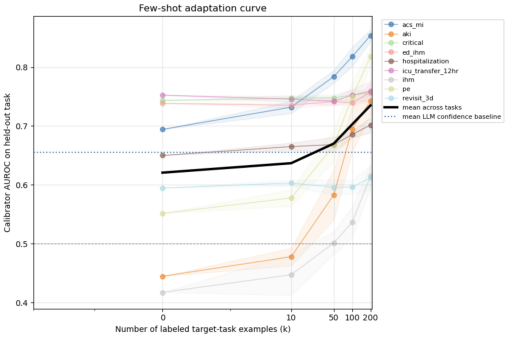


Figure 4: Gemini CoT Adaptation Curve

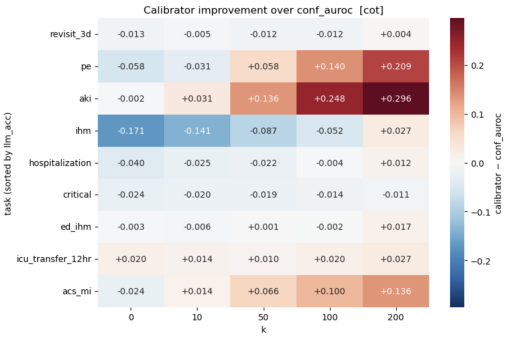


Figure 5: Gemini CoT Improvement Heatmap

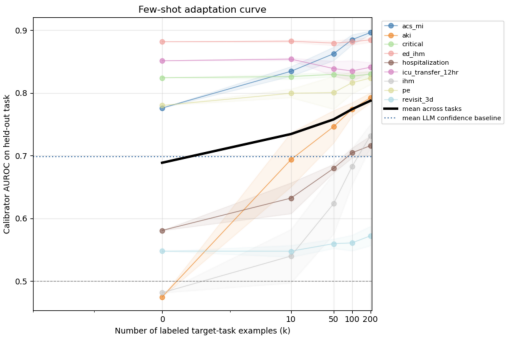


Figure 6: GPT ZS Adaptation Curve

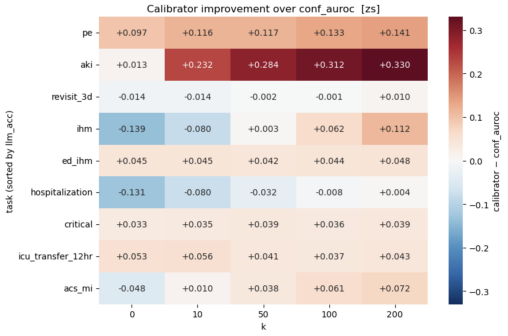


Figure 7: GPT ZS Improvement Heatmap

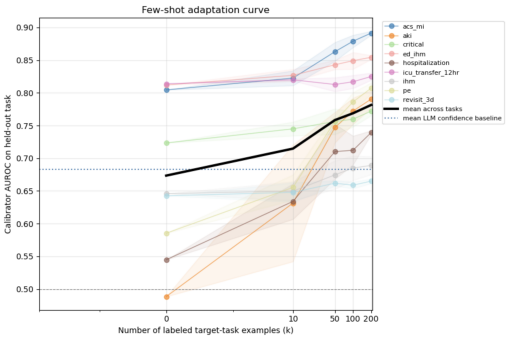


Figure 8: GPT FS Adaptation Curve

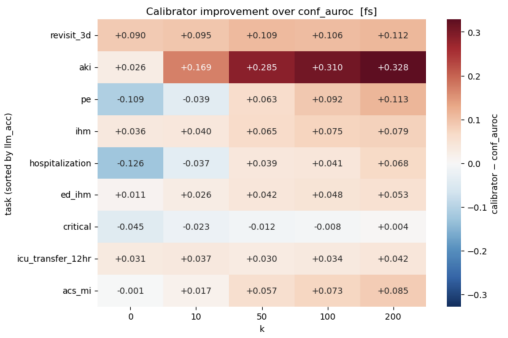


Figure 9: GPT FS Improvement Heatmap

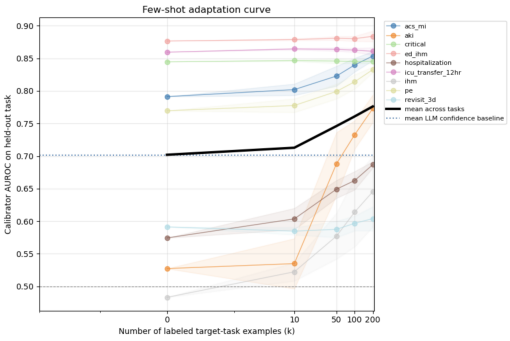


Figure 10: GPT CoT Adaptation Curve

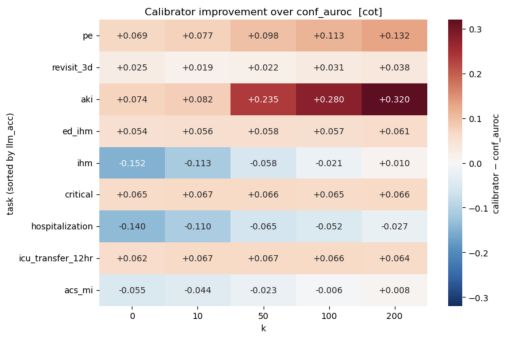


Figure 11: GPT CoT Improvement Heatmap

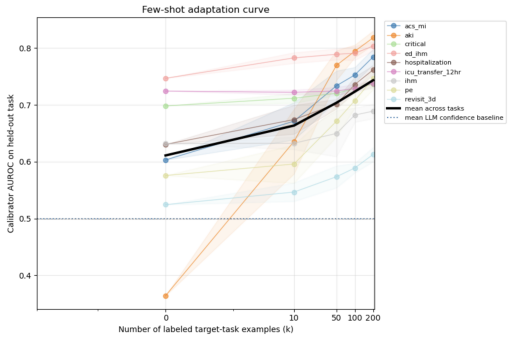


Figure 12: Qwen ZS Adaptation Curve

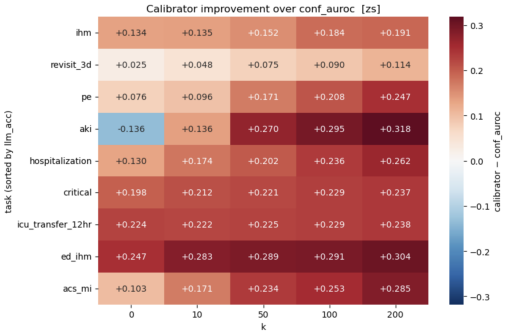


Figure 13: Qwen ZS Improvement Heatmap

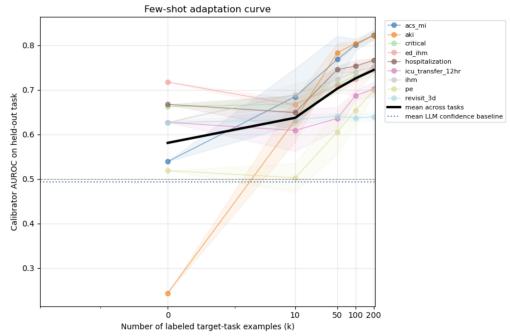


Figure 14: Qwen FS Adaptation Curve

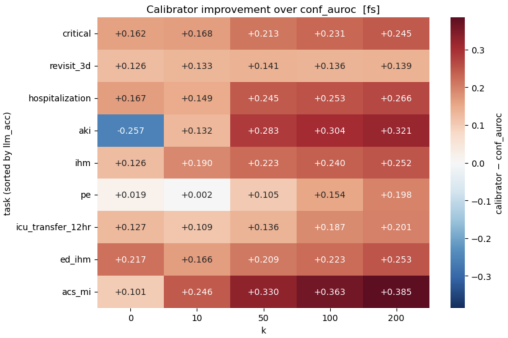


Figure 15: Qwen FS Improvement Heatmap

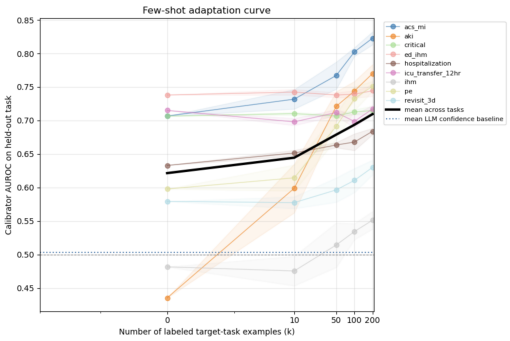


Figure 16: Qwen CoT Adaptation Curve

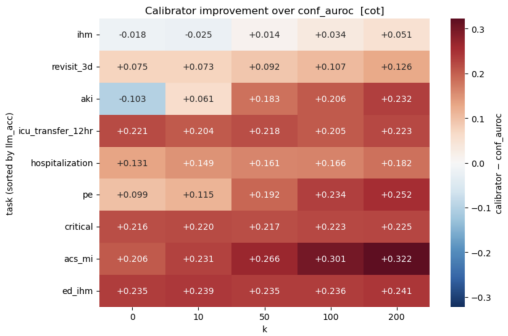


Figure 17: Qwen CoT Improvement Heatmap

# J ABLATIONS

## J.1 FEATURE CONTRIBUTIONS

To evaluate the utility of our proposed features in adjusting LLM confidence, we analyzed the global feature importance of the trained XGBoost calibrators using SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). Figures 18 and 19 illustrate the mean absolute SHAP values, representing the average impact of individual features on the calibrator's output magnitude during Leave-One-Task-Out (LOTO) cross-validation for ACS/MI, the ordering almost replicates across tasks although not exactly the same. For CoT, the top SHAP orders also contain reasoning dimensions.

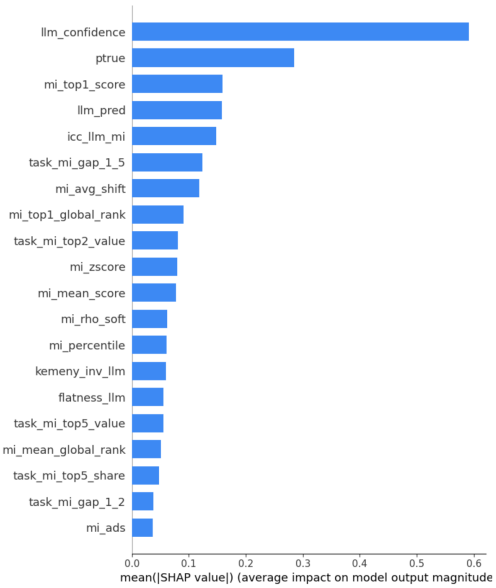


Figure 18: SHAP feature importance ranking for API-based models.

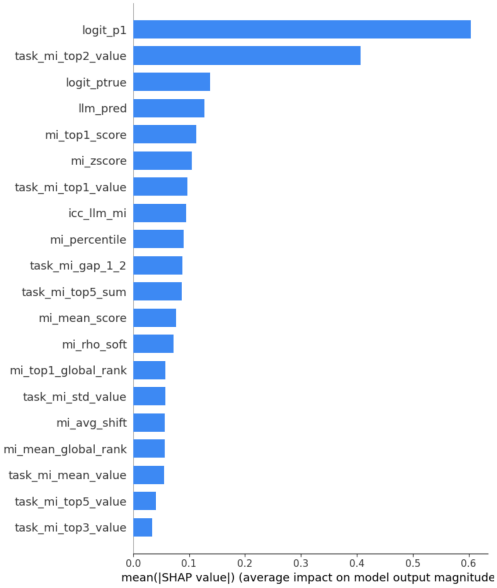


Figure 19: SHAP feature importance ranking for open-weight models.

The SHAP analysis reveals that the calibrator fuses signals across all proposed feature families rather than relying on any single one. The dominant feature differs by model access. For the API-based models (Figure 18), the verbalized outputs lead. `llm_confidence` carries the largest single impact, followed by `ptrue`, with `mi_top1_score`, `llm_pred` and `icc_llm_mi` next. For the open-weight model (Figure 19), the logit-derived features lead. `logit_p1` is the single most influential feature, above every attribution and task descriptor, with `logit_ptrue` also ranking above `llm_pred`. The fact that the token-level distribution outranks the model's own verbalized confidence where both are available is consistent with the degeneracy documented in Section 5, Qwen's stated confidence carries no correctness information, while the distribution it computes may. Attribution and task-descriptor features (`mi_top1_score`, `icc_llm_mi`, `task_mi_gap_1_5`, `mi_zscore`) contribute substantially in both regimes.

The feature `mi_rho_zone` was designed to measure coarse-grained ordering agreement by bucketing feature ranks into zones of width $\max(K, \lfloor d_t/5 \rfloor)$. However, because $K = 5$ and every evaluated task consists of $d_t \geq 26$ features, the calculated zone width strictly exceeds $K$ across all datasets. Consequently, all model-selected features fall entirely within the first zone, rendering the model-side zone vector constant for every patient instance. This feature therefore provides zero discriminative variance and contributes nothing to the calibrator. We report its definition as it may become informative in experiments that evaluate larger values of $K$.

To isolate the discriminative power of each feature subset, we performed a within-task ablation, predicting instance-level correctness using only one feature family at a time. Table 19 reports the Zero-Shot AUROC across all three models.

Table 19: Within-Task Feature Ablation (Zero-Shot AUROC). Performance of the calibrator when restricted to a single feature family: **Conf** (Model Outputs), **Base** (basic attribution divergence), **Adv** (includes rank correlation, positional shift, and MI distributions), and **Agg** (Rank Aggregation).

| | Gemini 3 Flash | | | | GPT-5.5 | | | | Qwen2.5-7B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Conf** | **Base** | **Adv** | **Agg** | **Conf** | **Base** | **Adv** | **Agg** | **Conf** | **Base** | **Adv** | **Agg** |
| AKI | 0.600 | **0.839** | 0.804 | 0.780 | 0.537 | **0.809** | 0.795 | 0.779 | 0.500 | **0.852** | 0.850 | 0.813 |
| PE | 0.828 | **0.833** | 0.821 | 0.759 | **0.832** | 0.723 | 0.723 | 0.656 | 0.500 | 0.772 | **0.780** | 0.638 |
| ACS / MI | 0.762 | **0.799** | 0.774 | 0.798 | **0.848** | 0.807 | 0.756 | 0.785 | 0.500 | **0.649** | 0.635 | 0.623 |
| Critical | **0.750** | 0.612 | 0.585 | 0.625 | **0.773** | 0.589 | 0.576 | 0.577 | 0.500 | **0.648** | 0.642 | 0.631 |
| ED Mort. | **0.773** | 0.647 | 0.615 | 0.529 | **0.875** | 0.694 | 0.686 | 0.652 | 0.500 | **0.649** | 0.626 | 0.619 |
| Hosp. | **0.733** | 0.649 | 0.630 | 0.523 | **0.725** | 0.609 | 0.587 | 0.552 | 0.500 | **0.616** | 0.607 | 0.534 |
| ICU (12h) | **0.728** | 0.579 | 0.559 | 0.619 | **0.813** | 0.636 | 0.619 | 0.615 | 0.500 | **0.645** | 0.640 | 0.616 |
| IHM | **0.665** | 0.518 | 0.567 | 0.550 | **0.711** | 0.606 | 0.597 | 0.635 | 0.500 | **0.594** | 0.585 | 0.559 |
| Revisit | **0.684** | 0.639 | 0.634 | 0.624 | **0.565** | 0.560 | 0.554 | 0.548 | 0.500 | 0.611 | 0.601 | **0.616** |

This resolves a critical vulnerability in deployment, what happens when verbalized confidence is uninformative? The behavior of Qwen2.5-7B illustrates the extreme limit of this failure. Across all nine tasks, Qwen's raw verbalized confidence provides mathematically zero discrimination ($0.500$ AUROC). However, passing its text-based outputs through the attribution divergence pipeline entirely rescues the calibrator, lifting performance to $0.852$ on AKI and $0.780$ on PE without accessing internal logits. Similarly, for the highly capable frontier models (GPT-5.5 and Gemini), raw confidence yields near-random discrimination on structurally complex etiologies like AKI (e.g., $0.537$ and $0.600$ AUROC). Yet, base attribution divergence features alone achieve $0.809$ and $0.839$ AUROC, respectively, proving that the models' internal reasoning structures retain a strong correctness signal even when their self-assessments are uninformative.

While raw confidence maintains a strong residual signal on broader prognostic endpoints (such as ED Mortality and ICU Transfer) for the proprietary APIs, it *cannot be trusted*. The calibrator's **combined feature-set** bridges this gap, relying on attribution divergence to adjust the decision boundary exactly where structural reasoning breaks down.

**The Role of Task Descriptors.** Finally, in this within-task evaluation, the task-level statistics yield exactly $0.500$ AUROC across all models. Because these dataset-level descriptors are constant for all patients within a specific task, they inherently possess zero within-task variance. This mathematically confirms their exclusive, specialized role, they are domain-bridging features designed solely to identify concept shift and anchor the calibrator during cross-task transfer.

### J.2 Per-Task Calibrator

To establish the upper-bound performance achievable when task-specific training data is available, we evaluate a **Per-Task Calibrator**. Unlike Leave-One-Task-Out (LOTO) cross-validation which evaluates cross-task transfer, this model is trained and evaluated strictly within each task using stratified 5-fold cross-validation.

Tables 20 reports the 5-fold CV calibrator AUROC across Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT) prompting for Gemini 3 Flash Preview, GPT-5.5, and Qwen2.5-7B, respectively.

**Comparative Insights on Within-Task Calibration.** Evaluating the calibrator via within-task 5-fold cross-validation establishes an upper-bound performance limit. It isolates how well the calibrator can learn task-specific overconfidence patterns when trained on the target distribution.

- For Qwen2.5-7B, training the calibrator on CoT features yields the highest absolute AUROC on distinct diagnostic tasks, notably ACS/MI ($0.931$) and PE ($0.919$). The explicit reasoning traces provide strong discriminative signals for these complex conditions.
- FS prompting improves or stabilizes calibration on general prognostic tasks such as IHM (rising to $0.743$) and Critical Outcome ($0.782$) compared to the Zero-Shot baseline.

Table 20: Per-task 5-fold CV Calibrator AUROC across prompting conditions.

| | Gemini 3 Flash Preview | | | GPT-5.5 | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **ZS** | **FS** | **CoT** | **ZS** | **FS** | **CoT** | **ZS** | **FS** | **CoT** |
| AKI | 0.838 | 0.828 | 0.822 | 0.828 | 0.827 | 0.840 | 0.856 | 0.856 | 0.831 |
| PE | 0.904 | 0.851 | 0.935 | 0.832 | 0.836 | 0.889 | 0.859 | 0.811 | 0.919 |
| IHM | 0.652 | 0.626 | 0.683 | 0.692 | 0.694 | 0.663 | 0.696 | 0.743 | 0.599 |
| Revisit (3d) | 0.658 | 0.660 | 0.679 | 0.577 | 0.645 | 0.639 | 0.642 | 0.635 | 0.665 |
| Hospitalization | 0.751 | 0.758 | 0.749 | 0.729 | 0.748 | 0.720 | 0.794 | 0.792 | 0.748 |
| Critical | 0.742 | 0.801 | 0.763 | 0.798 | 0.753 | 0.839 | 0.759 | 0.782 | 0.730 |
| ED Mortality | 0.775 | 0.788 | 0.799 | 0.877 | 0.850 | 0.903 | 0.795 | 0.766 | 0.777 |
| ICU (12h) | 0.710 | 0.732 | 0.754 | 0.827 | 0.820 | 0.856 | 0.731 | 0.710 | 0.755 |
| ACS/MI | 0.878 | 0.878 | 0.911 | 0.896 | 0.907 | 0.895 | 0.841 | 0.868 | 0.931 |

- While ZS performs well on AKI ($0.856$) and Hospitalization ($0.794$), it generally lacks the structured attribution signals generated by in-context examples or explicit reasoning, leading to lower peak performance on reasoning-heavy tasks.

### J.3 Target-Only vs. Cross-Task Transfer Adaptation Sweeps

To quantify the value of cross-task pretraining across different sample budgets, we compare the target-only calibrator (trained strictly on $k$ target examples) against our LOTO transfer framework (trained on source tasks + $k$ target examples) across $k \in \{10, 50, 100, 200\}$. Tables 21, 22, and 23 report the Target-Only AUROC (**Tgt**), the LOTO transfer AUROC (**+ src**), and the resulting transfer gain ($\mathbf{\Delta}$) for all 9 clinical tasks under Zero-Shot prompting.

This comparison yields the following findings:

- With low-label adaptation ($k = 10$), target-only calibrators suffer from severe variance and overconfidence, yielding poor mean discrimination across all three model families ($0.568$ for Gemini, $0.570$ for GPT-5.5, and $0.566$ for Qwen2.5-7B). Incorporating source-task pretraining bridges this deficit, providing average transfer gains of $+0.101$, $+0.164$, and $+0.098$ AUROC, respectively, with peak gains reaching $+0.281$ on ED Mortality for GPT-5.5 and $+0.216$ on Critical Outcome for Qwen2.5-7B.
- As in-domain target labels increase to $k = 200$, target-only models converge toward within-task performance ($\sim$ 0.72–0.75 mean AUROC). Crucially, source-task pretraining continues to outperform target-only models by $+0.032$ (Gemini), $+0.036$ (GPT-5.5), and $+0.026$ (Qwen2.5-7B). This confirms that cross-task pretraining contributes to the performance.
- While certain complex tasks (such as In-Hospital Mortality on GPT-5.5) experience slightly negative transfer at $k = 10$, expanding the target support to $k \geq 100$ enables the tree ensemble to restore positive transfer gains of $+0.049$ and $+0.048$ AUROC at $k = 200$.

### J.4 Cross-Condition Calibrator

In this ablation, we evaluate a Cross-Condition Calibrator that fuses multi-prompt feature representations combining Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT) signals to predict the correctness of the model's ZS prediction ($\hat{y}_i^{\text{ZS}}$).

**Feature Formulation.** For each patient instance $x_i$, the multi-prompt feature vector $\phi_i^{\text{cross}}$ integrates intra-condition metrics with cross-condition interactions:

- **Prediction Agreement (Self-Consistency):** Features indicating whether the model's predicted diagnostic class remains stable across prompting conditions ($\hat{y}^{\text{ZS}} == \hat{y}^{\text{FS}}$ and $\hat{y}^{\text{ZS}} == \hat{y}^{\text{CoT}}$).
- **Confidence & $P(true)$ Deltas:** The shifts in raw confidence and self-evaluation between conditions.

Table 21: Gemini 3 Flash Preview Calibrator AUROC with and without source-task pretraining across all four $k$ values

| | $\mathbf{k=10}$ | | | $\mathbf{k=50}$ | | | $\mathbf{k=100}$ | | | $\mathbf{k=200}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** |
| ACS / MI | 0.541 | 0.743 | +0.202 | 0.749 | 0.779 | +0.030 | 0.801 | 0.813 | +0.012 | 0.823 | 0.833 | +0.009 |
| ED mortality | 0.592 | 0.764 | +0.172 | 0.724 | 0.768 | +0.044 | 0.743 | 0.782 | +0.039 | 0.753 | 0.789 | +0.036 |
| ICU transfer | 0.581 | 0.745 | +0.165 | 0.660 | 0.721 | +0.061 | 0.683 | 0.732 | +0.049 | 0.688 | 0.739 | +0.051 |
| Critical outcome | 0.588 | 0.739 | +0.151 | 0.698 | 0.734 | +0.035 | 0.697 | 0.747 | +0.050 | 0.718 | 0.762 | +0.043 |
| PE | 0.547 | 0.629 | +0.082 | 0.760 | 0.770 | +0.010 | 0.803 | 0.817 | +0.014 | 0.831 | 0.845 | +0.014 |
| Revisit (3d) | 0.571 | 0.642 | +0.071 | 0.620 | 0.662 | +0.042 | 0.609 | 0.667 | +0.059 | 0.637 | 0.681 | +0.044 |
| Hospitalization | 0.553 | 0.614 | +0.061 | 0.641 | 0.662 | +0.021 | 0.666 | 0.687 | +0.021 | 0.674 | 0.712 | +0.038 |
| AKI | 0.584 | 0.590 | +0.006 | 0.690 | 0.715 | +0.025 | 0.750 | 0.782 | +0.032 | 0.788 | 0.810 | +0.022 |
| IHM | 0.558 | 0.558 | +0.000 | 0.649 | 0.631 | −0.017 | 0.631 | 0.644 | +0.013 | 0.636 | 0.670 | +0.034 |
| **Mean** | 0.568 | 0.669 | +0.101 | 0.688 | 0.716 | +0.028 | 0.709 | 0.741 | +0.032 | 0.728 | 0.760 | +0.032 |

Table 22: GPT-5.5 Calibrator AUROC with and without source-task pretraining across all four $k$ values.

| | $\mathbf{k=10}$ | | | $\mathbf{k=50}$ | | | $\mathbf{k=100}$ | | | $\mathbf{k=200}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** |
| ED mortality | 0.601 | 0.882 | +0.281 | 0.810 | 0.879 | +0.069 | 0.844 | 0.881 | +0.038 | 0.842 | 0.885 | +0.043 |
| ICU transfer | 0.593 | 0.854 | +0.261 | 0.743 | 0.839 | +0.096 | 0.786 | 0.835 | +0.049 | 0.786 | 0.841 | +0.055 |
| PE | 0.554 | 0.799 | +0.246 | 0.724 | 0.800 | +0.077 | 0.774 | 0.816 | +0.043 | 0.786 | 0.824 | +0.037 |
| ACS / MI | 0.619 | 0.834 | +0.216 | 0.802 | 0.862 | +0.060 | 0.848 | 0.885 | +0.037 | 0.870 | 0.896 | +0.026 |
| Critical outcome | 0.614 | 0.826 | +0.212 | 0.738 | 0.830 | +0.092 | 0.764 | 0.827 | +0.063 | 0.778 | 0.830 | +0.052 |
| AKI | 0.530 | 0.694 | +0.164 | 0.740 | 0.746 | +0.006 | 0.763 | 0.774 | +0.011 | 0.788 | 0.792 | +0.005 |
| Hospitalization | 0.548 | 0.632 | +0.084 | 0.656 | 0.680 | +0.024 | 0.666 | 0.704 | +0.038 | 0.671 | 0.716 | +0.045 |
| Revisit (3d) | 0.506 | 0.548 | +0.041 | 0.547 | 0.560 | +0.012 | 0.551 | 0.561 | +0.010 | 0.561 | 0.572 | +0.011 |
| IHM | 0.567 | 0.540 | −0.027 | 0.637 | 0.623 | −0.014 | 0.656 | 0.682 | +0.026 | 0.682 | 0.732 | +0.049 |
| **Mean** | 0.570 | 0.734 | +0.164 | 0.711 | 0.758 | +0.047 | 0.739 | 0.774 | +0.035 | 0.752 | 0.788 | +0.036 |

Table 23: Qwen2.5-7B Calibrator AUROC with and without source-task pretraining across all four $k$ values.

| | $\mathbf{k=10}$ | | | $\mathbf{k=50}$ | | | $\mathbf{k=100}$ | | | $\mathbf{k=200}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** | **Tgt** | **+Src** | **Δ** |
| Critical outcome | 0.495 | 0.712 | +0.216 | 0.644 | 0.721 | +0.077 | 0.669 | 0.729 | +0.060 | 0.692 | 0.737 | +0.045 |
| ED mortality | 0.568 | 0.783 | +0.215 | 0.711 | 0.789 | +0.078 | 0.731 | 0.791 | +0.060 | 0.759 | 0.804 | +0.044 |
| ICU transfer | 0.596 | 0.722 | +0.126 | 0.646 | 0.725 | +0.079 | 0.677 | 0.729 | +0.052 | 0.685 | 0.738 | +0.053 |
| Hospitalization | 0.572 | 0.674 | +0.102 | 0.688 | 0.702 | +0.013 | 0.723 | 0.736 | +0.013 | 0.744 | 0.762 | +0.018 |
| ACS / MI | 0.570 | 0.671 | +0.101 | 0.669 | 0.734 | +0.065 | 0.734 | 0.753 | +0.019 | 0.765 | 0.785 | +0.019 |
| IHM | 0.546 | 0.633 | +0.086 | 0.640 | 0.650 | +0.010 | 0.633 | 0.682 | +0.049 | 0.641 | 0.689 | +0.048 |
| Revisit (3d) | 0.528 | 0.547 | +0.019 | 0.557 | 0.574 | +0.016 | 0.579 | 0.589 | +0.010 | 0.606 | 0.613 | +0.007 |
| PE | 0.582 | 0.596 | +0.014 | 0.688 | 0.671 | −0.016 | 0.723 | 0.708 | −0.016 | 0.753 | 0.747 | −0.006 |
| AKI | 0.636 | 0.636 | +0.000 | 0.801 | 0.770 | −0.031 | 0.808 | 0.795 | −0.013 | 0.819 | 0.818 | −0.001 |
| **Mean** | 0.566 | 0.664 | +0.098 | 0.672 | 0.704 | +0.032 | 0.697 | 0.724 | +0.026 | 0.718 | 0.744 | +0.026 |

- **Cross-Prompt Attribution Divergence:** Rank correlation metrics (e.g., Kendall's $\tau$, Jaccard top-$k$ overlap) measuring the stability of the LLM's feature attribution rankings across different prompts.
- **CoT Reasoning Dimensions:** Implicit reasoning quality from the Chain-of-Thought narrative, covering logical validity, clinical premise alignment, and counterfactual robustness via sentence embeddings.
- **Baseline ZS & FS Metrics:** Standard attribution divergence against the data-derived reference $\rho_t$ and raw output probabilities for the individual conditions.

Table 24: Cross-Condition Calibrator Performance. Predicts Zero-Shot LLM correctness under LOTO cross-validation.

| | Gemini 3 Flash | | GPT-5.5 | | Qwen2.5-7B | |
|---|---|---|---|---|---|---|
| **Task** | **AUROC** | **Δ vs Conf** | **AUROC** | **Δ vs Conf** | **AUROC** | **Δ vs Conf** |
| Acute Kidney Injury (AKI) | 0.448 | +0.001 | 0.486 | +0.024 | 0.374 | −0.126 |
| Pulmonary Embolism (PE) | 0.617 | +0.067 | 0.799 | +0.116 | 0.577 | +0.077 |
| In-Hospital Mortality (IHM) | 0.572 | −0.016 | 0.545 | −0.075 | 0.632 | +0.134 |
| Revisit (3d) | 0.536 | −0.135 | 0.565 | +0.003 | 0.577 | +0.078 |
| Hospitalization | 0.688 | −0.023 | 0.635 | −0.077 | 0.692 | +0.192 |
| Critical Outcome | 0.734 | −0.014 | 0.803 | +0.012 | 0.682 | +0.182 |
| ED Mortality | 0.793 | +0.051 | 0.889 | +0.052 | 0.744 | +0.244 |
| ICU Transfer (12h) | 0.745 | +0.018 | 0.851 | +0.053 | 0.726 | +0.226 |
| ACS/MI | 0.773 | +0.028 | 0.811 | −0.013 | 0.561 | +0.061 |

**Key Findings.**

1. Although this would increase latency by having 6 inferences per record, incorporating auxiliary cross-prompt interaction features (such as prediction agreement and confidence shifts) yields substantial calibration improvements on clinically nuanced tasks. For GPT-5.5, this multi-condition fusion improves PE calibration AUROC to $0.799$ ($+0.116$ over raw confidence).
2. Disagreement between a model's ZS and FS prediction is a massive indicator of hallucination. The calibrator heavily weights these prediction agreement flags to discount falsely confident ZS outputs.
3. For Qwen2.5-7B, where raw zero-shot confidence provides near-random discrimination ($0.500$ AUROC), multi-condition feature fusion reconstructs the calibration surface. It achieves an average gain of $+0.120$ AUROC across all 9 clinical tasks by bypassing the collapsed ZS probabilities and relying on structural divergence and cross-prompt consistency.
4. Preliminary evaluations indicated that adding a fewshot adaptation to cross-condition calibrator had similar performances across tasks when compared with a standard single-condition fewshot adapted calibrator.

## K Why Calibrate an LLM Instead of Training a Tabular Classifier?

A natural question regarding the practical deployment of our external calibrator is whether the required labeling budget ($k = 10$–$200$) and dataset statistics would be better spent training a traditional supervised machine learning model directly on the tabular data. Dedicated XGBoost classifiers achieve high discrimination ($0.88 - 0.96$ AUROC) on these tasks (Brown et al., 2025), comfortably outperforming LLMs (Table 1). However, deploying an LLM calibrator serves a fundamentally different infrastructural purpose than training a standalone predictive model, driven by three key factors:

**Mixed-Modal LLM Deployments.** While traditional machine learning remains the gold standard for purely tabular clinical prediction, healthcare systems are actively integrating LLMs into existing workflows to handle complex, mixed-modal environments (e.g., processing unstructured clinical notes alongside vital signs). Our objective is not to argue that LLMs surpass XGBoost in tabular accuracy, but rather to acknowledge the reality of their ongoing deployment. If an LLM is utilized for its generative flexibility, providing a safety guardrail for its inferences is a strict prerequisite. Our task is to optimize the reliability estimate of the LLM, rather than replace it with well-performing models.

**Data Efficiency of Calibration.** Training a full clinical XGBoost model from scratch to reach $0.88-0.96$ AUROC requires a large number of patient records to understand the underlying patterns.

In contrast, the external calibrator does not learn the underlying clinical outcome from scratch, it learns the pattern of how the LLM expresses overconfidence and structural reasoning breakdown. Because the calibrator operates on the LLM's outputs rather than the raw patient physiology, it achieves robust reliability estimation with just 10 to 200 labeled examples.

**Cross-Task Generalization.** A standard tabular XGBoost model is strictly task-specific, a model trained to predict Acute Kidney Injury provides zero utility for predicting Pulmonary Embolism. Our calibrator, however, functions as a domain-agnostic meta-model. As demonstrated in our transfer ablations, source-task pretraining may leverage the generalized behaviors of language models, allowing the calibrator to transfer across entirely different clinical domains. This capability significantly reduces the target-task labeling budget required to achieve safe deployment.

# L Extension to Multiclass Tasks

We started with binary outcomes, which helped the comparison of metrics in this work. With two classes the signing operation of Eq. 1 is well-defined and the two metrics are directly comparable on the same score. The calibrator's intuition carries over to multiclass tasks without modification, its target $z = \mathbf{1}[\hat{y} = y]$ is binary however many classes the task has, and the attribution and aggregation features are class-agnostic. For open-weight models, the token-level probe requires only extending the candidate token set from $\{0, 1\}$ to the $C$ label tokens. The classifier metric does not generalize as cleanly, with $C > 2$ there is no single signed score, and one would need a multiclass AUROC generalization. Self-knowledge is unaffected.

# M Comparative Baseline Analysis

In this section, we analyze the theoretical and operational reasons why standard univariate post-hoc calibration methods and multi-query semantic uncertainty estimators cannot serve as direct drop-in solutions for instance-level tabular LLM calibration.

## M.1 Post-hoc Recalibration Methods

A natural question is whether standard post-hoc recalibration methods, specifically Platt scaling, temperature scaling (Guo et al., 2017), or isotonic regression on target-task examples ($S \subset \mathcal{L}_t$) could rescue verbalized LLM confidence without extracting attribution features.

Let $c \in [0.5, 1.0]$ denote the scalar verbalized confidence emitted by the model, and let $f : [0.5, 1.0] \to [0, 1]$ denote a univariate recalibration mapping parameterized by target examples.

1. Temperature scaling and Platt scaling are strictly monotonic increasing transformations of the original scalar input $c$. Applying standard parametric 1D scaling leaves the self-knowledge AUROC mathematically identical to the uncalibrated baseline.
2. On complex etiologies such as Acute Kidney Injury (AKI), raw model self-knowledge inverts below chance. A monotonic mapping cannot invert or reshape the underlying feature space to generate positive discrimination.
3. For open-weight models such as Qwen2.5-7B, verbalized confidence collapses to three discrete values ($c \in \{0.75, 0.85, 0.95\}$). Even non-parametric monotonic transforms like isotonic regression cannot recover ranking information among thousands of ties.

## M.2 Comparison with Multi-Inference Baselines

Recent literature has explored uncertainty quantification via semantic entropy (Quevedo et al., 2024) and auxiliary calibration (Shen et al., 2024). We contrast these approaches with our external calibrator:

1. Semantic entropy (Quevedo et al., 2024) estimates semantic uncertainty by sampling $M$ independent stochastic generations at high temperature ($T > 0$) and clustering outputs into bidirectional entailment classes using a secondary natural language inference (NLI) model.

In clinical EHR deployments, this requires $M$ LLM inferences per patient encounter plus additional NLI-model inference calls for semantic clustering. In contrast, our proposed calibrator operates within a strict two-call deterministic inferences ($T = 0.0$), eliminating stochastic generation variance and reducing inference latency and API cost. Furthermore, since our work focuses on structured tabular data, prediction outputs are not free-form natural language strings, making semantic clustering redundant.

2. Frameworks such as Thermometer (Shen et al., 2024) train an auxiliary model to predict an optimal temperature scaling parameter using the LLM's internal feature representations. While effective for open-weight models, these methods are fundamentally incompatible with proprietary models (e.g., Gemini 3 Flash and GPT-5.5), where model-internals are are inaccessible. Our calibrator relies strictly on externalized signals and data-derived reference statistics.