# Calibrating Agentic LLMs for Clinical Prediction

**Zheqi Lv**[1] **Fuying Ding**[1] **Meihan Wei**[1] **Haoyang Li**[1]
**Christos Faloutsos**[2] **Fei Wang**[1] **Chengxi Zang**[1]
[1]Weill Cornell Medicine, Cornell University
[2]Carnegie Mellon University

## Abstract

Agentic Large Language Models (LLMs) show great promise in clinical prediction using real-world clinical data. While recent studies have primarily examined their discrimination performance, their calibration performance—that is, whether predicted probabilities accurately reflect absolute clinical risk—remains largely unknown. This gap limits their use in high-stakes medical settings, where well-calibrated risk estimates are essential for guiding treatment and prevention decision-making. In this work, we provide conditional theoretical analyses of how agentic LLM architectures, task knowledge, and tool use affect calibration performance in clinical prediction. Motivated by these analyses, we propose *C²Align*, a Calibrated and Clinically Aligned framework, to optimize calibration while maintaining or even improving discrimination. *C²Align* aligns two complementary agent branches: an ensemble of task-homogeneous LLM agents and a set of clinical task-inspired LLM agents. This alignment is achieved through a novel discrimination- and calibration-aware loss that encourages clinically meaningful agreement while controlling probabilistic miscalibration. We evaluate this framework on four clinical prediction benchmarks—spanning prognostic prediction and disease diagnosis tasks—using diverse LLMs, agent architectures, and calibration methods over both MIMIC-III and IV datasets, and on an external in-hospital mortality cohort from an independent hospital (TJH). Extensive experiments, supported by theoretical motivation under assumptions, demonstrate that *C²Align* improves calibration while maintaining or enhancing discrimination. Our work provides a theoretical, methodological, and benchmark reference for developing well-calibrated agentic LLMs for trustworthy clinical prediction [1].

## 1 Introduction

Clinical prediction, namely using patient information to estimate the probability of a clinical phenotype, risk, or outcome, is at the core of clinical decision making [1, 2, 3]. Electronic health records (EHRs), which capture rich longitudinal information about patients' health status in real-world practice, with both structured data (e.g., age, race, gender, laboratory results) and unstructured data (e.g., clinical notes, discharge summaries), together with advances in machine learning (ML), deep learning, and generative artificial intelligence, have great potential to revolutionize medicine and healthcare. In particular, large language models (LLMs) excel at extracting information from and reasoning over unstructured clinical notes [4, 5], and agentic LLMs enable multiple LLMs to collaborate and integrate homogeneous or heterogeneous signals into a final decision [6].

However, clinical prediction differs from some general prediction tasks. It requires models to possess not only strong *discrimination* but also to ensure adequate *calibration*, namely, ensuring accurate prediction of absolute risk or probability.[7, 8, 9] A model can have excellent discrimination performance but still mislead clinical decision-making. As shown in Fig. 1(a), suppose the actual

[1]Our code is available at `https://github.com/HelloZicky/C2Align`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

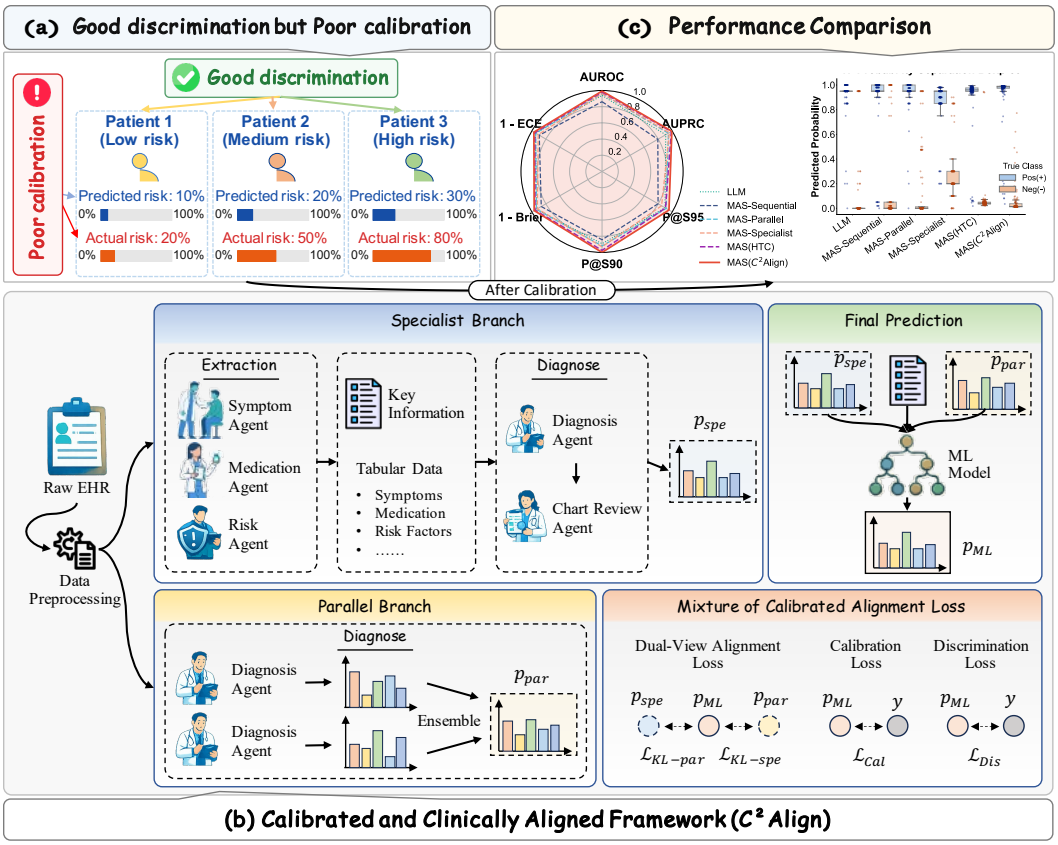


Figure 1: The Calibrated and Clinically Aligned Framework *C²Align* . (a) The clinical calibration gap. Good discrimination does not guarantee good calibration, which can mislead high-stakes decisions (e.g., cardiac surgery or prophylaxis). (b) An example architecture of the proposed *C²Align* framework. (c) Performance comparison. The Radar figure demonstrates that *C²Align* is better-calibrated while maintaining discrimination ability. The boxplot visualizes predicted probabilities. Ideally, positive (blue) and negative (orange) samples cluster near 1 and 0, respectively. *C²Align* yields compact, well-separated distributions and is well-calibrated.

probabilities of disease for three patients are 80%, 50%, and 20%, respectively, while the model predicts 30%, 20%, and 10%. Since the predicted probabilities preserve the same ranking as the actual probabilities, the model may appear perfect in terms of discrimination. However, it is clearly not a perfect model, as the predicted probabilities deviate substantially from the actual probabilities. This is known as poor calibration. Such poor calibration may lead physicians to make inappropriate decisions regarding cardiac surgery for high-risk patients due to a poorly calibrated inpatient mortality risk model [10], or fail to prescribe low-dose aspirin to pregnant women who are actually at high risk for preeclampsia (a life-threatening pregnancy complication for both the mother and the baby) [11]. Thus, the alignment between the predicted probabilities and the actual empirical frequencies of the outcomes across the population is critical for clinical decision-making. Some existing work has evaluated LLMs for clinical prediction [12, 4], and a paucity of work has investigated agent workflows for some disease prediction tasks [6]. However, none of them investigated or benchmarked their modeling calibration performance, and there is a lack of both theoretical and empirical analyses of using LLMs and agentic LLMs for clinical prediction tasks using real-world EHR data. One existing work explores agent calibration [13], however, by focusing on token probabilities returned by agent trajectories it performs poorly in knowledge-intensive clinical predictions. Furthermore, its applicability decreases for agentic LLMs with parallel or more complex structures. Research on the calibration of LLMs and agentic workflows for clinical prediction remains largely unexplored.

In this work, we theoretically analyze how different agentic LLM architectures, task knowledge, and ML tool use influence calibration using generalization bound theory [14, 15], with the analysis being conditional on the assumptions of each theorem. We start with typical agentic LLM architectures with a sequential chain structure, a parallel structure, their mixture designed by domain knowledge,

and using ML tools, and derive and compare their calibration generalization bounds. We find that: 1) controlling for a comparable discrimination performance (w.r.t. Area Under the Receiver Operating Characteristic Curve or C statistics), parallel agentic architectures potentially optimize calibration better than sequential ones; 2) incorporating knowledge of clinical task potentially improves calibration; and 3) tool use of machine learning (ML) models potentially further improves calibration, and use of for example interpretable ML, making it highly suitable for clinical settings. Guided by this theory, we propose *C²Align*, a Calibrated and Clinically *Align*ed Framework that aligns Parallel LLM branch and domain-specific branches via a multi-objective alignment loss, jointly optimizing discrimination and calibration. Specifically, *C²Align* first utilizes a homogeneous population of parallel LLM agents to obtain robust baseline predictive distributions, while simultaneously employing a heterogeneous system of specialist agents for deep clinical reasoning to extract fine-grained diagnostic probabilities. Subsequently, we design a mixture of calibrated alignment loss incorporating a discrimination loss, a calibration loss, and a dual-view knowledge alignment loss, aligning the soft labels generated by both agent architectures into an efficient, lightweight, and interpretable ML model. This design not only inherits the agentic LLM's comprehension of complex clinical texts but also mitigates the overconfidence and miscalibration prevalent in native LLM probabilities, achieving better calibration without sacrificing discrimination.

Empirically, we evaluate *C²Align* across 4 prognostic and diagnostic benchmarks derived from MIMIC-III and MIMIC-IV, including prognostic prediction (30-day readmission) and disease diagnosis or phenotyping (Alzheimer's disease and related dementias (ADRD), and Heart Failure (HF)), and further on in-hospital mortality from an independent hospital (TJH). Comparing multiple LLM backbones, architectures, and calibration methods, extensive experiments demonstrate that *C²Align* consistently performs better in both discrimination and calibration.

In summary, we systematically explore agentic LLM calibration in clinical prediction settings by providing theoretical analysis, calibration methods, and comparing several benchmarks. The main contributions of this paper are summarized as follows:

- We highlight and mathematically formulate the calibration problem of agentic LLMs with focus on clinical prediction settings—a critically important yet largely unexplored domain.
- We provide a conditional theoretical analysis of calibration trade-offs under agentic structures, knowledge incorporation, and ML tool use, aiming to offer a principled guideline for designing well-calibrated agentic LLMs for clinical tasks.
- We propose *C²Align*, a Calibrated and Clinically Aligned Framework that aligns two complementary agent branches: an ensemble of task-homogeneous LLM agents and a set of clinical subtask-inspired LLM agents, achieved through optimizing a novel discrimination- and calibration-aware loss that encourages clinically meaningful agreement while controlling miscalibration.
- We construct four benchmarks based on MIMIC-III and MIMIC-IV, covering diverse clinical prediction tasks including prognosis and disease diagnosis, and add an external validation on TJH. Extensive experiments across multiple LLM backbones, architectures, and calibration methods demonstrate that *C²Align* achieves better calibration while maintaining discrimination.

## 2 Related Work

**LLM for Clinical Prediction.** Clinical prediction has traditionally relied on structured EHRs processed by deep sequence models [1, 16, 2, 17] or disease-specific natural language processing (NLP) pipelines [18]. Recently, medical LLMs have been adapted for unstructured clinical phenotyping, risk prediction, and automated chart review [19, 20, 21, 22, 23, 24, 25, 26, 27, 4]. Subsequently, a few studies have explored how agentic LLMs can better extend the capabilities of LLMs to clinical prediction [6]. However, they all overlook a critical issue: clinical prediction requires models to possess not only strong discriminative capability but also excellent calibration, a dual requirement that is often neglected in current approaches.

**LLM Calibration.** Despite their advanced reasoning ability, LLMs typically suffer from overconfidence, producing poorly calibrated token probabilities. Traditional confidence calibration methods—such as Platt Scaling [28], Isotonic Regression [29], and BBQ [30]—were originally designed for the continuous logits of conventional discriminative models. Applying them directly to the generative token probabilities of LLMs often yields suboptimal alignment. While recent research has proposed agentic confidence calibration to mitigate hallucination and overconfidence in

general question-answering tasks [13], calibrating the complex, longitudinal reasoning required for high-stakes clinical tasks remains an open challenge.

**Distinctions of Our Work.** Existing clinical prediction pipelines face a persistent dilemma: traditional ML models provide well-calibrated risk estimates but struggle with unstructured narratives, whereas LLMs parse complex texts but suffer from poorly calibrated probabilities. Furthermore, prior clinical LLM studies predominantly optimize discrimination while neglecting calibration. In contrast, we mathematically formulate the calibration problem of agentic LLMs in clinical prediction settings, and provide a theoretical analysis of calibration trade-offs under varying discrimination constraints and agentic workflow architectures, offering principled guidelines for designing well-calibrated agentic LLMs for clinical tasks. Our *C²Align* provides conditional theoretical motivation and principled guidance for clinical predictions.

# 3 Methodology

## 3.1 Problem Formulation

Let $\mathbf{h}^{(t)}$ be the latent context state at step $t$, with initialization $\mathbf{h}^{(0)} = \mathcal{X}$, where $\mathcal{X} = \{\mathcal{X}_{\text{struct}}, \mathcal{X}_{\text{unstruct}}\}$ represents the patient's EHR, comprising structured covariates and unstructured narrative notes. We formulate the universal agentic transition as:

$$\mathbf{h}^{(t)} = \mathcal{G}\Big(\{\Phi(\mathbf{h}^{(t-1)};\ \rho_v,\ \epsilon_v)\}_{v\in\mathcal{V}_t}\Big) \tag{1}$$

where $\Phi$ is the base LLM transition kernel, $\mathcal{V}_t$ is the set of active agent nodes at step $t$, $\rho_v$ denotes the specific role prompt for agent $v$, $\epsilon_v$ is the stochastic sampling noise, and $\mathcal{G}(\cdot)$ is the graph topology function that routes and aggregates the agent outputs. The equation 1 encompasses sequential chain structure ($|\mathcal{V}_t| = 1, \mathcal{G} =$ Identity), parallel ensemble of homogeneous LLMs ($\mathcal{G} =$ Mean), and mixture of heterogeneous subtask LLMs ($\mathcal{G} =$ Concatenation) as typical special cases.

Let the clinical prediction task be formulated as estimating the conditional probability distribution $P(\mathbf{Y} \mid \mathcal{X})$ over $K$ distinct clinical outcomes, where $\mathbf{Y} \in \{0,1\}^K$ denotes the one-hot encoded ground truth. In the context of agentic LLMs, calibration focuses on the alignment between the model's output probabilities and the actual empirical frequencies of the outcomes. Let $f \in \mathcal{F}$ denote the predictive mapping of any unified architecture, such that $\mathbf{p}_{\text{pred}} = f(\mathcal{X}) \in \Delta^{K-1}$ is the predicted probability vector, where $\Delta^{K-1} = \{\mathbf{p} \in [0,1]^K : \sum_{k=1}^K p_k = 1\}$ denotes the probability simplex. A model is perfectly calibrated if $\mathbb{E}[\mathbf{Y} \mid f(\mathcal{X}) = \mathbf{c}] = \mathbf{c}$ for all $\mathbf{c} \in \Delta^{K-1}$.

We formulate the main problem as a constrained optimization problem to jointly optimize calibration and discrimination measures. Specifically, we aim to minimize the expected Brier risk $\mathcal{R}_{\text{Brier}}(f)$, subject to a strict information-theoretic discrimination constraint $\eta(>0)$ (e.g., maintaining a required Log-loss boundary):

$$\min_{f\in\mathcal{F}} \mathcal{R}_{\text{Brier}}(f) \quad \text{s.t.} \quad \mathcal{R}_{\text{Dis}}(f) \le \eta \tag{2}$$

where the expected risks are formally defined over the joint distribution $\mathbb{P}(\mathcal{X}, \mathbf{Y})$ as $\mathcal{R}_{\text{Brier}}(f) = \mathbb{E}\big[\frac{1}{K}\|\mathbf{Y} - f(\mathcal{X})\|_2^2\big]$ and $\mathcal{R}_{\text{Dis}}(f) = \mathbb{E}[-\sum_k Y_k \log f(\mathcal{X})_k]$.

### 3.1.1 Theoretical Motivation

Here, we develop a conditional generalization bound theory framework that motivates the design of C²Align and provides theoretical motivations on its discrimination and calibration performance.

To formalize the evaluation, let $\hat{\mathfrak{R}}_N(\mathcal{F}) = \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \sigma_i f(\mathcal{X}_i)]$ denote the empirical Rademacher complexity of the hypothesis class $\mathcal{F}$ over $N$ samples, quantifying the model's capacity to overfit. Detailed mathematical proofs are provided in Appendix F. Our analysis establishes three theorems that motivate our structural choices; each holds only under its stated assumptions and is not a general guarantee. To isolate the effect of calibration-relevant stochasticity, we compare sequential and parallel architectures under a controlled discrimination setting. Specifically, for a fixed discrimination level $\eta$, we decompose the constrained Brier risk into a deterministic discrimination term $A_a(\eta)$ and a stochastic decoding-variance term $V_a$, where $a \in \{\text{seq}, \text{par}\}$. We assume that, after controlling for $\eta$, the sequential architecture does not have a smaller deterministic discrimination term than the parallel architecture, i.e., $A_{\text{seq}}(\eta) \ge A_{\text{par}}(\eta)$. This allows the comparison to focus on the variance contribution to calibration.

**Theorem 1** (Parallel **Structure** Lower Brier Risk Under Controlled Discrimination)**.** *Let $\eta$ denote a fixed discrimination constraint. Under additive independent decoding noise, suppose the constrained Brier risk of an architecture $a \in \{\mathrm{seq}, \mathrm{par}\}$ decomposes as*

$$\inf \mathcal{R}_{\mathrm{Brier}}(f_a(\eta)) = A_a(\eta) + V_a, \tag{3}$$

*where $A_a(\eta)$ is the deterministic discrimination component and $V_a$ is the stochastic decoding-variance component. Under this controlled comparison, sequential agentic architectures accumulate decoding noise across consecutive reasoning steps, resulting in $V_{\mathrm{seq}} = \sigma^2_{\mathrm{seq}} > \sigma^2_0$. Conversely, the homogeneous parallel branch reduces the base stochastic sampling noise to $V_{\mathrm{par}} = \sigma^2_0/M$. Therefore,*

$$\inf \mathcal{R}_{\mathrm{Brier}}(f_{\mathrm{seq}}(\eta)) - \inf \mathcal{R}_{\mathrm{Brier}}(f_{\mathrm{par}}(\eta)) \geq \sigma^2_{\mathrm{seq}} - \frac{\sigma^2_0}{M} \geq 0. \tag{4}$$

This theorem on agentic structures implies that, after controlling for discrimination performance, replacing a cascading chain of reasoners with an ensemble of parallel reasoners improves calibration by shrinking the stochastic sampling-variance component, directly motivating our design of the Homogeneous Parallel Branch ($A_{\text{par}}$). If the $M$ parallel outputs have pairwise error correlation $\rho$, then $V_{\mathrm{par}} = \sigma^2_0[\rho + (1-\rho)/M]$, so the benefit vanishes as $\rho \to 1$.

**Theorem 2** (Clinical Task **Knowledge** Lowers Brier Risk)**.** *Let $\mathbf{u}_{traj}$ denote task-agnostic information derived only from the agentic structure (e.g., intrinsic token likelihoods or sequential generative fluctuations), and $\mathbf{x}_{task}$ denote explicit clinical task-inspired knowledge extracted by domain specialists. Clinically, diagnostic and prognostic decisions are primarily determined by patient-level evidence, such as symptoms, medications, risk factors, laboratory results, and chart-derived findings. We therefore model $\mathbf{x}_{\mathrm{task}}$ as an idealized sufficient summary of outcome-relevant clinical evidence, while $\mathbf{u}_{\mathrm{traj}}$ captures residual trajectory-level fluctuations that do not contain additional information about $\mathbf{Y}$ once $\mathbf{x}_{\mathrm{task}}$ is conditioned on. Assuming $\mathbf{x}_{task}$ forms a causal Markov Blanket (an idealized sufficiency assumption) for the true clinical outcome $\mathbf{Y}$ with respect to $\mathbf{u}_{traj}$ (i.e., $\mathbf{Y} \perp\!\!\!\perp \mathbf{u}_{traj} \mid \mathbf{x}_{task}$), the minimum Brier risk of models only relying on task-agnostic information is lower-bounded by the knowledge-based model plus a non-negative information loss penalty $\Delta_{loss}$:*

$$\inf \mathcal{R}_{Brier}(\mathcal{M}(\mathbf{u}_{traj})) = \inf \mathcal{R}_{Brier}(\mathcal{M}(\mathbf{x}_{task})) + \underbrace{\mathbb{E}_{\mathbf{u}_{traj}}\Big[Var_{\mathbf{x}_{task}|\mathbf{u}_{traj}}\big(\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{task}]\big)\Big]}_{\Delta_{loss} \geq 0} \tag{5}$$

This theorem on clinical task knowledge implies that using clinical task-inspired knowledge helps calibration better than only using task-agnostic information. Consequently, only relying on internal LLM token fluctuations may leave additional diagnostic uncertainty when the clinical evidence in $\mathbf{x}_{\mathrm{task}}$ is not recoverable from $\mathbf{u}_{\mathrm{traj}}$. Incorporating the Heterogeneous Specialist Branch to extract explicit domain semantics can help calibration.

**Theorem 3** (ML **Tool Use** and Alignment Tighten Generalization Bounds)**.** *Let $\mathcal{R}(f)$ denote the expected generalization error. A frozen zero-shot agentic LLM ($f_{LLM}$) has zero empirical Rademacher complexity ($\hat{\mathfrak{R}}_N = 0$) but suffers from a high structural bias (Approximation Error):* $\inf \mathcal{R}_{Brier}(f_{LLM})$.

*Appending a parameterized ML tool $\mathcal{M}$ lowers this approximation error to $A_{\mathcal{M}}(\eta) \geq \mathcal{C}_{bound}(\eta)$ (Lemma 5), yielding a bias reduction $\Delta_{approx} = \inf \mathcal{R}_{Brier}(f_{LLM}) - A_{\mathcal{M}}(\eta) \geq 0$ if the class of $\mathcal{M}$ can match $f_{LLM}$. Optimizing $\mathcal{M}$ via hard-label cross-entropy expands the parameter space to a maximal norm $C_w$, introducing a positive estimation error penalty $\Delta_{est} = 4\frac{C_w R_x}{\sqrt{N}}$ (where $R_x = \max_i \|\mathbf{x}_i\|_2$, for norm-bounded linear $\mathcal{M}$ (Appendix F.3)).*

*Co-optimizing the dual Kullback–Leibler (KL) divergence ($\mathcal{L}_{Align}$) with agentic soft targets restricts logit inflation, compressing the parameter radius to $C^*_w \leq C_w$, so the dual-KL term can further improve calibration. This requires that the agentic LLM branches provide non-adversarial and outcome-informative clinical representations or soft targets, so that the downstream ML tool learns from an enriched representation rather than from scratch.*

This theorem on using ML tools implies that appending an ML-based calibration layer through agentic tool use reduces the structural bias. Further encouraging the alignment of ML and agent workflows (via dual-KL soft targets) can help reduce the Rademacher complexity penalty, thereby yielding tighter generalization bounds. Furthermore, lightweight and interpretable ML tools are particularly suitable for clinical prediction tasks.

### 3.2 Calibrated and Clinically Aligned Framework (C²Align)

Motivated by the above three theorems, we propose *C²Align*, a Calibrated and Clinically Aligned framework, to optimize calibration while maintaining or even improving discrimination. As shown in Figure 1b, *C²Align* runs in six steps: (1) frozen parallel agents return a probability $\mathbf{p}_{\text{par}}$; (2) frozen specialist agents extract fixed task indicators $\mathbf{x}_{\text{spe}}$ and a probability $\mathbf{p}_{\text{spe}}$; (3) both outputs are joined with structured covariates by patient ID and schema-validated; (4) a task-level allowlist builds the feature vector, excluding labels, annotations, outcomes, post-index fields, and identifiers; (5) only the downstream XGBoost model is trained; (6) $\mathbf{p}_{\text{ML}}$ is evaluated. The details are as follows.

#### 3.2.1 Mixture of Dual-branch Agents

We formalize the diagnostic extraction as a dual mapping function $\Phi_{\text{dual}} : \mathcal{X} = \{\mathcal{X}_{\text{struct}}, \mathcal{X}_{\text{unstruct}}\} \to (\mathbf{p}_{\text{par}}, \mathbf{p}_{\text{spe}}, \mathbf{x}_{\text{spe}})$, comprising two concurrent streams:

**Homogeneous Parallel Branch ($A_{\text{par}}$, frozen):** This branch consists of $M$ independent, homogeneous stochastic agent policies $\pi_{\text{par}}^{(m)}$. Rather than relying on a single, potentially biased generation, $A_{\text{par}}$ establishes a robust baseline consensus distribution $\mathbf{p}_{\text{par}} \in \Delta^{K-1}$ via independent Monte Carlo ensemble averaging:

$$\mathbf{p}_{\text{par}} = \frac{1}{M} \sum_{m=1}^{M} \pi_{\text{par}}^{(m)}(\mathbf{Y} \mid \mathcal{X}_{\text{struct}}, \mathcal{X}_{\text{unstruct}}) \tag{6}$$

This independent ensemble averaging operation effectively smooths predictive variance and mitigates idiosyncratic LLM overconfidence.

**Heterogeneous Specialist Branch ($A_{\text{spe}}$, frozen):** To ensure depth of clinical reasoning, this branch executes domain-specific evidence extraction across specialized semantic sub-spaces $\mathcal{S} = \{\text{symptom, intervention, risk}\}$. Let $f_s$ denote the specialized extractor agent for domain $s \in \mathcal{S}$, which parses unstructured text into explicit clinical indicators (e.g., symptom severity, medication risk flags). Their fields are fixed by a task-level JSON schema (Appendix C). The extracted structured clinical feature manifold $\mathbf{x}_{\text{spe}} \in \mathbb{R}^d$ is formulated via the direct sum of the sub-space embeddings: $\mathbf{x}_{\text{spe}} = \bigoplus_{s \in \mathcal{S}} f_s(\mathcal{X}_{\text{struct}}, \mathcal{X}_{\text{unstruct}})$, where $\bigoplus$ denotes the vector concatenation operator. Subsequently, a high-level diagnostic agent $\pi_{\text{spe}}$ aggregates this evidence manifold to deduce a fine-grained, specialist-driven probability distribution:

$$\mathbf{p}_{\text{spe}} = \pi_{\text{spe}}(\mathbf{Y} \mid \mathbf{x}_{\text{spe}}) \in \Delta^{K-1} \tag{7}$$

#### 3.2.2 Mixture of Calibrated Alignment Loss

To synthesize the dual-view insights, we employ a lightweight, interpretable ML model $\mathcal{M}$ (e.g., a gradient boosted tree), parameterized by $\theta$. The input to $\mathcal{M}$ is a sophisticated augmented feature manifold $\mathbf{x}_{\text{aug}}$ that encapsulates both the detailed clinical feature space and the multi-dimensional cognitive probabilities. The comprehensive input manifold is constructed by concatenating the raw structured covariates $\mathbf{x}_{\text{struct}}$, the specialist-extracted clinical evidence $\mathbf{x}_{\text{spe}}$, and the soft-label probabilities generated by both agent branches: $\mathbf{x}_{\text{aug}} = \mathbf{x}_{\text{struct}} \oplus \mathbf{x}_{\text{spe}} \oplus \mathbf{p}_{\text{par}} \oplus \mathbf{p}_{\text{spe}}$

By explicitly embedding the dual-view probabilities directly into the ML model's input space alongside the evidence manifold, the architecture dynamically infuses multi-agent generalization boundaries into the final decision process to yield the ultimate prediction $\mathbf{p}_{\text{ML}} = \mathcal{M}(\mathbf{x}_{\text{aug}}; \theta) \in \Delta^{K-1}$. To guarantee that the final estimated probabilities $\mathbf{p}_{\text{ML}}$ are both highly discriminative and empirically trustworthy, we define a custom multi-objective loss space incorporating three rigorously constrained variables evaluated over $N$ training samples.

**Discrimination Loss ($\mathcal{L}_{\text{Dis}}$).** Operating as the foundational predictive penalty, $\mathcal{L}_{\text{Dis}}$ enforces the decision-making accuracy of the model against the definitive ground truth $\mathbf{y}^{(i)}$. We optimize this via standard Cross-Entropy Loss: $\mathcal{L}_{\text{Dis}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log([\mathbf{p}_{\text{ML}}^{(i)}]_k)$, where $y_{i,k}$ is the $k$-th element of the true label vector for sample $i$, and $[\mathbf{p}_{\text{ML}}^{(i)}]_k$ is the corresponding predicted probability.

**Calibration Loss ($\mathcal{L}_{\text{Cal}}$).** Operating as a proper scoring rule, the calibration loss penalizes overconfidence and bounds the probability dispersion via the Brier Score computation:

$$\mathcal{L}_{\text{Cal}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \left( [\mathbf{p}_{\text{ML}}^{(i)}]_k - y_{i,k} \right)^2 \tag{8}$$

**Dual-View Alignment Loss ($\mathcal{L}_{\text{Align}}$).** To align the wisdom of the dual-architecture system, we propose $\mathcal{L}_{\text{Align}}$ to actively minimize the KL divergence between the ML model's output and the agentic priors. Acting as the soft-target regularizer, we formulate this dual injection by aligning $\mathbf{p}_{\text{ML}}$ concurrently with the parallel ensemble baseline $\mathbf{p}_{\text{par}}$ and the specialized deep-reasoning distribution $\mathbf{p}_{\text{spe}}$:

$$\mathcal{L}_{\text{Align}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\beta_{\text{par}}[\mathbf{p}_{\text{par}}^{(i)}]_k \log\left(\frac{[\mathbf{p}_{\text{par}}^{(i)}]_k}{[\mathbf{p}_{\text{ML}}^{(i)}]_k}\right) + \beta_{\text{spe}}[\mathbf{p}_{\text{spe}}^{(i)}]_k \log\left(\frac{[\mathbf{p}_{\text{spe}}^{(i)}]_k}{[\mathbf{p}_{\text{ML}}^{(i)}]_k}\right)\right] \tag{9}$$

where $\beta_{\text{par}}$ and $\beta_{\text{spe}}$ control the distillation intensity from the respective agent branches.

The final optimal parameters $\theta^*$ are learned by minimizing the holistic loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dis}} + \alpha\mathcal{L}_{\text{Cal}} + \mathcal{L}_{\text{Align}} \tag{10}$$

# 4 Experiment

## 4.1 Experimental Setup

**Baselines.** To evaluate our proposed framework, we compare it with a diverse set of baseline models and calibration techniques. (1) ***Model Selection.*** We include (i) ML models: Logistic Regression (LR) and XGBoost [31]. These models are restricted to taking structured features as inputs, as they inherently struggle to process extremely long unstructured texts and fail to capture the highly sparse, fine-grained features embedded within them. (ii) LLMs: We evaluate both closed-source and open-source models, including Qwen3-Max [32], ChatGPT-4o [33, 34], Gemini-2.5 [35], DeepSeek-V4 [36], and LLaMA-3.1-8B [37]. (2) ***Architecture of Multi-Agent System (MAS).*** (i) Sequential: A chain-of-reasoning pipeline where agents process clinical notes iteratively. (ii) Parallel: An ensemble of independent agents operating simultaneously. (3) ***Calibration Methods.*** We evaluate *C²Align* against traditional post-hoc scaling methods and recent agentic calibration methods: (i) Post-hoc Calibration: Platt Scaling [28], Isotonic Regression [29], and BBQ [30], reporting the best-performing method for each baseline. However, these methods merely rescale terminal probabilities, failing to resolve the underlying LLM hallucinations or improve the intrinsic semantic representations. (ii) Trajectory Calibration ("MAS Calibration" in the tables, MAS(HTC) in Figure 2): agentic Confidence Calibration [13], which represents one of the latest and very few works focusing on agentic calibration by utilizing execution trajectories to estimate reliability. Nevertheless, it relies on fluctuations in token probabilities, fundamentally ignoring the critical impact of explicit domain knowledge on calibration and struggling with complex clinical reasoning.

**Datasets.** We evaluate our method on four primary benchmarks derived from MIMIC-III [38] and MIMIC-IV [39], with external validation on an in-hospital mortality cohort from TJH [40], covering disease diagnosis and prognostic prediction. (1) *Diagnosis and Phenotyping.* To identify ADRD and HF patients from MIMIC-IV, we extracted inpatient admissions with and without target diseases via ICD codes. Two clinicians (plus a senior expert for discrepancies) manually annotated. Excluding records whose evidence remained ambiguous or insufficient for a confident label after adjudication (ADRD: 50; HF: 27), the finalized records were partitioned into train/test sets for ADRD (712/238, total 950) and HF (730/243, total 973). These are retrospective chart-based phenotyping tasks rather than forecasts of future onset. (2) *Prognostic Prediction.* This category encompasses 30-day unplanned readmission based on both MIMIC-III (Readmission (III)) and MIMIC-IV (Readmission (IV)). For each task, we constructed a cohort of 1,000 cases, which were randomly allocated into 800 samples for training and 200 samples for testing. All inputs end at discharge from the index hospitalization. We also performed mortality prediction on the TJH dataset. Settings are the same as those for MIMIC.

**Metrics.** We evaluate on discrimination and calibration metrics. (1) *Discrimination Metrics.* We measure discrimination using Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), and positive predictive value (PPV) at a sensitivity of 0.95 (P@S95). (2) *Calibration Metrics.* We evaluate probability alignment visually via Reliability Curves and quantitatively using two key metrics: the Brier Score ($BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$, where $\hat{p}_i$ is the predicted probability and $y_i$ is the true label) and the Expected Calibration Error ($ECE = \sum_{m=1}^{M}\frac{|B_m|}{N}|\text{acc}(B_m) - \text{conf}(B_m)|$, where $B_m$ is the $m$-th bin out of $M = 10$. For

Table 1: Comparison of our method with baselines across multiple benchmarks (Qwen3-Max backbone). *Improvement*: absolute (relative) gain of C²Align over the strongest baseline for each metric; for BS and ECE it is a reduction.

| | Model | AUROC↑ | AUPRC↑ | P@S95↑ | BS↓ | ECE↓ |
|---|---|---|---|---|---|---|
| | *Heart Failure Diagnosis and Phenotyping* | | | | | |
| ML | LR | 0.9191 | 0.9250 | 0.6920 | 0.1029 | 0.0498 |
| | XGBoost | 0.9157 | 0.9240 | 0.6415 | 0.0957 | 0.0528 |
| LLM | Base | 0.9386 | 0.9069 | 0.8967 | 0.0670 | 0.0535 |
| | Calibration | 0.9386 | 0.9069 | 0.8967 | 0.0689 | 0.0545 |
| MAS | Sequential | 0.8619 | 0.7943 | 0.7977 | 0.1345 | 0.1059 |
| | Parallel | 0.9614 | 0.9449 | 0.9130 | 0.0657 | 0.0547 |
| | Calibration | 0.9806 | 0.9865 | 0.9034 | 0.0459 | 0.0428 |
| | **Ours** | **0.9927** | **0.9940** | **0.9530** | **0.0336** | **0.0294** |
| | *Improvement* | *+.0121 (1.23%)* | *+.0075 (0.76%)* | *+.0400 (4.38%)* | *+.0123 (26.80%)* | *+.0134 (31.31%)* |
| | *Alzheimer's Disease and Related Dementias (ADRD) Diagnosis and Phenotyping* | | | | | |
| ML | LR | 0.9357 | 0.9589 | 0.6482 | 0.0912 | 0.0464 |
| | XGBoost | 0.9372 | 0.9562 | 0.7247 | 0.0901 | 0.0569 |
| LLM | Base | 0.9680 | 0.9707 | 0.8944 | 0.0523 | 0.0868 |
| | Calibration | 0.9680 | 0.9707 | 0.8944 | 0.0412 | 0.0280 |
| MAS | Sequential | 0.9501 | 0.9523 | 0.8567 | 0.1006 | 0.0821 |
| | Parallel | 0.9727 | 0.9747 | 0.9604 | 0.0505 | 0.0841 |
| | Calibration | 0.9804 | 0.9855 | 0.8837 | 0.0641 | 0.0900 |
| | **Ours** | **0.9962** | **0.9971** | **1.0000** | **0.0210** | **0.0273** |
| | *Improvement* | *+.0158 (1.61%)* | *+.0116 (1.18%)* | *+.0396 (4.12%)* | *+.0202 (49.03%)* | *+.0007 (2.50%)* |
| | *Readmission (IV)* | | | | | |
| ML | LR | 0.5540 | 0.5821 | 0.5360 | 0.4016 | 0.3831 |
| | XGBoost | 0.6394 | 0.6344 | 0.5505 | 0.2638 | 0.1657 |
| LLM | Base | 0.5788 | 0.6058 | 0.5285 | 0.3533 | 0.3220 |
| | Calibration | 0.4212 | 0.5317 | 0.5280 | 0.2509 | 0.1400 |
| MAS | Sequential | 0.5842 | 0.6250 | 0.5285 | 0.3691 | 0.3358 |
| | Parallel | 0.5874 | 0.6439 | 0.5285 | 0.3477 | 0.3048 |
| | Calibration | 0.5805 | 0.6023 | 0.5505 | 0.2456 | 0.1304 |
| | **Ours** | **0.7128** | **0.7098** | **0.6076** | **0.2210** | **0.1214** |
| | *Improvement* | *+.0734 (11.48%)* | *+.0659 (10.23%)* | *+.0571 (10.37%)* | *+.0246 (10.02%)* | *+.0090 (6.90%)* |
| | *Readmission (III)* | | | | | |
| ML | LR | 0.5240 | 0.3126 | 0.2803 | 0.3277 | 0.3100 |
| | XGBoost | 0.5864 | 0.3255 | 0.3064 | 0.2467 | 0.1802 |
| LLM | Base | 0.6331 | 0.3675 | 0.3444 | 0.3670 | 0.4012 |
| | Calibration | 0.6331 | 0.3675 | 0.3444 | 0.1882 | 0.0830 |
| MAS | Sequential | 0.6420 | 0.3809 | 0.3352 | 0.3875 | 0.4228 |
| | Parallel | 0.6414 | 0.3738 | 0.3408 | 0.3833 | 0.4184 |
| | Calibration | 0.4583 | 0.2643 | 0.2775 | 0.2216 | 0.1131 |
| | **Ours** | **0.6674** | **0.4068** | **0.3598** | **0.1851** | **0.0702** |
| | *Improvement* | *+.0254 (3.96%)* | *+.0259 (6.80%)* | *+.0154 (4.47%)* | *+.0031 (1.65%)* | *+.0128 (15.42%)* |

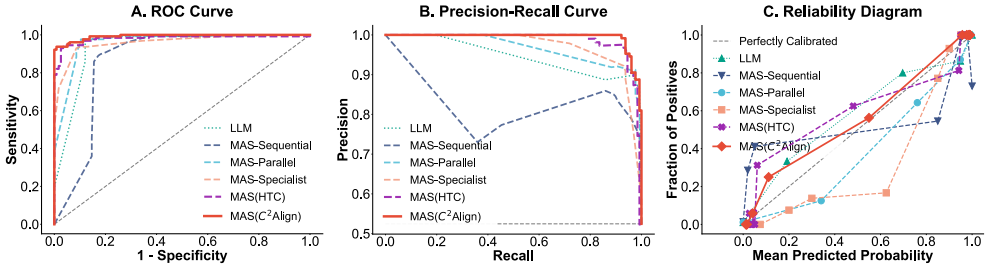


Figure 2: Performance evaluation of various architectures using Qwen3-Max on the HF benchmark. (A) ROC Curve. (B) Precision-Recall Curve (Precision=PPV, Recall=Sensitivity). (C) Reliability Diagram (10 equal-mass bins). *C²Align* tightly adheres to the perfectly calibrated diagonal while simultaneously dominating both discrimination curves.

consistency, the visual Reliability Diagrams were also generated using 10 equal-frequency quantile bins. Every system outputs a continuous probability per patient; bins are used only for ECE and reliability diagrams, and equal-mass bins reduce estimation bias [41]. We also report calibration intercept and slope [9] (Appendix D). For results reported with 95% confidence intervals (CIs), values in square brackets give the lower and upper bounds, in that order.

### 4.2 Main Results

Table 1 and Figure 2 present the performance of C²Align against baselines across four clinical benchmarks, yielding three insights into agentic architectures. (1) **LLMs vs. Traditional ML.** A persistent dilemma exists: traditional ML (e.g., XGBoost) exhibits reasonable calibration (ADRD ECE: 0.0569) but lower discrimination (AUROC: 0.9372). Conversely, Base LLMs achieve higher discrimination

Table 2: Comparison of our method with baselines across different foundation models with 95% CIs.

| | Model | AUROC↑ | BS↓ | ECE↓ |
|---|---|---|---|---|
| | *ChatGPT-4o* | | | |
| LLM | Base | 0.6277 [0.5507, 0.7047] | 0.3610 [0.2944, 0.4276] | 0.3250 [0.2557, 0.3943] |
| | Calibration | 0.6280 [0.5510, 0.7050] | 0.2417 [0.1824, 0.3010] | 0.1590 [0.0897, 0.2283] |
| MAS | Sequential | 0.6009 [0.5227, 0.6791] | 0.3882 [0.3207, 0.4557] | 0.3473 [0.2780, 0.4166] |
| | Parallel | 0.6426 [0.5664, 0.7188] | 0.2947 [0.2315, 0.3579] | 0.2326 [0.1633, 0.3019] |
| | Calibration | 0.6075 [0.5296, 0.6854] | 0.2241 [0.1663, 0.2819] | 0.1147 [0.0454, 0.1840] |
| | **Ours** | **0.6548** [0.5793, 0.7303] | **0.1921** [0.1375, 0.2467] | **0.0511** [0.0000, 0.1204] |
| | *Deepseek-V4* | | | |
| LLM | Base | 0.6556 [0.5802, 0.7310] | 0.2959 [0.2326, 0.3592] | 0.2601 [0.1908, 0.3294] |
| | Calibration | 0.6558 [0.5804, 0.7312] | 0.2417 [0.1824, 0.3010] | 0.1499 [0.0806, 0.2192] |
| MAS | Sequential | 0.6426 [0.5664, 0.7188] | 0.3155 [0.2511, 0.3799] | 0.2833 [0.2140, 0.3526] |
| | Parallel | 0.6689 [0.5944, 0.7434] | 0.2626 [0.2016, 0.3236] | 0.1822 [0.1129, 0.2515] |
| | Calibration | 0.5547 [0.4751, 0.6343] | 0.2488 [0.1889, 0.3087] | **0.0564** [0.0000, 0.1257] |
| | **Ours** | **0.6707** [0.5963, 0.7451] | **0.1943** [0.1395, 0.2491] | 0.1005 [0.0312, 0.1698] |

| | Model | AUROC↑ | BS↓ | ECE↓ |
|---|---|---|---|---|
| | *Gemini-2.5* | | | |
| LLM | Base | 0.5723 [0.4931, 0.6515] | 0.3815 [0.3142, 0.4488] | 0.3435 [0.2742, 0.4128] |
| | Calibration | 0.4344 [0.3550, 0.5138] | 0.2509 [0.1908, 0.3110] | 0.1515 [0.0822, 0.2208] |
| MAS | Sequential | 0.6114 [0.5336, 0.6892] | 0.3938 [0.3261, 0.4615] | 0.3514 [0.2821, 0.4207] |
| | Parallel | 0.6019 [0.5237, 0.6801] | 0.3753 [0.3082, 0.4424] | 0.3538 [0.2845, 0.4231] |
| | Calibration | 0.5397 [0.4598, 0.6196] | 0.2525 [0.1923, 0.3127] | 0.1052 [0.0359, 0.1745] |
| | **Ours** | **0.6420** [0.5658, 0.7182] | **0.2029** [0.1472, 0.2586] | **0.1048** [0.0355, 0.1741] |
| | *LLaMA-3.1-8B* | | | |
| LLM | Base | 0.5930 [0.5145, 0.6715] | 0.3075 [0.2435, 0.3715] | 0.2724 [0.2031, 0.3417] |
| | Calibration | 0.5842 [0.5054, 0.6630] | 0.2433 [0.1838, 0.3028] | 0.0997 [0.0304, 0.1690] |
| MAS | Sequential | 0.6235 [0.5463, 0.7007] | 0.3330 [0.2677, 0.3983] | 0.3027 [0.2334, 0.3720] |
| | Parallel | 0.5933 [0.5148, 0.6718] | 0.3081 [0.2441, 0.3721] | 0.2455 [0.1762, 0.3148] |
| | Calibration | 0.4953 [0.4151, 0.5755] | 0.2495 [0.1895, 0.3095] | 0.0689 [0.0000, 0.1382] |
| | **Ours** | **0.6483** [0.5724, 0.7242] | **0.1948** [0.1399, 0.2497] | **0.0536** [0.0000, 0.1229] |

(AUROC: 0.9680) but suffer from overconfidence (ECE: 0.0868), highlighting the need to bridge text comprehension with statistical reliability. (2) **Agentic Architectures and Calibration Trade-offs.** Agentic baselines are consistent with our conditional analysis: (i) *Parallel* architectures achieve lower BS than *Sequential* architectures on all four benchmarks and lower ECE on three (ADRD is the exception), consistent with the view that ensemble averaging mitigates generative noise for better probability estimation (Theorem 1). (ii) Trajectory-based calibration can collapse discrimination (e.g., *Readmission III* AUROC drops to 0.4583). Conversely, C²Align injects clinical semantics ($\mathbf{x}_{\text{spe}}$) to resolve diagnostic uncertainty, achieving a 0.6674 AUROC and lower BS (Theorem 2). (iii) Via dual-architecture distillation, C²Align regularizes the ML head parameter space, concurrently maximizing discrimination (HF AUROC: 0.9927; Figure 2A) and minimizing calibration errors (HF ECE: 0.0294; Figure 2C) (Theorem 3). (3) **Overall Clinical Utility.** C²Align preserves performance at stringent operating points (P@S95), as visually corroborated by its dominance in the high-sensitivity region of the Precision-Recall curve (Figure 2B). Compared to the strongest competing baselines, it yields average relative improvements of 4.57% (AUROC), 4.74% (AUPRC), and 5.84% (P@S95), alongside reductions of 21.88% (BS) and 14.03% (ECE). This confirms its robustness and reliability for clinical prediction.

### 4.3 Generalization Across Foundation Models

To evaluate backbone-agnosticism, we test C²Align across proprietary (ChatGPT-4o, Gemini-2.5) and open-weight (Deepseek-V4, LLaMA-3.1-8B) models on *Readmission (IV)* (Table 2). C²Align maintains its structural superiority regardless of LLM capacity and find that: (1) **Mitigating the Calibration-Discrimination Trade-off.** Existing calibration methods frequently degrade discrimination (e.g., trajectory calibration collapses AUROC to 0.4953 in LLaMA-3.1-8B and 0.5547 in Deepseek-V4). This is consistent with Theorem 2: relying on statistical rescaling or token fluctuations without explicit clinical semantics fails to resolve diagnostic uncertainty. Conversely, C²Align breaks this trade-off, achieving the highest AUROC across all backbones while minimizing the BS. (2) **Consistency of Architectural Theorems.** Cross-model results are consistent with our analysis: (i) *Parallel* architectures yield lower BS than *Sequential* baselines for all four backbones and lower ECE for three (Gemini-2.5 is the exception), consistent with generative variance reduction (Theorem 1). (ii) Distilling parallel distributions and specialist semantics into the ML head is consistent with the tighter bound (Theorem 3). Even when trajectory calibration achieves a lower ECE on Deepseek-V4 (0.0564 vs. 0.1005), it sacrifices 11.6 AUROC points (0.5547 vs. 0.6707). C²Align provides a robust balance between discrimination and probabilistic reliability across all foundation models.

### 4.4 External Validation

For a concise external-validation summary, Table 3 reports TJH results for two selected backbones, ChatGPT-4o and DeepSeek-V4. Every method uses the same structured inputs. On ChatGPT-4o, C²Align has the highest AUROC and lowest BS and ECE. On DeepSeek-V4, it has the lowest BS and ties for the highest AUROC, while four baselines have lower ECE (lowest: MAS Parallel, 0.0159 vs. 0.0208). Given the small cohort and overlapping CIs, these results should be interpreted descriptively.

### 4.5 Ablation Study

We ablate the Mixture of Dual-Branch Agents (MoDBA) and Mixture of Calibrated Alignment Loss (MoCAL) on the *Readmission (IV)* benchmark (Table 4) and find that: (1) **Impact of MoCAL.** Direct LLM inference (Rows 1–2) yields suboptimal calibration and discrimination. By distilling agentic

Table 3: Selected TJH in-hospital mortality results for ChatGPT-4o and Deepseek-V4 with 95% CIs. Best results within each backbone are bolded, including ties.

| Backbone | Model | | AUROC↑ | BS↓ | ECE↓ |
|---|---|---|---|---|---|
| ChatGPT-4o | LLM | Base | 0.9897 [0.9720, 1.0000] | 0.0268 [0.0026, 0.0547] | 0.0351 [0.0087, 0.0586] |
| | | Calibration | 0.9881 [0.9691, 1.0000] | 0.0267 [0.0022, 0.0538] | 0.0426 [0.0216, 0.0723] |
| | MAS | Sequential | 0.9905 [0.9732, 1.0000] | 0.0269 [0.0024, 0.0550] | 0.0356 [0.0084, 0.0588] |
| | | Parallel | 0.9897 [0.9720, 1.0000] | 0.0269 [0.0025, 0.0552] | 0.0353 [0.0084, 0.0588] |
| | | Calibration | 0.9913 [0.9746, 1.0000] | 0.0263 [0.0019, 0.0535] | 0.0412 [0.0189, 0.0683] |
| | | **Ours** | **0.9944** [0.9830, 1.0000] | **0.0242** [0.0020, 0.0497] | **0.0165** [0.0084, 0.0503] |
| Deepseek-V4 | LLM | Base | 0.9905 [0.9728, 1.0000] | 0.0277 [0.0021, 0.0578] | 0.0356 [0.0085, 0.0582] |
| | | Calibration | 0.9889 [0.9691, 1.0000] | 0.0260 [0.0015, 0.0536] | 0.0378 [0.0160, 0.0665] |
| | MAS | Sequential | 0.9897 [0.9714, 1.0000] | 0.0260 [0.0039, 0.0523] | 0.0183 [0.0093, 0.0561] |
| | | Parallel | **0.9921** [0.9755, 1.0000] | 0.0245 [0.0015, 0.0510] | **0.0159** [0.0081, 0.0534] |
| | | Calibration | 0.9905 [0.9736, 1.0000] | 0.0249 [0.0032, 0.0501] | 0.0206 [0.0150, 0.0576] |
| | | **Ours** | **0.9921** [0.9768, 1.0000] | **0.0239** [0.0016, 0.0497] | 0.0208 [0.0172, 0.0604] |

Table 4: Ablation study of the proposed framework.

| **MoDBA** (Parallel) | **MoDBA** (Specialist) | **MoCAL** | AUROC↑ | AUPRC↑ | P@S95↑ | BS↓ | ECE↓ |
|---|---|---|---|---|---|---|---|
| ✓ | ✗ | ✗ | 0.5874 | 0.6439 | 0.5285 | 0.3477 | 0.3048 |
| ✗ | ✓ | ✗ | 0.6157 | 0.6324 | 0.5280 | 0.3621 | 0.3228 |
| ✓ | ✗ | ✓ | 0.7124 | 0.7052 | 0.5676 | 0.2253 | 0.1238 |
| ✗ | ✓ | ✓ | 0.7060 | 0.6894 | 0.6035 | 0.2248 | 0.1258 |
| ✓ | ✓ | ✓ | **0.7128** | **0.7098** | **0.6076** | **0.2210** | **0.1214** |

priors into the ML model, MoCAL reduces BS and ECE (Rows 3–4), consistent with Theorem 3. (2) **Parallel vs. Specialist Branches.** Independent deployment is consistent with our analysis: the Parallel branch reduces generative variance, yielding better calibration (lower ECE/BS) than the Specialist branch (Row 1 vs. 2; Theorem 1). Conversely, while the Specialist branch exhibits higher calibration variance independently, its explicit clinical semantics drive critical improvements in discrimination (e.g., boosting P@S95 from 0.5676 to 0.6035; Row 3 vs. 4, at a slightly lower AUROC of 0.7060 vs. 0.7124), consistent with Theorem 2. (3) **Synergy of the Full Framework.** The complete C²Align (Row 5) achieves an optimal balance. By utilizing MoCAL to fuse the variance-reduced baseline of the Parallel branch with the fine-grained semantics of the Specialist branch, it mitigates individual architectural weaknesses and achieves the best value on every metric.

## 5 Conclusion

We introduce C²Align, a Calibrated and Clinically Aligned framework to calibrate agentic LLM for clinical prediction. Synergizing a Parallel Branch (variance reduction) and a Specialist Branch (task knowledge extraction), C²Align distills agentic distributions into an interpretable ML tool via a multi-objective loss. Evaluations across four benchmarks spanning prognostic and diagnostic tasks over two real-world clinical datasets, together with an external hospital (TJH), show improved calibration while maintaining or improving discrimination. Crucially, our conditional theoretical analysis of agentic architectures, task knowledge, and tool use provides a principled guidance for designing agentic LLM calibration, with C²Align serving as a concrete instantiation of these guidelines. Future research can explore advanced architectural optimizations guided by this theory. Overall, this work provides an essential theoretical, methodological, and benchmark foundation for developing well-calibrated and trustworthy medical AI.

## 6 Limitations

While C²Align demonstrates robust empirical performance and theoretical motivation, certain limitations exist, including but not limited to: (1) Our study cannot evaluate all architectural permutations under the proposed theoretical framework; C²Align is one concrete instantiation, and future research will explore diverse agentic architectures and a broader range of clinical scenarios. (2) Although our approach is applicable to the field of clinical prediction—a domain characterized by small sample sizes—the ML tool is prone to underfitting when dealing with extremely small samples (e.g., a few dozen instances). (3) Theorems 1–3 hold only under their stated assumptions. (4) Test sets are small, MIMIC-III and MIMIC-IV may share patients, and TJH is a single external cohort.

## Acknowledgments and Disclosure of Funding

C.Z. would like to acknowledge the support from NIH award 1R01HD120978. F.W. would like to acknowledge the support from NSF award 2212175 and NIH awards RF1AG072449, RF1AG084178, R01AG080991, R01AG080624, R01AG076448, R01AG076234.

## References

[1] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, pages 3504–3512, 2016.

[2] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021.

[3] Chengxi Zang, Yu Hou, Edward J Schenck, Zhenxing Xu, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk, Dhruv Khullar, Anna S Nordvig, et al. Identification of risk factors of Long COVID and predictive modeling in the RECOVER EHR cohorts. *Communications Medicine*, 4(1):130, 2024.

[4] Yinghao Zhu, Junyi Gao, Zixiang Wang, Weibin Liao, Xiaochen Zheng, Lifang Liang, Miguel O Bernabeu, Yasha Wang, Lequan Yu, Chengwei Pan, Ewen M. Harrison, and Liantao Ma. ClinicRealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. *npj Digital Medicine*, 9(1):319, 2026.

[5] Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Shuang Zhou, Yuan Luo, Rui Zhang, Danielle S. Bitterman, Fei Wang, and Kai Shu. ClinicalBench: Can LLMs beat traditional ML models in clinical prediction? In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26)*, 2026.

[6] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Khandekar, Nicholas Wan, Xuguang Ai, W John Wilbur, et al. AgentMD: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377, 2025.

[7] Ana Carolina Alba, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, PJ Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*, 318(14):1377–1384, 2017.

[8] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

[9] Ben Van Calster, David J McLernon, Maarten van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

[10] Fabio Barili, Davide Pacini, Francesco Rosato, Maurizio Roberto, Alberto Battisti, Claudio Grossi, Francesco Alamanni, Roberto Di Bartolomeo, and Alessandro Parolari. In-hospital mortality risk assessment in elective and non-elective cardiac surgery: a comparison between EuroSCORE II and age, creatinine, ejection fraction score. *European Journal of Cardio-Thoracic Surgery*, 46(1):44–48, 2014.

[11] Haoyang Li, Yaxin Li, Chengxi Zang, Weishen Pan, He S Yang, Tracy B Grossman, Zhen Zhao, and Fei Wang. Machine learning for dynamic and short-term prediction of preeclampsia using routine clinical data. *JAMA Network Open*, 9(3):e260359, 2026.

[12] 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.

[13] Jiaxin Zhang, Caiming Xiong, and Chien-Sheng Wu. Agentic confidence calibration. *arXiv preprint arXiv:2601.15778*, 2026.

[14] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2018.

[15] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[16] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020.

[17] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

[18] Arash A. Nargesi, Philip Adejumo, Lovedeep Singh Dhingra, Benjamin Rosand, Astrid Hengartner, Andreas Coppi, Simon Benigeri, Sounok Sen, Tariq Ahmad, Girish N. Nadkarni, Zhenqiu Lin, Faraz S. Ahmad, Harlan M. Krumholz, and Rohan Khera. Automated identification of heart failure with reduced ejection fraction using deep learning-based natural language processing. *JACC: Heart Failure*, 13(1):75–87, 2025.

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

[20] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

[21] Emily Alsentzer, Matthew J. Rasmussen, Romy Fontoura, Alexis L. Cull, Brett Beaulieu-Jones, Kathryn J. Gray, David W. Bates, and Vesela P. Kovacheva. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *npj Digital Medicine*, 6:212, 2023.

[22] Chao Yan, Henry H Ong, Monika E Grabowska, Matthew S Krantz, Wu-Chen Su, Alyson L Dickson, Josh F Peterson, QiPing Feng, Dan M Roden, C Michael Stein, et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. *Journal of the American Medical Informatics Association*, 31(9):1994–2001, 2024.

[23] Ofir Ben Shoham and Nadav Rappoport. CPLLM: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024.

[24] Jiankun Wang, Sumyeong Ahn, Taykhoom Dalal, Xiaodan Zhang, Weishen Pan, Qiannan Zhang, Bin Chen, Hiroko H. Dodge, Fei Wang, and Jiayu Zhou. Augmented risk prediction for the onset of Alzheimer's disease from electronic health records with large language models. *arXiv preprint arXiv:2405.16413*, 2024.

[25] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.

[26] Zilong Bai, Zihan Xu, Cong Sun, Chengxi Zang, H. Timothy Bunnell, Catherine Sinfield, Jacqueline Rutter, Aaron Thomas Martinez, L. Charles Bailey, Mark Weiner, Thomas R. Campion, Thomas W. Carton, Christopher B. Forrest, Rainu Kaushal, Fei Wang, and Yifan Peng. Extracting post-acute sequelae of SARS-CoV-2 infection symptoms from clinical notes via hybrid natural language processing. *npj Health Systems*, 2:31, 2025.

[27] Trejeeve Martyn, Mazen Hanna, Steven E. Nissen, Harsh Patolia, Judah Rajendran, Weiqi Sun, Eirini Schlosser, Jason Cannavale, Konstantinos Bakogiannis, W. H. Wilson Tang, Amanda Vest, Olga Tasopoulou, Aikaterini Karathanasopoulou, Leslie Cho, Lars G. Svensson, Vaibhav Mavi, Aaron Gerds, Samir Kapadia, Lara Jehi, and Ashish Sarraju. Automating chart review using an artificial intelligence-enabled system for assessing transthyretin amyloid cardiomyopathy trial eligibility. *Journal of Cardiac Failure*, 2026.

[28] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[29] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[30] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[31] 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*, pages 785–794, 2016.

[32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[34] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[35] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[36] DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

[37] Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[38] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[39] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[40] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2(5):283–288, 2020.

[41] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4036–4054. PMLR, 2022.

[42] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[43] PhysioNet. PhysioNet credentialed health data use agreement 1.5.0. `https://physionet.org/content/mimiciii/view-dua/1.4/`. Accessed October 5, 2026.

[44] PhysioNet. Use of MIMIC data with large language models and online services. `https://physionet.org/news/post/llm-responsible-use/`, September 2025.

# A Broader Impact

The integration of agentic LLMs into healthcare presents immense opportunities, yet their inherent overconfidence poses profound risks in high-stakes scenarios. As a critically important but largely underexplored direction, clinical calibration ensures that predictive confidence strictly aligns with empirical frequencies, thereby preventing clinicians from being misled by automated hallucinations. This paper represents a pioneering step toward mitigating these risks. By providing mathematically grounded mechanisms and comprehensive benchmarks to evaluate and enhance agentic LLM calibration, C²Align directly addresses automation bias and alarm fatigue. Ultimately, this early research lays the necessary groundwork for transitioning complex agentic reasoning from theoretical text comprehension to statistically reliable clinical decision-making.

Calibrated probabilities can still be wrong for individual patients, may degrade under distribution shift or in under-represented subgroups, and may encourage over-reliance. Any deployment requires local validation, monitoring of calibration over time and across subgroups, and clinician oversight. LLM inference on patient data must follow the applicable data-use agreements and institutional privacy controls.

C²Align is designed for clinical decision support, not autonomous decision-making. For HF and ADRD, potential users are physicians and trained reviewers who assess diagnoses or phenotypes from long clinical records; presenting calibrated risks together with the specialist indicators and their source evidence lets users verify the rationale and reject extractions that are not supported by the chart. For readmission, predictions are most useful before discharge plans are finalized; for mortality or care-escalation risk, before inpatient review and escalation decisions.

# B Data, Ethics, and Privacy

**IRB.** The study, including the physician annotation of HF and ADRD records and the physician audit of specialist extractions, was reviewed and approved by the Weill Cornell Medicine Institutional Review Board (protocol no. 25-07029127-02).

**Data access.** MIMIC-III [38] and MIMIC-IV [39] were accessed through PhysioNet [42] by credentialed researchers under the PhysioNet Credentialed Health Data Use Agreement [43], following PhysioNet's guidance on the responsible use of credentialed data with online services [44]. The TJH data [40] were used under their access conditions.

**LLM inference environment.** Model inference was conducted only in institutionally approved, security-controlled environments, including security-tagged local compute and approved enterprise API endpoints such as secure Azure OpenAI services. No consumer-facing chat product was used. The endpoint configurations used for the reported experiments were governed by enterprise terms and institutional controls that prohibit provider training and human review and prevent persistent retention of prompts and outputs. Only results generated under configurations compatible with the applicable data-use agreement are reported.

**Release.** The code repository (`https://github.com/HelloZicky/C2Align_repo`) contains the cohort-construction logic, prompts and fixed JSON schemas, task-specific feature allowlists, the fail-closed feature-policy module and its tests, the callable XGBoost objective, random seeds, dependency specifications, and entry points. It does not contain patient records, identifiers, raw notes, credentials, or completed physician annotation forms, which may contain restricted record-level information. Researchers with approved data access can reconstruct the processing and training pipeline.

# C Implementation Details

**Hyperparameters.** All hyperparameters can be found in Table 5.

**Custom objective for XGBoost.** For patient $i$ with raw margin $z_i$, $p_i = \sigma(z_i)$, label $y_i$, and branch probabilities $q_i^{\text{spe}}, q_i^{\text{par}}$ (clipped to $[10^{-6}, 1-10^{-6}]$), the per-patient objective $\ell_i = \mathrm{CE}(y_i, p_i) + \alpha(p_i - y_i)^2 + \beta_{\text{spe}} D_{\mathrm{KL}}(q_i^{\text{spe}} \| p_i) + \beta_{\text{par}} D_{\mathrm{KL}}(q_i^{\text{par}} \| p_i)$, i.e., the binary form of $\mathcal{L}_{\text{total}}$ in Section 3.2.2,

Table 5: Experimental settings and hyperparameters.

| Parameter | Value / Search Space |
|---|---|
| ***LLM agents*** | |
| Parallel agents | $\{2, 3, 4\}$ |
| Temperature ($T$) | $\{0.0, 0.05, 0.1, 0.2\}$ |
| Token log probabilities (top-$k$) | $\{3, 5\}$ |
| #(Clinical specialist agents) | 3 |
| ***ML tool*** | |
| Candidate algorithms | XGBoost, Logistic Regression, GBT |
| `n_estimators` | $\{100, 200, 300\}$ |
| `learning_rate` | $\{0.01, 0.03, 0.05\}$ |
| `max_depth` | $\{3, 4, 5\}$ |
| ***Joint objective*** | |
| $\alpha, \beta_{\text{par}}, \beta_{\text{spe}}$ (search) | $\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.50\}$ |

has gradient and Hessian

$$g_i = (p_i - y_i) + 2\alpha(p_i - y_i)\, p_i(1 - p_i) + \beta_{\text{spe}}(p_i - q_i^{\text{spe}}) + \beta_{\text{par}}(p_i - q_i^{\text{par}}), \tag{11}$$

$$h_i = (1 + \beta_{\text{spe}} + \beta_{\text{par}})\, p_i(1 - p_i) + 2\alpha\, p_i(1 - p_i)\big[(p_i - y_i)(1 - 2p_i) + p_i(1 - p_i)\big], \tag{12}$$

and $h_i$ is floored at $10^{-6}$ because the Brier contribution to the curvature can be negative. The function `get_custom_xgb_obj_bin_dual_kl` returns $(g_i, h_i)$ at every boosting iteration and is passed directly as the `objective` of `XGBClassifier`, so all four terms influence split selection, leaf weights, and tree construction.

**Feature construction and missing values.** The specialist branch works like a fixed multidisciplinary assessment. For each task, the roles are defined in advance and each role is responsible for a predetermined set of fields, with each field's name, meaning, type, unit, and JSON schema specified in the prompt (Appendix E). Two patients may have different values or missing evidence, but never different feature names. After parsing, the downstream matrix is built from an explicit, ordered, task-level allowlist containing only (i) structured covariates available at the prediction time, (ii) the predefined specialist features, and (iii) the specialist and parallel probabilities. Training, validation, and test matrices use the same names and order. Missing values are handled as follows:

1. A missing, non-numeric, or non-finite value of an allowlisted feature is imputed with that feature's median estimated on the training set only.
2. The stored training median is reused unchanged for validation and test rows; no imputation statistic is computed from held-out data.
3. Zero is used only when an allowlisted feature is entirely non-finite in the training set, so that no median exists.
4. If a feature column required by the allowlist is absent from the input table, matrix construction raises an error instead of inventing the column.
5. Unexpected extra columns are ignored. Label, candidate-label, cohort-inclusion, annotation, adjudication, outcome-derived, post-index, identifier, and split or fold fields are rejected even if they are added to the allowlist by mistake; a released test injects such fields and fails if any reaches the matrix.
6. Malformed JSON, missing requested fields, and API failures are tracked separately. When no valid value is recovered, the field remains missing in the fixed schema; free text never becomes a new column.

# D Additional Results

**Calibration intercept and slope.** Table 6 complements Table 2. All intercepts are negative and all slopes are below 1, so every method, including C²Align, tends to overestimate risk and to produce overly extreme probabilities on this task; C²Align has the intercept closest to 0 for every backbone (tied on DeepSeek-V4).

**Subgroups.** Table 7 reports AUROC, BS, and ECE by age, sex, and race on Readmission (IV). C²Align has the lowest ECE in all 11 subgroups, the lowest BS in 10 (MAS Parallel is lower for Black patients), and the highest AUROC in 8.

Table 6: Calibration intercept (ideal 0) and slope (ideal 1) with 95% CIs on Readmission (IV) across foundation models (complements Table 2). Point estimates closest to the ideal values within each backbone are bolded, including ties.

| Backbone | Model | | Intercept | Slope |
|---|---|---|---|---|
| ChatGPT-4o | LLM | Base | −1.100 [−2.080, −0.675] | 0.116 [−0.156, 0.675] |
| | | Calibration | −0.684 [−1.236, −0.308] | 0.434 [0.109, 0.934] |
| | MAS | Sequential | −2.083 [−4.404, −1.299] | 0.640 [0.252, 1.807] |
| | | Parallel | −1.238 [−3.099, −1.065] | 0.285 [0.166, 1.424] |
| | | Calibration | −0.745 [−1.504, −0.366] | 0.595 [0.262, 1.239] |
| | | **Ours** | **−0.413** [−0.688, −0.070] | **0.660** [0.281, 1.103] |
| Gemini-2.5 | LLM | Base | −0.997 [−2.147, −0.698] | 0.036 [−0.112, 0.509] |
| | | Calibration | −0.626 [−1.055, −0.317] | 0.427 [0.120, 0.881] |
| | MAS | Sequential | −1.683 [−2.750, −1.278] | 0.409 [0.220, 0.890] |
| | | Parallel | −1.391 [−2.643, −1.125] | 0.308 [0.164, 0.938] |
| | | Calibration | −0.538 [−0.792, −0.226] | **0.522** [0.177, 0.972] |
| | | **Ours** | **−0.536** [−0.789, −0.224] | **0.522** [0.177, 0.972] |
| Deepseek-V4 | LLM | Base | −0.883 [−1.532, −0.493] | −0.017 [−0.195, 0.286] |
| | | Calibration | −0.492 [−0.909, −0.136] | 0.543 [0.205, 0.972] |
| | MAS | Sequential | −1.038 [−1.463, −0.973] | 0.101 [0.039, 0.413] |
| | | Parallel | −0.871 [−1.210, −0.627] | −0.040 [−0.248, 0.239] |
| | | Calibration | **−0.317** [−0.630, 0.024] | **0.794** [0.384, 1.279] |
| | | **Ours** | **−0.317** [−0.630, 0.024] | **0.794** [0.384, 1.279] |
| LLaMA-3.1-8B | LLM | Base | −0.993 [−1.920, −0.622] | 0.045 [−0.154, 0.490] |
| | | Calibration | −0.542 [−0.959, −0.202] | 0.524 [0.189, 0.986] |
| | MAS | Sequential | −1.601 [−2.872, −1.183] | 0.390 [0.170, 1.037] |
| | | Parallel | −1.167 [−2.317, −0.939] | 0.184 [0.027, 0.867] |
| | | Calibration | −0.489 [−0.834, −0.157] | 0.627 [0.268, 1.106] |
| | | **Ours** | **−0.422** [−0.702, −0.090] | **0.652** [0.281, 1.118] |

Table 7: Descriptive subgroup analysis on Readmission (IV) (ChatGPT-4o backbone for the LLM-based methods). Best values per subgroup are in bold. Subgroup sizes are small, so these results are descriptive; all methods have AUROC below 0.5 in the Asian subgroup.

| Group | Subgroup | XGBoost | | | LR | | | LLM | | | MAS Sequential | | | MAS Parallel | | | C²Align | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | BS | ECE | AUROC | BS | ECE | AUROC | BS | ECE | AUROC | BS | ECE | AUROC | BS | ECE | AUROC | BS | ECE |
| Age | 18–49 | .6415 | .2209 | .1214 | .6192 | .2409 | .1464 | .5182 | .4987 | .5424 | .5830 | .4576 | .4841 | .5580 | .4180 | .4520 | **.6660** | **.2050** | **.0720** |
| | 50–64 | .5976 | .2639 | .2181 | .5753 | .2839 | .2431 | .4947 | .5337 | .6319 | .5990 | .5296 | .5771 | .5740 | .4900 | .5450 | **.6030** | **.2030** | **.1580** |
| | 65–74 | .6535 | .2382 | .1766 | .6312 | .2582 | .2016 | .5442 | .5077 | .5859 | .6410 | .4896 | .5331 | .6160 | .4500 | .5010 | **.6600** | **.1910** | **.1100** |
| | 75–84 | .6975 | .2062 | .0992 | .6752 | .2262 | .1242 | .5887 | .4727 | .5139 | **.7110** | .4146 | .4241 | .6860 | .3750 | .3920 | .6790 | **.1960** | **.0750** |
| | 85+ | .6071 | .2345 | .1519 | .6048 | .2545 | .1769 | .5032 | .4957 | .5589 | .6080 | .4486 | .4541 | .5830 | .4090 | .4220 | **.6110** | **.2080** | **.1350** |
| Sex | F | .6001 | .2415 | .1276 | .5778 | .2615 | .1526 | .5032 | .5047 | .5374 | **.6160** | .4656 | .4991 | .5910 | .4260 | .4670 | .6030 | **.2090** | **.0470** |
| | M | .6708 | .2304 | .1239 | .6485 | .2504 | .1489 | .5552 | .5047 | .5634 | .6470 | .4826 | .5121 | .6220 | .4430 | .4800 | **.6760** | **.1920** | **.0860** |
| Race | White | .6140 | .2502 | .1436 | .5917 | .2702 | .1686 | .5092 | .5237 | .5814 | **.6260** | .5086 | .5531 | .6010 | .4690 | .5210 | .6050 | **.2040** | **.0810** |
| | Black | .6700 | .1574 | .1131 | .6477 | .1774 | .1281 | .5877 | .4087 | .5049 | .6870 | .1876 | .1491 | .6620 | **.1480** | .1170 | **.7010** | .1950 | **.1020** |
| | Asian | .3550 | .2385 | .2722 | .3550 | .2385 | .2972 | .2392 | .4802 | .6484 | .3580 | .3956 | .3821 | .3330 | .3560 | .3500 | **.3773** | **.2300** | **.2660** |
| | Unknown/other | .7906 | .2160 | .1922 | .7683 | .2360 | .2172 | .7152 | .4872 | .6004 | .7980 | .5136 | .5791 | .7730 | .4740 | .5470 | **.8450** | **.1260** | **.0930** |

**Efficiency.** Table 8 reports mean per-patient token use, estimated API charge under the prices used for the experiments, and measured end-to-end latency. The physician row is an order-of-magnitude reference for a one-minute review at an assumed wage of $115 per hour; it is not a claim that one automated prediction replaces a physician encounter.

# E Prompts

We utilized specific prompt templates to guide the LLM baselines and the MAS. Below are the full, original instructions provided to each module.

For the diagnosis and phenotyping and prognostic prediction, ICD codes are treated as legitimate structured EHR features rather than label leakage. This setting follows standard computational phenotyping practice, where phenotype algorithms integrate billing codes, medications, procedures, laboratory results, and clinical narratives to identify whether a patient satisfies a chart-review phenotype definition. All methods that use structured EHR inputs, including LR, XGBoost, LLM baselines, MAS baselines, and C²Align, were provided access to the same structured code features when applicable. Thus, performance differences do not arise from C²Align having exclusive access to ICD information. For HF and ADRD, these codes include the target-disease codes used to select candidate

Table 8: Mean per-patient token use, estimated API cost, and latency.

| Backbone | Method | Tokens | Cost (USD) | Latency (s) |
|---|---|---|---|---|
| ChatGPT-4o | LLM | 5054.29 | 0.0148 | 1.13 |
| | MAS Sequential | 4474.15 | 0.0119 | 2.07 |
| | MAS Parallel | 4516.93 | 0.0121 | 1.17 |
| | C²Align | 9571.21 | 0.0269 | 2.24 |
| Gemini-2.5 | LLM | 5332.12 | 0.0025 | 6.03 |
| | MAS Sequential | 4513.53 | 0.0016 | 13.54 |
| | MAS Parallel | 4580.45 | 0.0017 | 6.06 |
| | C²Align | 9912.57 | 0.0042 | 15.59 |
| DeepSeek-V4 | LLM | 5103.41 | 0.0016 | 4.62 |
| | MAS Sequential | 4486.81 | 0.0013 | 9.56 |
| | MAS Parallel | 4542.40 | 0.0013 | 5.82 |
| | C²Align | 9645.81 | 0.0030 | 9.81 |
| US physician | – | – | 1.9167 | 60.00 |

admissions; the labels are the physicians' chart-review decisions, so the tasks assess phenotype confirmation rather than code detection.

### E.1 Diagnosis and Phenotyping

For Diagnosis and Phenotyping tasks, the models typically require both structured variables (such as demographics, laboratory results, and prescription data) and unstructured clinical narratives (such as discharge summaries and radiology reports).

#### E.1.1 LLM/MAS-Sequential/MAS-Parallel

The following is the full instruction used for the holistic LLM baseline.

**System Prompt**

```
You are a Senior Expert Cardiologist. Analyze the [Clinical Data] provided for
    Heart Failure (HF) diagnosis.

### 1. DIAGNOSTIC CHECKLIST & KEYWORDS (Strictly Scan for These):
Use the following lists as your primary extraction targets.

A. SYMPTOMS & SIGNS:
* Specific Symptoms: [Dyspnea, Orthopnea, PND, Edema, Fatigue, Rapid weight gain,
     JVD, Ascites, Cough, Wheezing].
* Supportive Signs: [PND, Orthopnea, Abnormal JVD, Pulmonary rales, Cardiomegaly,
     Edema, Night cough, Dyspnea on exertion, Hepatomegaly, Pleural effusion, HR
    >120, Weight loss >4.5kg/5days, Pulmonary edema, S3 gallop].
* Clinical Status: Check if patient is described as "Warm and wet" or "Fluid
    overloaded".

B. TESTS & IMAGING:
* Procedures: [Echocardiogram, ECG, Chest X-Ray, Stress Test, Cardiac MRI,
    Cardiac Cath, Holter, ABPM].
* Key Findings: specifically check for EF (LVEF) value (e.g., "EF 35%", "LVEF <
    40%").

C. MEDICATIONS:
* Drug Classes: [Diuretics (Loop/Thiazide), ACEi, ARBs, Beta-blockers,
    Aldosterone Antagonists (MRA), Digoxin, Vasodilators, Hydralazine, Inotropes,
     ARNI (Sacubitril/Valsartan), SGLT-2, GLP-1].
* Critical Checks: For each drug found, verify context (dosage, frequency, reason
    ). *Note: BB/ACEi alone for HTN does not equal HF.*

D. PROCEDURES & DEVICES:
* Devices: [ICD, CRT, LVAD].
```

```
* Surgeries: [CABG, PCI, Heart Transplantation, Valve surgery].

E. RISK FACTORS & EXCLUSIONS:
* History: [HTN, Diabetes, Hyperlipidemia, Obesity, Smoking, CKD, CAD].
* Exclusions/Differentials: [Cardiogenic shock, PE (Pulmonary Embolism), Stroke/
    CVA/TIA, AMI/ACS, Liver disease, Renal dysfunction (eGFR<20), DKA].

F. SOCIAL DETERMINANTS (SDOH):
* Factors: [Housing instability, Homelessness, Economic/Income barriers,
    Unemployment, Insurance issues, Substance use].

---

### 2. INSTRUCTIONS:

Step 1: Feature Assessment (Independent)
For `has_hf` field, determine if that specific data source provides evidence for
    HF.
* Judgment: Assign 1 (Supports HF), 0 (Does not support/Normal/Irrelevant).
* Reasoning (`reason_hf`):
    * MUST be < 40 words.
    * If 1: Quote the specific keyword or value found (e.g., "Found ICD code I50
    .9", "Echo shows EF 35%").
    * If 0: State "No HF evidence" or explain the negative finding (e.g., "Lungs
    clear", "Normal EF").

Step 2: LVEF Structured Extraction (CRITICAL)
Do not simplify; preserve the full topology of the data. Extract LVEF into 1
    fields:
ef_value_extracted: The exact string found (e.g., "30-35%", ">55%", "Normal", "
    severely reduced").

Step 3: Holistic Synthesis (The "Purity" Filter)
Determine `has_hf` based on the convergence of evidence.
* 1 (Confident POSITIVE): Definitive active or treated Heart Failure. Requires
    specific signs/symptoms AND objective evidence (Echo/Labs) OR specific HF
    devices (LVAD/CRT).
* 0 (Confident NEGATIVE): Definitive absence. Symptoms explained by other causes
    (e.g., COPD, PE) or objective data rules out HF (Normal Echo + Normal BNP).

Step 4: Confidence Scoring (CRITICAL SEMANTICS)
For the `hf_confidence` field, you MUST output the PROBABILITY THAT THE PATIENT
    HAS HF.
* 0 - 49: You believe the patient DOES NOT have HF. (e.g., 0 means you are 100%
    certain they are perfectly healthy / have no HF).
* 50 - 100: You believe the patient HAS HF. (e.g., 100 means you are 100% certain
     they have HF).
WARNING: Do NOT output a high number just because you are "confident they do not
    have it". High numbers ALWAYS mean "High probability of having the disease".

Step 5: Phenotyping (Subtype Logic)
For `hf_subtype` field, determine the specific subtype based on the extracted
    LVEF value and additional clinical context if available.
* IF `has_hf` is 0: The subtype MUST be "na".
* IF `has_hf` is 1: Assign subtype based on LVEF:
    * HFrEF: LVEF <= 40%.
    * HFmrEF: LVEF 41-49%.
    * HFpEF: LVEF >= 50% (requires accompanying diastolic dysfunction or elevated
     pressures).
    * Others: HF confirmed, but LVEF not reported or right-sided HF only.
* Reasoning (`reason_subtype`):
    * MUST be < 40 words.
```

```
---

### 3. JSON RESPONSE FORMAT (Strict Adherence Required):
Return ONLY a JSON object:
{
  "has_hf": <bool>,
  "hf_confidence": <0-100 integer representing the probability of HAVING HF (0=
    definitely healthy, 100=definitely diseased)>,
  "hf_subtype": "HFrEF" | "HFmrEF" | "HFpEF" | "Others" | "na",
  "subtype_probabilities": {
      "na": <0-100 integer>,
      "HFrEF": <0-100 integer>,
      "HFmrEF": <0-100 integer>,
      "HFpEF": <0-100 integer>,
      "Others": <0-100 integer>
  },
  "ef_value_extracted": "string or 'na'",
  "reason_hf": "<reasoning for HF diagnosis>",
  "reason_subtype": "<reasoning for subtype classification>"
}
Note: Ensure the 5 probability values in 'subtype_probabilities' sum to exactly
    100.
```

#### E.1.2 MAS-Specialist

In the MAS framework, clinical feature extraction is delegated to independent specialist agents using the following prompts.

**Symptom Agent**

```
You are a 'Symptom Specialist' for Heart Failure.
TASK: Extract specific symptoms and signs. Return JSON only.

OUTPUT JSON FORMAT:
{
    "symptoms_list": ["list of detected general symptoms"],
    "supportive_signs_15": ["list of detected 15 signs"],
    "is_warm_wet": <bool>
}
```

**Pharmacologist Agent**

```
You are a 'Pharmacologist Agent'.
TASK: Extract medications from these specific classes. Return JSON only.
TARGET CLASSES:
1. Diuretics (e.g., Furosemide, Bumex)
2. ACE Inhibitors (e.g., Lisinopril)
3. ARBs (e.g., Losartan)
4. Beta-Blockers (e.g., Carvedilol)
5. Aldosterone Antagonists/MRA (e.g., Spironolactone)
6. Digoxin
7. Vasodilators
8. Inotropes
9. ARNI
10. SGLT-2 Inhibitors
11. GLP-1 Agonists

OUTPUT JSON FORMAT:
{
    "medications": [
        {
```

```
            "class": "Diuretic",
            "name": "Bumex",
            "dose_freq": "1mg IV",
            "route": "IV",
            "reason": "fluid overload",
            "ade": "gut edema"
        }
    ]
}
```

### Diagnostic Test & Risk Factor Agent

```
You are a 'Diagnostic Test & Risk Factor Specialist'.
TASK: Extract Tests, Procedures, Risks, and SDOH. Return JSON only.
1. TESTS & IMAGING: Procedures: [Echo, ECG, Chest X-Ray, Stress Test, MRI, Cath,
    Holter].
   - LVEF: Extract the numeric Ejection Fraction % (e.g., "55%", "35-40%").
   - Labs: Note abnormal Creatinine or BNP levels.
2. PROCEDURES & DEVICES: [ICD, CRT, LVAD, CABG, PCI, Transplant, Valve Surgery].
3. RISK FACTORS: [HTN, Diabetes, Hyperlipidemia, Obesity, Smoking, CKD, CAD].
4. EXCLUSIONS: [Cardiogenic shock, PE, Stroke/CVA/TIA, AMI/ACS, Liver disease,
    Renal dysfunction, DKA].
5. SDOH: [Housing instability, Homelessness, Economic barriers, Unemployment,
    Insurance, Substance use].
6. FUNCTIONAL CLASS: NYHA Class (I-IV) or ACC/AHA Stage (A-D).

OUTPUT JSON FORMAT:
{
    "procedures_found": [],
    "lvef_extracted": "string or 'na'",
    "devices_surgeries": [],
    "risks_history": [],
    "exclusion_flags": [],
    "sdoh_factors": [],
    "nyha_class": "string"
}
```

### Diagnostic Agent

```
You are a Heart Failure Diagnostic Specialist.
Your Goal: Determine if the patient has Heart Failure (HF) based on the
    Specialist Reports provided.
Input: Summarized findings from Symptom, Pharmacologist, and Risk Specialists.

WARNING - CONFIDENCE SCORING:
For 'hf_confidence', you MUST output the PROBABILITY THAT THE PATIENT HAS HF.
* 0 - 49: You believe the patient DOES NOT have HF. (e.g., 0 means you are 100%
    certain they are healthy).
* 50 - 100: You believe the patient HAS HF. (e.g., 100 means you are 100% certain
     they have HF).
Do NOT output a high number just because you are "confident they do not have it".
     High numbers ALWAYS mean "High probability of having the disease".

Output JSON format:
{
    "has_hf": <bool>,
    "hf_confidence": <0-100 integer representing the probability of HAVING HF (0=
    definitely healthy, 100=definitely diseased)>,
    "reason_hf": "<Brief medical reasoning citing specific symptoms or tests>"
}
Do NOT default to Yes unless evidence supports it.
```

### Phenotyping Agent

```
You are a Heart Failure Phenotyping Specialist.
Your Goal: Classify the HF subtype based on the Diagnosis and Extracted EF.
Input: Diagnostic Agent result and Risk Specialist (EF) result.

Rules:
- HFrEF: EF <= 40%
- HFmrEF: EF 41-49%
- HFpEF: EF >= 50%
- Others: HF is present but EF is missing/ambiguous.
- na: Patient does NOT have HF.

Output JSON format:
{
    "hf_subtype": "HFrEF" | "HFmrEF" | "HFpEF" | "Others" | "na",
    "subtype_probabilities": {
        "na": <0-100 integer>,
        "HFrEF": <0-100 integer>,
        "HFmrEF": <0-100 integer>,
        "HFpEF": <0-100 integer>,
        "Others": <0-100 integer>
    },
    "reason_subtype": "<Brief reasoning based on EF value>"
}
Note: Ensure the 5 probability values in 'subtype_probabilities' sum to exactly
    100.
```

### Senior Chart Review Agent

```
You are a Senior Heart Failure Diagnostic Specialist specializing in Chart Review
    .
Your Goal: Review the findings of the previous agents (Specialists + Diagnosis)
    against the original patient data.
Correct any hallucinations or logic errors.

Input will include:
1. Original Clinical Note
2. Specialist Findings (Symptoms, Meds, Risks)
3. Preliminary Diagnosis & Subtype

Output JSON format (Final Validated Result):
{
    "has_hf": <bool>,
    "hf_confidence": <0-100 integer representing the probability of HAVING HF (0=
    definitely healthy, 100=definitely diseased)>,
    "ef_value_extracted": "string or 'na'",
    "hf_subtype": "<subtype string>",
    "subtype_probabilities": {
        "na": <0-100 integer>,
        "HFrEF": <0-100 integer>,
        "HFmrEF": <0-100 integer>,
        "HFpEF": <0-100 integer>,
        "Others": <0-100 integer>
    },
    "audit_reason": "<Why you kept or changed the decision>"
}

WARNING - STRICT JSON SEMANTICS:
1. Your output JSON MUST represent the PATIENT'S actual condition, NOT your
    agreement with the previous AI.
   - "has_hf": 1 (or true) means the PATIENT HAS HF.
```

```
   - "has_hf": 0 (or false) means the PATIENT DOES NOT HAVE HF.
   DO NOT output "true" just to mean "I agree with the previous agent".
2. "hf_confidence" MUST represent the probability of HAVING the disease. If you
     confidently agree the patient is healthy, this value must be close to 0.

Note: Ensure the 5 probability values in 'subtype_probabilities' sum to exactly
     100.
```

### E.2 Prognostic Prediction

For Prognostic Prediction tasks (such as 30-Day Unplanned Readmission), the models utilize structured features (including demographics, length of stay, and discharge location) and unstructured clinical notes. For Prognostic Prediction tasks, providing ICD codes and prior clinical history is typically required.

#### E.2.1 LLM/MAS-Sequential/MAS-Parallel

The following is the full instruction used for the holistic LLM baseline in prognostic prediction.

**LLM Baseline System Prompt**

```
You are a Senior Clinical AI Assistant. Analyze the provided clinical data to [
    TASK_OBJECTIVE].

### INSTRUCTIONS & CLINICAL GUIDELINES:
[CLINICAL_FOCUS]

### JSON RESPONSE FORMAT (Strict Adherence Required):
Return ONLY a valid JSON object with EXACTLY these three keys:
{
  "prediction": <int>, // 1 for YES (Event will occur), 0 for NO (Event will not
    occur)
  "probability": <float>, // A float between 0.00 and 1.00 representing the
    clinical risk/probability of the event occurring.
  "reasoning": "<string>" // MUST be < 40 words. Briefly quote the strongest
    evidence supporting your prediction.
}
```

#### E.2.2 MAS-Specialist

In the MAS framework, clinical feature extraction for prognosis is delegated to three specialized agents, each focusing on different aspects of the patient's trajectory.

**Clinical Status Agent**

```
You are a Clinical Status Specialist. Your goal is to review the notes to help [
    TASK_OBJECTIVE]
TASK: Extract acute symptoms, vital sign instability, and overall physical
    trajectory.
GUIDELINE: [CLINICAL_FOCUS]
OUTPUT JSON ONLY:
{
    "key_symptoms": ["list of severe/unresolved symptoms"],
    "has_vital_instability": <bool>,
    "clinical_trajectory": "improving" | "stable" | "deteriorating"
}
```

### Treatment & Intervention Agent

```
You are a Treatment & Intervention Specialist. Your goal is to review the notes
    to help [TASK_OBJECTIVE]
TASK: Extract major life support, critical surgeries, and high-risk medications.
OUTPUT JSON ONLY:
{
    "key_interventions": ["list of major interventions (e.g., intubation,
    vasopressors, dialysis)"],
    "is_on_life_support": <bool>,
    "medication_risk_level": "high" | "medium" | "low"
}
```

### Medical History Agent

```
You are a Medical History Specialist. Your goal is to review the notes to help [
    TASK_OBJECTIVE]
TASK: Extract severe chronic conditions, historical risk factors, and critical
    lab/imaging findings.
OUTPUT JSON ONLY:
{
    "chronic_conditions": ["list of severe chronic illnesses"],
    "critical_lab_imaging": ["list of severe abnormal findings"],
    "high_baseline_risk_flag": <bool>
}
```

### Senior Diagnostician Agent

```
You are the Senior Diagnostician.
Goal: [TASK_OBJECTIVE]
Input: Summaries from Clinical, Intervention, and Risk Specialists.

OUTPUT JSON FORMAT (Strict Adherence):
{
    "prediction": <1 for YES/High Risk, 0 for NO/Low Risk>,
    "probability": <float between 0.00 and 1.00 representing the clinical risk/
    probability of the event occurring>,
    "reasoning": "<Brief reasoning quoting specific evidence, max 40 words>"
}
```

### Senior Chart Review Agent

```
You are a Senior Chart Review Auditor.
Goal: [TASK_OBJECTIVE]
Input: The original note, Specialist Summaries, and a Preliminary Prediction.
Task: Audit the preliminary prediction. Look for missed nuances, false
    assumptions, or hallucinations.

OUTPUT JSON FORMAT (Strict Adherence):
{
    "prediction": <1 or 0>,
    "probability": <float between 0.00 and 1.00 representing the finalized risk/
    probability>,
    "audit_reason": "<Why you kept or changed the prediction, max 40 words>"
}
```

# F Theoretical Framework and Architecture Formalization

To rigorously analyze the generalization bounds and calibration capabilities of various clinical prediction architectures, we formalize the diagnostic process within a unified Statistical Learning framework.

## F.1 Notation, Operators, and Probability Space

Let the raw patient clinical record be a random variable $\mathcal{X} = \{\mathcal{X}_{\text{struct}}, \mathcal{X}_{\text{unstruct}}\}$, comprising structured covariates and unstructured narrative notes. Let the ground-truth diagnostic label be $\mathbf{Y} \in \{0,1\}^K$, where $K$ is the number of distinct clinical outcomes. We assume the dataset $S_N = \{(\mathcal{X}_i, \mathbf{Y}_i)\}_{i=1}^N$ is drawn i.i.d. from an unknown joint distribution $\mathbb{P}$ over $\mathcal{X} \times \{0,1\}^K$. All architectures aim to learn a predictive mapping $f : \mathcal{X} \to \Delta^{K-1}$.

In NLP, an LLM mapping is traditionally conditioned on a prompt instruction $c \in \mathcal{C}$. However, within a fixed architecture, the prompt $c$ acts as a constant prior. Because $c$ does not alter the hypothesis class complexity or the asymptotic generalization bounds with respect to sample size $N$, we absorb it into the operator definitions to maintain notational clarity.

We define the following mathematical operators and structural variables:

- $\Phi(\cdot)$: The base LLM forward-mapping transition kernel.
- $\Phi_{\text{chain}}(\cdot)$: A composite operator representing the Markovian sequential reasoning chain of multiple LLMs.
- $g_{\text{cal}}(\cdot)$: A post-hoc calibration function (e.g., Platt Scaling or Isotonic Regression).
- $\mathcal{M}(\cdot)$: A trainable downstream ML classifier parameterized by $\theta$.
- $\mathbf{u}_{\text{traj}} \in \mathbb{R}^{d_u}$: The vector representing intrinsic fluctuations in auto-regressive token probabilities derived from the agentic trajectory.
- $\mathbf{x}_{\text{spe}} \in \mathbb{R}^d$: The explicitly extracted clinical semantic features from the heterogeneous specialist branch.
- $\oplus$: The vector concatenation operator.

Specific to the C²Align framework, we formally define the augmented manifold as $\mathbf{x}_{\text{aug}} = \mathbf{x}_{\text{struct}} \oplus \mathbf{x}_{\text{spe}} \oplus \mathbf{p}_{\text{par}} \oplus \mathbf{p}_{\text{spe}}$, where $\mathbf{p}_{\text{par}}, \mathbf{p}_{\text{spe}} \in \Delta^{K-1}$ represent the diagnostic probability distributions derived from the homogeneous parallel branch and the heterogeneous specialist branch, respectively.

## F.2 Formalization of Clinical Prediction Architectures

Based on these operators, we mathematically construct the functional forms of the nine cognitive architectures evaluated in this study. Let $\mathcal{F}_m$ denote the hypothesis class for Method $m$.

**M1. Single LLM:** Direct zero-shot inference.

$$f_1(\mathcal{X}) = \Phi(\mathcal{X})$$

**M2. Agentic LLM (Sequential):** A multi-turn reasoning trajectory.

$$f_2(\mathcal{X}) = \Phi_{\text{chain}}(\mathcal{X})$$

**M3. Agentic LLM (Parallel):** Aggregating $M$ independent stochastic generations, where $\epsilon_m$ denotes the i.i.d. sampling perturbation.

$$f_3(\mathcal{X}) = \frac{1}{M} \sum_{m=1}^{M} \Phi(\mathcal{X}; \epsilon_m)$$

**M4. Agentic LLM (Hybrid):** Parallel generations are concatenated and fed into a sequential summarizer.

$$f_4(\mathcal{X}) = \Phi_{\text{chain}} \left( \bigoplus_{m=1}^{M} \Phi^{(m)}(\mathcal{X}) \right)$$

**M5. Agentic LLM (Hybrid) + Trajectory Calibration:** Fits a linear boundary parameterized by weights $\mathbf{w}$ and bias $b$ over the extracted trajectory dynamics $\mathbf{u}_{\text{traj}}$.

$$f_5(\mathcal{X}) = \sigma(\mathbf{w}^\top \mathbf{u}_{\text{traj}} + b)$$

**M6. Agentic LLM (Specialist Divide & Conquer):** The input is partitioned into $C$ clinical sub-domains. Specialized sub-agents process each domain before a final fusion.

$$f_6(\mathcal{X}) = \Phi\left(\bigoplus_{c=1}^{C} \Phi(\mathcal{X}^{(c)})\right)$$

**M7. Dual-Architecture Agentic LLM + ML tool (Base):** Insights form the augmented manifold $\mathbf{x}_{\text{aug}}$ as input to the downstream ML model.

$$f_7(\mathcal{X}) = \mathcal{M}(\mathbf{x}_{\text{aug}})$$

**M8. Dual-Architecture Agentic LLM + ML tool (Posthoc calibration):** Applies post-hoc calibration over the fully trained M7 model.

$$f_8(\mathcal{X}) = g_{\text{cal}}\left(\mathcal{M}(\mathbf{x}_{\text{aug}})\right)$$

**M9. Dual-Architecture Agentic LLM + ML tool (Joint loss calibration) [C²Align]:** Shares the hypothesis space of M7, but the optimal parameters $\theta^*$ are explicitly constrained by the C²Align composite objective.

$$\hat{f}_9(\mathcal{X}) = \mathcal{M}(\mathbf{x}_{\text{aug}}; \theta^*), \quad \text{where } \theta^* = \arg\min_{\theta} \left[\mathcal{L}_{\text{Dis}} + \alpha \mathcal{L}_{\text{Cal}} + \mathcal{L}_{\text{Align}}\right]$$

By unifying these methodologies under a strict probability space, we lay the groundwork to theoretically bound their empirical Rademacher complexities $\hat{\mathfrak{R}}_N(\mathcal{F}_m)$ and elucidate the structural advantages of loss-calibrated agentic frameworks.

### F.3 Generalization Bounds and Theoretical Analysis

To compare the proposed architectures under explicit statistical assumptions, we analyze their expected risk on unseen data using generalization bound theory. We specifically focus on how structural changes affect *discrimination accuracy* (measured by AUROC, AUPRC, PPV) and *calibration* (measured by BS, ECE).

**Theorem 4** (Generalization Error Bound)**.** *Let the evaluation scoring rule $\ell : [0,1]^K \times \{0,1\}^K \to \mathbb{R}_+$ be $L_\ell$-Lipschitz continuous with respect to the $\ell_2$-norm and bounded by a constant $B > 0$. For the averaged multiclass BS $\ell_{\text{Brier}}(\mathbf{p}, \mathbf{y}) = \frac{1}{K}\|\mathbf{p} - \mathbf{y}\|_2^2$, the loss is bounded on $\Delta^{K-1} \times \{0,1\}^K$; throughout the following bounds, $B$ and $L_\ell$ denote the corresponding boundedness and Lipschitz constants. For any prediction architecture $m \in \{1, \ldots, 9\}$ with an associated hypothesis class $\mathcal{F}_m$, with probability at least $1 - \delta_{\text{conf}}$ over the random draw of the training sample $S_N$, every predictor $\hat{f}_m \in \mathcal{F}_m$ satisfies:*

$$\mathcal{R}(\hat{f}_m) \leq \underbrace{\hat{\mathcal{R}}(\hat{f}_m)}_{\textit{Empirical Risk}} + \underbrace{2L_\ell \hat{\mathfrak{R}}_N(\mathcal{F}_m)}_{\textit{Complexity Penalty}} + \underbrace{3B\sqrt{\frac{\log(2/\delta_{\text{conf}})}{2N}}}_{\textit{Confidence Bound}} \tag{13}$$

*Furthermore, if $\hat{f}_m$ is specifically chosen as the empirical risk minimizer (ERM), it satisfies the oracle inequality:*

$$\mathcal{R}(\hat{f}_m) \leq \underbrace{\inf_{f \in \mathcal{F}_m} \mathcal{R}(f)}_{\textit{Best-in-class Risk}} + \underbrace{2L_\ell \hat{\mathfrak{R}}_N(\mathcal{F}_m) + 4B\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}}}_{\textit{Estimation Penalty}} \tag{14}$$

*where $\mathcal{R}(f) = \mathbb{E}_{(\mathcal{X},\mathbf{Y})\sim\mathbb{P}}[\ell(f(\mathcal{X}), \mathbf{Y})]$ is the expected risk, $\hat{\mathcal{R}}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell(f(\mathcal{X}_i), \mathbf{Y}_i)$ is the empirical risk, and $\hat{\mathfrak{R}}_N(\mathcal{F}_m)$ is the empirical Rademacher complexity, defined for a sample $S_N = \{\mathcal{X}_i\}_{i=1}^N$ and independent Rademacher variables $\sigma_i \in \{-1, 1\}$ as:*

$$\hat{\mathfrak{R}}_N(\mathcal{F}_m) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}_m} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f(\mathcal{X}_i)\right] \tag{15}$$

The theoretical objective in designing C²Align is to minimize the approximation error by extracting a richer information manifold, while explicitly bounding the Rademacher complexity to control the estimation error.

However, comparing unconstrained Brier risks directly is theoretically flawed. A naive model could output the constant marginal prior $\mathbb{P}(\mathbf{Y})$ to arbitrarily drive its calibration penalty to zero while failing to discriminate. To guarantee a mathematically valid and fair comparison across architectures, we must establish their risk bounds under a controlled Information-Theoretic accuracy constraint. We formalize this universal lower bound for a specific binary margin in the following Lemma.

**Lemma 5** (Information-Theoretic Brier Risk Lower Bound)**.** *Let $Y \in \{0, 1\}$ be a true binary target (or a specific class margin) and $Z$ be an arbitrary latent information state. Let $p(Z) = \mathbb{P}(Y = 1 \mid Z)$ be the true conditional posterior. Let $q(Z) \in [\tau, 1 - \tau]$ be a model's predicted probability, symmetrically clipped by a predefined stabilization constant $\tau \in (0, 0.5)$. If the predictor is constrained to a fixed cross-entropy Log-loss level $\eta$, i.e., $\mathbb{E}[L_{log}(q)] = \eta$ (an equality constraint; the bound does not hold under $\mathbb{E}[L_{log}(q)] \leq \eta$), then its Brier risk is strictly bounded from below by the intrinsic properties of the information state $Z$:*

$$\mathcal{R}_{Brier}(q) \geq \mathbb{E}_Z[\mathit{Var}(Y \mid Z)] + \tau(1 - \tau)\Big(\eta - \mathbb{E}_Z[H(Y \mid Z)]\Big) \tag{16}$$

*where $H(Y \mid Z)$ is the inherent conditional entropy.*

For a hypothesis class $\mathcal{F}_m$, we write $\mathcal{F}_m(\eta) = \{f \in \mathcal{F}_m : \mathbb{E}[L_{log}(f)] = \eta\}$ for its predictors at this level.

*Proof.* By definition, the expected cross-entropy Log-loss is given by:

$$\begin{aligned}
\mathbb{E}[L_{log}(q)] &= \mathbb{E}_Z\Big[ - p(Z) \log q(Z) - (1 - p(Z)) \log(1 - q(Z))\Big] \\
&= \mathbb{E}_Z\Big[ - p(Z) \log p(Z) - (1 - p(Z)) \log(1 - p(Z))\Big] \\
&\quad + \mathbb{E}_Z\Big[p(Z) \log \frac{p(Z)}{q(Z)} + (1 - p(Z)) \log \frac{1 - p(Z)}{1 - q(Z)}\Big] \\
&= \mathbb{E}_Z\left[H(Y \mid Z)\right] + \mathbb{E}_Z\left[D_{KL}(p(Z) \parallel q(Z))\right] = \eta
\end{aligned} \tag{17}$$

Isolating the KL-divergence term yields the exact equality:

$$\mathbb{E}_Z\left[D_{KL}(p(Z) \parallel q(Z))\right] = \eta - \mathbb{E}_Z\left[H(Y \mid Z)\right] \tag{18}$$

Simultaneously, the BS $\mathcal{R}_{Brier}(q) = \mathbb{E}_{Z,Y}[(Y - q(Z))^2]$ algebraically decomposes into the irreducible variance and the expected squared calibration error:

$$\mathcal{R}_{Brier}(q) = \mathbb{E}_Z[\mathrm{Var}(Y \mid Z)] + \mathbb{E}_Z[(p(Z) - q(Z))^2] \tag{19}$$

To bridge the Log-loss constraint and the BS, we bound the squared error using $D_{KL}(p \parallel q)$. Since $\log x \leq x - 1$, the KL-divergence between Bernoulli distributions is bounded by the $\chi^2$-divergence, $D_{KL}(p \parallel q) \leq \frac{(p-q)^2}{q(1-q)}$, and since $q(Z) \in [\tau, 1 - \tau]$ implies $q(1 - q) \geq \tau(1 - \tau)$:

$$D_{KL}(p \parallel q) \leq \frac{(p - q)^2}{\min_z q(z)(1 - q(z))} \leq \frac{(p - q)^2}{\tau(1 - \tau)} \tag{20}$$

Taking the expectation over $Z$ establishes a strict lower bound on the expected squared calibration error: $\mathbb{E}_Z[(p(Z) - q(Z))^2] \geq \tau(1 - \tau)\mathbb{E}_Z[D_{KL}(p(Z) \parallel q(Z))]$. Substituting this bound into Equation 19, and replacing the KL-divergence term via Equation 18, directly yields the analytical lower bound, concluding the proof. □

#### F.3.1 Evaluating Global-Context agentic Structures (M1, M2, M3, M4)

We first evaluate the theoretical limits of internal agentic structures that process the unpartitioned global context $\mathcal{X}$: Single-pass inference (M1), Sequential (M2), Parallel (M3), and Hybrid (M4).

**Proposition 1** (M1, M2, M3, M4: Information Loss and Variance Reduction)**.** *Operating over the identical global context $\mathcal{X}$, parallel aggregation (M3) suggest a reduction in generative variance under the stated assumptions, lowering the generalization bound compared to single-pass inference (M1). Conversely, sequential architectures (M2) structurally accumulate discretization noise, monotonically degrading mutual information and inflating the constrained lower bound of the generalization error.*

*Proof.* **0. Bounding the Estimation Error (Rademacher Complexity):** Before analyzing the approximation error, we must bound the estimation error. Because architectures M1, M2, M3, and M4 are constructed entirely from deterministic combinations of the frozen pre-trained LLM operator $\Phi$ without any parameterized training or optimization on the dataset $S_N$, their hypothesis classes are effectively fixed algorithmic compositions. Consequently, their capacity to overfit random empirical noise is strictly zero:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_1) = \hat{\mathfrak{R}}_N(\mathcal{F}_2) = \hat{\mathfrak{R}}_N(\mathcal{F}_3) = \hat{\mathfrak{R}}_N(\mathcal{F}_4) = 0 \tag{21}$$

By substituting this into the generalization decomposition, the estimation error term completely vanishes. Therefore, comparing the theoretical generalization bounds of these agentic architectures reduces strictly to comparing their approximation errors (i.e., Bayes risk).

**1. The Error Propagation of Sequential Agents (M2):** Let $\sigma_0^2 \geq 0$ denote the intrinsic irreducible variance of the LLM's stochastic decoding step. M2 models reasoning as a sequential chain where step $t$ depends on both the full context $\mathcal{X}$ and the previous discrete generation $Z_{t-1}$. Let the latent information state be mapped as $Z_t = \Phi(\mathcal{X}, Z_{t-1}) = \mu(\mathcal{X}, Z_{t-1}) + \xi_t$, where $\xi_t$ is an independent zero-mean stochastic noise with $\text{Var}(\xi_t) = \sigma_0^2$.

By the variance addition rule, since the current sampling noise $\xi_T$ is conditionally independent of the deterministic historical mapping $\mu$, their cross-covariance is zero. Using the Delta Method (first-order Taylor expansion) around the expected historical state, the propagated variance is approximated by:

$$\text{Var}(Z_T \mid \mathcal{X}) \approx \left(\frac{\partial \mu}{\partial Z_{T-1}}\right)^2 \text{Var}(Z_{T-1} \mid \mathcal{X}) + \sigma_0^2 \tag{22}$$

By recursively expanding this equation from $t = 1$ to $T$, the baseline noise $\sigma_0^2$ accumulates at each generation step. Let $\sigma_{seq}^2 = \text{Var}(Z_T \mid \mathcal{X})$ denote this final accumulated generative variance. Assuming the reasoning chain does not exhibit perfect information collapse (i.e., the Jacobian $\partial\mu/\partial Z \neq 0$), it is guaranteed that $\sigma_{seq}^2 > \sigma_0^2$ for any sequence length $T > 1$.

Orthogonally decomposing the expected Brier risk for M2 into its deterministic reasoning path and the accumulated zero-mean stochastic noise yields:

$$\inf_{f \in \mathcal{F}_2} \mathcal{R}_{Brier}(f) = \mathbb{E}_{\mathcal{X},\mathbf{Y}}\left[\|\mathbf{Y} - \mu^{(T)}(\mathcal{X})\|_2^2\right] + \sigma_{seq}^2 \tag{23}$$

Assuming the optimal sequential mapping $\mu^{(T)}(\mathcal{X})$ asymptotically achieves an equivalent empirical bias to the single-pass mapping $\mu(\mathcal{X})$ in M1, the difference is dictated by the generative noise penalty. Because $\sigma_{seq}^2 > \sigma_0^2$, we establish:

$$\inf_{f \in \mathcal{F}_2} \mathcal{R}_{Brier}(f) \approx \inf_{f \in \mathcal{F}_1} \mathcal{R}_{Brier}(f) + (\sigma_{seq}^2 - \sigma_0^2) > \inf_{f \in \mathcal{F}_1} \mathcal{R}_{Brier}(f) \tag{24}$$

**2. Variance Reduction in Parallel Agents (M3):** M3 queries the LLM $M$ times. Let the output of a single pass be decomposed into the expected latent response and the unified sampling noise: $\Phi(\mathcal{X}; \xi) = \mu(\mathcal{X}) + \xi$, where $\text{Var}(\xi) = \sigma_0^2 \geq 0$. For M3, $f_3(\mathcal{X}) = \frac{1}{M}\sum_{m=1}^{M} \Phi(\mathcal{X}; \xi_m)$. Orthogonally decomposing the expected Brier risk across independent sampling:

$$\begin{aligned} \inf_{f \in \mathcal{F}_3} \mathcal{R}_{Brier}(f) &= \mathbb{E}_{\mathcal{X},\mathbf{Y}}\left[\|\mathbf{Y} - \mu(\mathcal{X})\|_2^2\right] + \mathbb{E}_{\xi}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} \xi_m\right\|_2^2\right] \\ &= \inf_{f \in \mathcal{F}_1} \mathcal{R}_{Brier}(f) - \frac{M-1}{M}\sigma_0^2 \end{aligned} \tag{25}$$

Because $\sigma_0^2 \geq 0$ and $M \geq 1$, we analytically establish $\inf \mathcal{R}_{Brier}(f_3) \leq \inf \mathcal{R}_{Brier}(f_1)$. If the $M$ decoding errors are exchangeable with pairwise correlation $\rho \in [0, 1]$ rather than independent, the second term equals $\sigma_0^2[\rho + (1-\rho)/M]$, so the reduction becomes $\frac{M-1}{M}(1-\rho)\sigma_0^2$, which vanishes as $\rho \to 1$.

**3. The Error Propagation in Hybrid Integration (M4):** M4 aggregates $M$ parallel outputs into a consensus state and subsequently feeds it to a final sequential agent. Although M4 mitigates the initial sampling variance through $M$, the inescapable accumulation of the unmitigated final decoding noise mathematically bounds its performance strictly between M3 and M2.

**4. Brier Risk under Strict Accuracy Constraints:** Using the Information-Theoretic cross-entropy Log-loss constraint $L_{log}(q_m) = \eta$ for $m \in \{1, 2, 3\}$, and invoking Lemma 5, the constrained Brier risk lower bound for any architecture operating on latent state $Z_m$ is given by:

$$\inf_{f \in \mathcal{F}_m(\eta)} \mathcal{R}_{Brier}(f) \geq V_m + c(\eta - H_m) \equiv \mathcal{B}_{min}(Z_m) \tag{26}$$

where $c = \tau(1-\tau)$, $V_m = \mathbb{E}[\text{Var}(\mathbf{Y} \mid Z_m)]$ is the irreducible variance, and $H_m = \mathbb{E}[H(\mathbf{Y} \mid Z_m)]$ is the conditional entropy.

We link the conditional entropy to mutual information via $H_m = H(\mathbf{Y}) - I(\mathbf{Y}; Z_m)$:

- *Sequential Generation (M2):* Viewing generation as an information channel, mutual information is inversely proportional to generative noise. Because the sequential auto-regressive chain inflates the conditional variance ($\text{Var}(Z_2 \mid \mathcal{X}) \geq \text{Var}(Z_1 \mid \mathcal{X})$), the effective channel noise increases, degrading the mutual information: $I(\mathbf{Y}; Z_2) \leq I(\mathbf{Y}; Z_1) \implies H_2 \geq H_1$.
- *Parallel Consensus (M3):* Conversely, M3 suppresses generative variance via the Law of Large Numbers ($\text{Var}(Z_3 \mid \mathcal{X}) \leq \text{Var}(Z_1 \mid \mathcal{X})$). Reducing additive noise yields a higher signal-to-noise ratio, structurally increasing the mutual information: $I(\mathbf{Y}; Z_3) \geq I(\mathbf{Y}; Z_1) \implies H_3 \leq H_1$.

We define the variance differential $\Delta V = V_1 - V_3 \geq 0$, and the entropy differential $\Delta H = H_1 - H_3 \geq 0$. The difference in their constrained lower bounds is:

$$\Delta_{1,3} = \mathcal{B}_{min}(Z_1) - \mathcal{B}_{min}(Z_3) = \Delta V - c\Delta H \tag{27}$$

The sign of $\Delta_{1,3}$ depends on the relative magnitudes of the variance reduction $\Delta V$ and the scaled entropy difference $c\Delta H$. Under the controlled-discrimination comparison used in Theorem 1, we isolate the stochastic-variance contribution and assume that the variance reduction dominates the scaled entropy term:

$$\Delta V \geq c\Delta H. \tag{28}$$

Then

$$\mathcal{B}_{\min}(Z_3; \eta) \leq \mathcal{B}_{\min}(Z_1; \eta). \tag{29}$$

Applying identical algebraic bounding to M2 (where $V_2 \geq V_1$ and $H_2 \geq H_1$), and assuming analogously that the variance increase dominates, $V_2 - V_1 \geq c(H_2 - H_1)$, we obtain $\mathcal{B}_{min}(Z_1) \leq \mathcal{B}_{min}(Z_2)$. Substituting these relations back into Equation 26 establishes the final hierarchy:

$$\mathcal{B}_{\min}(Z_3; \eta) \leq \mathcal{B}_{\min}(Z_1; \eta) \leq \mathcal{B}_{\min}(Z_2; \eta) \tag{30}$$

Empirically, the Sequential MAS has higher test BS than the single-pass LLM baseline on the HF and ADRD tasks (0.1345 vs. 0.0670 and 0.1006 vs. 0.0523, respectively; Table 1). On Readmission (IV), Sequential also has higher BS than Parallel across all four backbones (Table 2). These findings are directionally consistent with poorer Brier performance under sequential reasoning, but they do not test the condition $V_2 - V_1 \geq c(H_2 - H_1)$, since the experiments neither match log-loss at a common $\eta$ nor estimate $V_m$ and $H_m$; we therefore retain the condition as an explicit assumption.

**5. Synthesis:** By confirming that empirical estimation errors ($\hat{\mathfrak{R}}_N$) vanish across unparameterized zero-shot LLM structures, their generalization capabilities are strictly bounded by their Approximation Errors. By imposing the Log-loss constraint $\eta$, we showed, under the stated assumptions, that sequential reasoning (M2) monotonically inflates both noise variance and conditional entropy. Conversely, parallel sampling (M3) intrinsically suppresses noise and recovers mutual information. Thus, under these assumptions, parallel structures have a tighter constrained lower bound than sequential chains; the ordering need not hold when the assumptions fail, for example when the parallel errors are perfectly correlated. □

### F.3.2 Trajectory-Awareness vs. Knowledge-Based Strategies

We now formally evaluate architectures incorporating downstream ML models. To ensure a rigorous and fair comparison—avoiding confounding variables related to model capacity—we evaluate M5 (Pure Trajectory Calibration) and M7 (C²Align Base) under the *exact same hypothesis class*.

Let the extracted trajectory dynamics be $\mathbf{u}_{\text{traj}} \in \mathbb{R}^{d_u}$, and the augmented clinical features be $\mathbf{x}_{\text{aug}} \in \mathbb{R}^{d_{\text{aug}}}$. Assuming all input features are boundedly normalized, let $C_w$ be the $\ell_2$ regularization bound on the linear model weights. Under Theorem 4, utilizing the BS ($L_\ell = 2$, $B = 1$), the generalization error for either architecture is bounded by:

$$\mathcal{R}_{Brier}(\hat{f}) \leq \underbrace{\inf_{f \in \mathcal{F}} \mathcal{R}_{Brier}(f)}_{\text{Approximation Error}} + \underbrace{4\hat{\mathfrak{R}}_N(\mathcal{F})}_{\text{Estimation Error}} + 4\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}} \tag{31}$$

To formalize the domain distinction between explicit knowledge and fluctuations in token probabilities, we establish a premise regarding their information structure:

**Assumption 1** (Causal Markov Property of Semantic Features)**.** *Assuming the explicit clinical features $\mathbf{x}_{aug}$ sufficiently capture the patient's physiological state, they form a causal Markov Blanket for the true diagnosis $\mathbf{Y}$ with respect to the trajectory fluctuations $\mathbf{u}_{traj}$. That is, once the clinical facts are observed, the model's auto-regressive token volatility provides no additional causal information regarding the actual disease state. Mathematically, they are conditionally independent: $\mathbf{Y} \perp\!\!\!\perp \mathbf{u}_{traj} \mid \mathbf{x}_{aug}$.*

**Proposition 2** (M5 vs. M7: The Information Manifold Dominance)**.** *Under Assumption 1, when restricted to identical linear capacities, both pure trajectory calibration (M5) and clinical knowledge integration (M7) have estimation-error bounds of the same order. However, M7 structurally dominates M5 by achieving a lower approximation error due to the causal completeness of the clinical semantic manifold.*

*Proof.* We restrict both architectures to the family of $L_2$-regularized linear models mapped through a 1-Lipschitz activation function $\sigma$ (e.g., Sigmoid). For M5: $\mathcal{F}_5 = \{\mathbf{u} \mapsto \sigma(\mathbf{w}^\top \mathbf{u}) \mid \|\mathbf{w}\|_2 \leq C_w\}$. For M7: $\mathcal{F}_7 = \{\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}_{\text{aug}}) \mid \|\mathbf{w}\|_2 \leq C_w\}$.

**1. Estimation Error (Rademacher Complexity):** We derive the empirical Rademacher complexity for a generic linear class $\mathcal{F} = \{\mathbf{v} \mapsto \sigma(\mathbf{w}^\top \mathbf{v}) \mid \|\mathbf{w}\|_2 \leq C_w\}$ over an arbitrary input feature matrix $\mathbf{V} \in \mathbb{R}^{N \times d}$. By Talagrand's Contraction Lemma, because the activation $\sigma$ is 1-Lipschitz, the complexity is bounded by the underlying linear function:

$$\hat{\mathfrak{R}}_N(\mathcal{F}) \leq \frac{1}{N}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\mathbf{w}\|_2 \leq C_w} \mathbf{w}^\top \sum_{i=1}^{N} \epsilon_i \mathbf{v}_i\right] \tag{32}$$

where $\epsilon_i \in \{-1, 1\}$ are independent Rademacher variables. By the Cauchy-Schwarz inequality, the supremum is $C_w \| \sum \epsilon_i \mathbf{v}_i \|_2$. Applying Jensen's Inequality ($\mathbb{E}[\sqrt{x}] \leq \sqrt{\mathbb{E}[x]}$) and expanding the squared $\ell_2$-norm:

$$\hat{\mathfrak{R}}_N(\mathcal{F}) \leq \frac{C_w}{N}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i \epsilon_j] \mathbf{v}_i^\top \mathbf{v}_j} \tag{33}$$

Because the Rademacher variables are independent and zero-mean, $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ for $i \neq j$, and $\mathbb{E}[\epsilon_i^2] = 1$. The double summation collapses to the diagonal:

$$\hat{\mathfrak{R}}_N(\mathcal{F}) \leq \frac{C_w}{N}\sqrt{\sum_{i=1}^{N} \|\mathbf{v}_i\|_2^2} \leq \frac{C_w R_v}{\sqrt{N}} \tag{34}$$

where $R_v = \max_i \|\mathbf{v}_i\|_2$. Applying this to our specific architectures, let $R_u \propto \sqrt{d_u}$ and $R_x \propto \sqrt{d_{\text{aug}}}$:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_5) \leq \mathcal{O}\left(\sqrt{\frac{d_u}{N}}\right), \quad \hat{\mathfrak{R}}_N(\mathcal{F}_7) \leq \mathcal{O}\left(\sqrt{\frac{d_{\text{aug}}}{N}}\right) \tag{35}$$

Because both $d_u$ and $d_{\text{aug}}$ are finite, low-dimensional structured vectors, the estimation errors for M5 and M7 are mathematically symmetric, decaying at the optimal $\mathcal{O}(1/\sqrt{N})$ rate.

**2. The Approximation Error Floor (Bayes Risk):** Because both architectures exhibit symmetric estimation errors, their generalization bounds diverge entirely at the approximation error floor (i.e., the Bayes Risk, governed by the expected conditional variance).

We analyze the Bayes Risk of M5 ($\mathbb{E}[\text{Var}(\mathbf{Y} \mid \mathbf{u}_{\text{traj}})]$) by projecting it through the clinical semantic space $\mathbf{x}_{\text{aug}}$. Applying the Law of Total Variance by conditioning on $\mathbf{x}_{\text{aug}}$:

$$\mathbb{E}_{\mathbf{u}}\Big[\text{Var}(\mathbf{Y} \mid \mathbf{u}_{\text{traj}})\Big] = \mathbb{E}_{\mathbf{u}}\Big[\mathbb{E}_{\mathbf{x}|\mathbf{u}}[\text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}}, \mathbf{u}_{\text{traj}})]\Big] + \mathbb{E}_{\mathbf{u}}\Big[\text{Var}_{\mathbf{x}|\mathbf{u}}(\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{\text{aug}}, \mathbf{u}_{\text{traj}}])\Big] \tag{36}$$

By invoking the Causal Markov Property established in Assumption 1 ($\mathbf{Y} \perp\!\!\!\perp \mathbf{u}_{\text{traj}} \mid \mathbf{x}_{\text{aug}}$), the joint conditioning simplifies to the marginal semantic conditioning: $\text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}}, \mathbf{u}_{\text{traj}}) = \text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}})$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{\text{aug}}, \mathbf{u}_{\text{traj}}] = \mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{\text{aug}}]$.

Substituting these causal equivalences back into Equation 36, the first term collapses exactly into the Bayes Risk of M7. We obtain an analytical decomposition:

$$\underbrace{\mathbb{E}\Big[\text{Var}(\mathbf{Y} \mid \mathbf{u}_{\text{traj}})\Big]}_{\text{M5 Bayes Risk}} = \underbrace{\mathbb{E}\Big[\text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}})\Big]}_{\text{M7 Bayes Risk}} + \underbrace{\mathbb{E}_{\mathbf{u}}\Big[\text{Var}_{\mathbf{x}|\mathbf{u}}\big(\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{\text{aug}}]\big)\Big]}_{\Delta_{\text{loss}} \geq 0} \tag{37}$$

Because clinical semantics govern the diagnosis, $\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_{\text{aug}}]$ is a discriminative function. When conditioned only on token fluctuations $\mathbf{u}_{\text{traj}}$, the variance of this expectation ($\Delta_{\text{loss}}$) is non-negative. Thus, we mathematically establish:

$$\mathbb{E}\Big[\text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}})\Big] \leq \mathbb{E}\Big[\text{Var}(\mathbf{Y} \mid \mathbf{u}_{\text{traj}})\Big] \tag{38}$$

The approximation errors of the norm-bounded linear classes are at least these Bayes risks, $\inf_{f \in \mathcal{F}_m} \mathcal{R}_{Brier}(f) \geq \mathbb{E}[\text{Var}(\mathbf{Y} \mid \cdot)]$, with equality only if $\mathcal{F}_m$ contains the corresponding Bayes predictor; the comparison above is therefore between the information contents of the two inputs.

**3. Brier Risk under Strict Accuracy Constraints:** To guarantee a valid comparison, we enforce an exactly equal Information-Theoretic accuracy by fixing the cross-entropy Log-loss to a level $\eta$. By invoking Lemma 5, the constrained Approximation Error bounds for M5 and M7 are:

$$\inf_{f \in \mathcal{F}_5(\eta)} \mathcal{R}_{Brier}(f) \geq V_{\mathbf{u}} + c\Big(\eta - H_{\mathbf{u}}\Big) \tag{39}$$

$$\inf_{f \in \mathcal{F}_7(\eta)} \mathcal{R}_{Brier}(f) \geq V_{\mathbf{x}} + c\Big(\eta - H_{\mathbf{x}}\Big) \tag{40}$$

where $c = \tau(1-\tau)$, $V_{(\cdot)}$ is the expected conditional variance, and $H_{(\cdot)}$ is the conditional entropy.

Because observing $\mathbf{x}_{\text{aug}}$ fundamentally resolves diagnostic uncertainty better than trajectory fluctuations $\mathbf{u}_{\text{traj}}$, the intrinsic variance and entropy satisfy:$V_{\mathbf{x}} \leq V_{\mathbf{u}}$ and $H_{\mathbf{x}} \leq H_{\mathbf{u}}$ under the semantic-sufficiency assumption. This dictates two optimization regimes:

**Case A (Mathematical Infeasibility of M5):** If the clinical constraint $\eta$ is highly stringent such that $H_{\mathbf{x}} \leq \eta < H_{\mathbf{u}}$, the required KL divergence for M5 becomes strictly negative ($D_{KL} = \eta - H_{\mathbf{u}} < 0$). Since this violates information-theoretic axioms, M5 mathematically cannot satisfy this constraint ($\mathcal{F}_5(\eta) = \emptyset$), rendering $\inf \mathcal{R}_{Brier}(f_5) = +\infty$. Conversely, M7 safely satisfies the constraint.

**Case B (Brier Risk Dominance):** If the constraint is relaxed ($\eta \geq H_{\mathbf{u}}$), both architectures are feasible. The disparity in their risk bounds is:

$$\Delta\mathcal{R}_{bound} = \Big(V_{\mathbf{u}} - V_{\mathbf{x}}\Big) - c\Big(H_{\mathbf{u}} - H_{\mathbf{x}}\Big) \tag{41}$$

For a standard probability clipping (e.g., $\tau = 10^{-3}$), the coefficient $c \approx 10^{-3}$ is highly attenuated. Thus, the magnitude of the difference in the irreducible variance ($V_{\mathbf{u}} - V_{\mathbf{x}}$) vastly dominates the scaled entropy difference. When the variance reduction term dominates the scaled entropy term, i.e., $\Delta\mathcal{R}_{bound} > 0$, the lower-bound comparison favors M7 over M5. This should be interpreted as a conditional bound comparison rather than an unconditional guarantee on the true risks.

**4. Synthesis:** By formulating the comparison under identical linear complexities (Step 1), we equated their Estimation Errors. By enforcing the strict Log-loss constraint (Step 3), we proved that because M5 lacks causal clinical semantics, forcing it to achieve high diagnostic accuracy either violates entropy limits or exposes a drastically higher irreducible variance floor. M7, anchored by clinical features, structurally minimizes this constrained lower bound. □

### F.3.3 Evaluating M6, M7, M8, and M9: From Zero-Shot LLMs to Calibrated ML

Having established the informational dominance of explicitly extracted clinical semantics ($\mathbf{x}_{\text{aug}}$), we now formally evaluate the architectural mechanisms used to map this information state to final diagnostic probabilities. We compare four distinct architectures operating over the same input manifold $\mathbf{x}_{\text{aug}}$: 1) **M6**: The zero-shot Specialist Divide & Conquer LLM. 2) **M7**: Base ML integration ($L_2$-regularized linear baseline). 3) **M8**: Post-hoc calibration over M7. 4) **M9**: Joint Loss calibration (C²Align).

Under Theorem 4, utilizing the BS ($L_\ell = 2$, $B = 1$) and given confidence $\delta$, the total generalization error is bounded by:

$$\mathcal{R}_{Brier}(\hat{f}) \leq \underbrace{\inf_{f\in\mathcal{F}} \mathcal{R}_{Brier}(f)}_{\text{Approximation Error}} + \underbrace{4\hat{\mathfrak{R}}_N(\mathcal{F})}_{\text{Estimation Error}} + 4\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}} \tag{42}$$

**Proposition 3** (Architectural Optimization over Clinical Semantics)**.** *When restricted to identical base hypothesis capacities, M8 expands the underlying parameter space via sequential post-hoc scaling, increasing the estimation error bound. M9 acts as a topological regularizer via the Joint Loss formulation, compressing the Rademacher complexity, thereby tightening the generalization upper bound under the norm-control condition $C_{w,9} \leq C_w$.*

*Proof.* Let the true Bayes-optimal posterior be $p^*(\mathbf{x}_{\text{aug}}) = \mathbb{P}(\mathbf{Y} = 1 \mid \mathbf{x}_{\text{aug}})$. Let the irreducible Bayes risk floor be $V_{\mathbf{x}} = \mathbb{E}[\text{Var}(\mathbf{Y} \mid \mathbf{x}_{\text{aug}})]$.

**1. Estimation Error (Rademacher Complexity):**

*For M6:* As a frozen pre-trained LLM without parameterized training on the empirical dataset $S_N$, its hypothesis space is a singleton set. The empirical supremum over a single function yields zero random variation, making its capacity to overfit mathematically zero:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_6) = 0 \tag{43}$$

*For M7:* Consistent with our prior analysis, we define the base hypothesis class as an $L_2$-regularized linear function mapped through a 1-Lipschitz activation $\sigma$: $\mathcal{F}_7 = \{\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}_{\text{aug}}) \mid \|\mathbf{w}\|_2 \leq C_w\}$. Its empirical Rademacher complexity is bounded by:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_7) \leq \frac{C_w R_x}{\sqrt{N}} \tag{44}$$

where $R_x = \max_i \|\mathbf{x}_{\text{aug},i}\|_2$.

*For M8:* M8 applies a post-hoc calibrator over the trained M7 model. When non-parametric post-hoc calibration (e.g., Isotonic Regression) is performed sequentially on the same training dataset $S_N$ without sample splitting, the hypothesis class effectively becomes a composition $\mathcal{F}_8 = \{g \circ f \mid f \in \mathcal{F}_7, g \in \mathcal{G}_{\text{mono}}\}$, where $\mathcal{G}_{\text{mono}}$ is the family of monotonic functions bounded in $[0, 1]$. By the composition properties of Rademacher complexity, applying a non-parametric calibrator on the empirical data amplifies the capacity to overfit. As a working assumption (this is not a general property of Rademacher complexity), we treat the calibrator's contribution as additive:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_8) \leq \hat{\mathfrak{R}}_N(\mathcal{F}_7) + \hat{\mathfrak{R}}_N(\mathcal{G}_{\text{cal}}), \tag{45}$$

where $\mathcal{G}_{\text{cal}}$ is the chosen calibrator class. This additional term depends on the calibrator and fitting protocol. With held-out calibration or strong regularization, the additional capacity can be controlled; without sample splitting, flexible post-hoc calibration may overfit calibration residuals.

*For M9:* M9 optimizes the same base linear capacity $\mathcal{F}_7$, but is implicitly regularized by the multi-objective dynamics: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dis}} + \alpha\mathcal{L}_{\text{Cal}} + \mathcal{L}_{\text{Align}}$. In the unconstrained M7 architecture, relying

purely on the Discrimination loss (Cross-Entropy) against hard labels $\mathbf{Y} \in \{0, 1\}$ mathematically incentivizes the optimizer to push the logit magnitudes toward infinity ($\|\mathbf{w}^\top \mathbf{x}_{\text{aug}}\| \to \infty$) to maximize margins. This logit inflation drives the empirical weight norm to the maximal allowable radius $C_w$. Conversely, M9 incorporates KL-divergence penalties ($\mathcal{L}_{\text{Align}}$) against the dual agentic priors. Crucially, these priors output *soft labels* strictly bounded away from the extremes ($\mathbf{p}_{\text{par}}, \mathbf{p}_{\text{spe}} \in (0, 1)$). The optimal logit targets for the dual-KL term are therefore intrinsically finite. During optimization, the dual-KL penalty acts as an implicit Euclidean regularizer, terminating unbounded logit inflation. It anchors the optimization trajectory, bounding the empirical parameters to a compressed effective radius, $C_{w,9} \leq C_w$. Substituting this constrained norm back into the Rademacher formula, M9 achieves a tighter estimation error bound:

$$\hat{\mathfrak{R}}_N(\mathcal{F}_9) \leq \frac{C_{w,9} R_x}{\sqrt{N}} \leq \frac{C_w R_x}{\sqrt{N}} \tag{46}$$

**2. Approximation Error (Bayes Risk):** By the standard orthogonal decomposition of the BS, we expand the Approximation Error for any function $f$ operating over $\mathbf{x}_{\text{aug}}$:

$$\inf_{f \in \mathcal{F}} \mathcal{R}_{Brier}(f) = \underbrace{V_{\mathbf{x}}}_{\text{Irreducible Variance}} + \underbrace{\inf_{f \in \mathcal{F}} \mathbb{E}\left[(p^*(\mathbf{x}_{\text{aug}}) - f(\mathbf{x}_{\text{aug}}))^2\right]}_{\text{Structural Bias } (B_{\mathcal{F}})} \tag{47}$$

No linear model can perfectly match the non-linear true posterior $p^*$. Thus, M7 and M9 share an identical, positive irreducible structural bias native to the linear hypothesis class, which post-hoc calibration (M8) cannot increase::

$$\inf_{f \in \mathcal{F}_8} \mathcal{R}_{Brier}(f) \leq \inf_{f \in \mathcal{F}_7} \mathcal{R}_{Brier}(f) = \inf_{f \in \mathcal{F}_9} \mathcal{R}_{Brier}(f) = V_{\mathbf{x}} + B_{\text{lin}} \tag{48}$$

Conversely, M6 relies entirely on a frozen zero-shot LLM mapping $p_{\text{M6}}$. Without empirical gradient descent over the clinical dataset, it suffers from an architectural calibration gap $B_6 = \mathbb{E}[(p^* - p_{\text{M6}})^2]$. We assume that empirically optimized linear functions adapt to domain marginals better than zero-shot LLMs, i.e., $B_{\text{lin}} < B_6$. Thus, the approximation error for M6 is structurally higher:

$$V_{\mathbf{x}} + B_{\text{lin}} < V_{\mathbf{x}} + B_6 \tag{49}$$

**3. Brier Risk under Accuracy Constraints:** To rigorously benchmark the Brier risks, we enforce an exactly equal Information-Theoretic accuracy by fixing the cross-entropy Log-loss to a level $\eta$. Because M6, M7, M8, and M9 all operate over the same manifold $\mathbf{x}_{\text{aug}}$, their intrinsic irreducible variance $V_{\mathbf{x}}$ and conditional entropy $H_{\mathbf{x}}$ are identical. Invoking Lemma 5, the Approximation Error is bounded from below:

$$\inf_{f \in \mathcal{F}(\eta)} \mathcal{R}_{Brier}(f) \geq \underbrace{V_{\mathbf{x}} + \tau(1-\tau)\Big(\eta - H_{\mathbf{x}}\Big)}_{\mathcal{C}_{\text{bound}}(\eta)} \tag{50}$$

We evaluate the feasibility and Approximation Error for each architecture under this universal constraint $\mathcal{C}_{\text{bound}}(\eta)$:

- **M7, M8, and M9:** As parameterized functional approximators trained empirically over $\mathbf{x}_{\text{aug}}$, their hypothesis spaces natively allow convergence toward this minimum. Their constrained approximation errors satisfy $A_8(\eta) \leq A_7(\eta) = A_9(\eta) =: A_{\text{lin}}(\eta) \geq \mathcal{C}_{\text{bound}}(\eta)$, where $A_m(\eta) = \inf_{f \in \mathcal{F}_m(\eta)} \mathcal{R}_{Brier}(f)$.
- **M6 (Zero-Shot LLM):** M6 utilizes a frozen mapping, dictating a fixed empirical Log-loss $\eta_6$. If the clinical requirement is stricter than the LLM's zero-shot capability ($\eta < \eta_6$), M6 cannot satisfy the constraint ($\mathcal{F}_6(\eta) = \emptyset$), forcing its constrained risk to diverge: $\inf \mathcal{R}_{Brier}(f_6) = +\infty$. Even if feasible ($\eta \geq \eta_6$), its zero-shot structural bias ($B_6 > B_{\text{lin}}$) bounds its Brier risk above the empirically optimized models.

**4. Synthesis of the Total Generalization Risk:** We assemble the final generalization bounds by substituting the constrained Approximation Errors (Step 3) and Estimation Errors (Step 1) into Equation 42. Assuming the clinical constraint $\eta$ renders M6 strictly suboptimal, we evaluate the remaining architectures:

- **M7 (Base ML):** Satisfies the accuracy constraint to reach the baseline approximation bound, yielding the estimation penalty governed by $C_w$:

$$\mathcal{R}(\hat{f}_7) \leq A_{\text{lin}}(\eta) + 4\frac{C_w R_x}{\sqrt{N}} + 4\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}} \tag{51}$$

- **M8 (Post-hoc Calibration):** Shares the exact same approximation bound $A_{\text{lin}}(\eta)$ (an upper bound for M8 as well, since $A_8(\eta) \leq A_{\text{lin}}(\eta)$), but sequentially accumulates the VC-subgraph dimension of the non-parametric calibrator, inflating the estimation error (the $\mathcal{O}(N^{-1/3})$ term is a heuristic rate for isotonic regression under the working assumption above, not a derived bound):

$$\mathcal{R}(\hat{f}_8) \leq A_{\text{lin}}(\eta) + 4\frac{C_w R_x}{\sqrt{N}} + \mathcal{O}(N^{-1/3}) + 4\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}} \tag{52}$$

- **M9 (Joint Loss / C²Align):** Satisfies the accuracy constraint while suppressing the Rademacher complexity via the restricted topological norm $C_{w,9}$:

$$\mathcal{R}(\hat{f}_9) \leq A_{\text{lin}}(\eta) + 4\frac{C_{w,9} R_x}{\sqrt{N}} + 4\sqrt{\frac{\log(4/\delta_{\text{conf}})}{2N}} \tag{53}$$

Under the effective norm-control condition $C_{w,9} \leq C_w$, the complexity term of M9 is no larger than that of M7:

$$4\frac{C_{w,9} R_x}{\sqrt{N}} \leq 4\frac{C_w R_x}{\sqrt{N}}. \tag{54}$$

Thus, M9 has a no-larger generalization upper bound than M7 under this condition. For M8, an additional calibrator-capacity term may appear when post-hoc calibration is fitted on the same training set without sample splitting. Therefore, the analysis supports a conditional upper-bound comparison rather than an unconditional ordering of the true risks.

This derivation shows that Joint Loss distillation can tighten the estimation-error term when the alignment loss induces a smaller effective parameter norm. □

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the formulation, the conditional theoretical analysis, the method, and the empirical results; Sections 3 and 4 support them, and the theory is presented as conditional motivation (Sections 3.1.1 and 6).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the theoretical assumptions and failure cases, data overlap and sample sizes, residual miscalibration, interpretability, scope, and cost.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem states its assumptions (Section 3.1.1); complete proofs are in Appendix F.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix C describe cohorts, splits, features, agent configuration, the custom objective, and hyperparameters; Appendix E gives the prompts.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available. MIMIC-III, MIMIC-IV, TJH are also available after dataset application (Appendix B).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix C.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4.2, Appendix D.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D reports per-patient tokens, cost, and latency of LLM inference; the downstream XGBoost model is lightweight.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study was IRB-approved and used credentialed data under the applicable data-use agreements in privacy-controlled inference environments (Appendix B).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix A.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release code only; no patient data, pretrained models, or scraped datasets are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: MIMIC-III, MIMIC-IV, TJH, and all LLMs are cited; MIMIC data are used under the PhysioNet Credentialed Health Data Use Agreement and LLMs under their providers' enterprise terms (Appendix B).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code repository documents cohort construction, prompts, schemas, the feature policy, and the training objective (Appendix B).

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study uses retrospective records and involves no crowdsourcing or experiments with human participants; physician annotation was performed by the study team.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Appendix B.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: LLM agents are a core component of the method; their roles, configuration, and prompts are described in Section 3 and Appendices C and E.