On-Policy Hindsight Distillation for Early Risk Prediction

Qiannan Zhang, Fei Wang*
Weill Cornell Medicine, Cornell University
{qiz4005,few2001}@med.cornell.edu

Abstract

Early prediction of chronic disease risk from electronic health records (EHRs) is challenging because early clinical signals are often sparse, noisy, and indirectly related to future outcomes. Existing methods typically learn from prediction-time records and final outcome labels alone, which provides limited supervision for identifying weak early signals from heterogeneous and confounded clinical records. LLM-generated rationales offer an intermediate form of prediction-time reasoning, but rationales derived from early EHR may still miss weak signals whose relevance becomes clearer only in later records. We therefore use follow-up EHR observed after the prediction time but before outcome assessment as *training-time hindsight* and propose *On-Policy Hindsight Distillation* (OPHD), a self-distillation framework that converts follow-up EHR into training signals for LLM-generated prediction-time rationales. Specifically, OPHD first uses an LLM to generate a rationale from prediction-time EHR alone, then re-evaluates the same rationale under privileged follow-up views to derive token-level signals that are distilled back to the prediction-time policy. To handle noisy follow-up records, OPHD emphasizes hindsight signals that are stable across perturbed privileged views. A task-grounded scorer further prioritizes rationales that improve downstream risk discrimination. Experiments on multiple real-world neurodegenerative disease cohorts show that OPHD achieves the best long-horizon prediction performance compared with a variety of baselines.

1 Introduction

Early prediction of chronic diseases from electronic health records (EHRs) is clinically important [21, 28, 14] because routinely collected EHR data are widely available, cost-effective, and of a longitudinal nature. Identifying at-risk patients before diagnosis may support earlier monitoring and intervention, especially for slowly progressive neurodegenerative conditions such as Alzheimer's disease [24, 50]. Existing EHR-based prediction methods typically model future risk directly from *prediction-time records*, largely optimizing the same objective: mapping recorded clinical events to outcome labels [49, 57]. Although these methods may differ in how they encode patient medical histories, this direct mapping is difficult for long-horizon prediction because, when prediction is made *far* in advance of the outcome, disease-relevant evidence is often sparse, incomplete, and clinically nonspecific [7]. Moreover, early EHR events are usually shaped by healthcare utilization and documentation practices rather than direct measures of disease progression [47, 11]. As a result, models may rely on stable but nonspecific correlates, e.g., generic comorbidities, instead of clinically meaningful early signals.

LLM-generated rationales provide an intermediate representation for reasoning over prediction-time EHR, instead of directly producing a risk score [5, 29, 35, 34, 23]. In disease risk prediction,

*Corresponding Author

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

pretrained medical knowledge in LLMs may help interpret the patient history, connect clinical evidence to possible future risk, and make indirect early signals more explicit [60]. However, long-horizon risk prediction presents a more difficult setting: rationales based only on prediction-time records may remain generic, overemphasize common comorbidities, or miss weak but outcome-informative signals. We refer to this mismatch as the *reasoning-prediction gap*: a rationale may appear clinically plausible yet fail to improve downstream risk discrimination.

The retrospective cohort formulation [50, 13], commonly adopted for EHR-based early prediction, provides a natural source of *training-time hindsight*: follow-up EHR observed after the prediction time but before outcome assessment can reveal how prediction-time patterns unfold before the outcome is determined. This mirrors how clinicians learn from patient follow-up: after an initial assessment, later visits may reveal which early impressions were supported, which were incidental, and which should be revised. Although such future records are unavailable when assessing a new patient, accumulated follow-up experience can improve how similar early evidence is interpreted in future patients. This is particularly useful for addressing the reasoning-prediction gap: follow-up EHR reveals whether an LLM-generated rationale captures early patterns that later become disease-relevant. Follow-up records provide richer intermediate evidence for learning prediction-time rationales, creating an opportunity to improve reasoning while maintaining a prediction setting based only on prediction-time EHR.

However, follow-up EHR cannot be used as rationale supervision in a straightforward manner. Directly conditioning the model on follow-up records during training would encourage reliance on evidence unavailable at inference, leading to a *train-test mismatch.* At the same time, follow-up records are not ground-truth explanations of the outcome: they may contain EHR routine-care noise rather than generalizable disease signals. Thus, the challenge is to convert imperfect follow-up EHR into robust training signals that improve prediction-time reasoning without depending on future-only evidence at test time.

To address this challenge, we propose *On-Policy Hindsight Distillation* (OPHD), a training framework that transforms follow-up EHR into robust supervision for prediction-time reasoning without exposing follow-up information at inference. The key idea is to keep the rationale trajectory *on-policy*: OPHD first generates a rationale with an LLM using only prediction-time EHR, under the same information condition used at inference. Instead of generating a separate follow-up-conditioned rationale, which may encode future-only evidence, OPHD re-evaluates the same prediction-time rationale under privileged views constructed with follow-up EHR. Hindsight is therefore transferred through token-level preference shifts along the model's own reasoning trajectory. To account for noisy evidence in follow-up EHR, OPHD compares a prediction-time reference view with multiple perturbed privileged views and concentrates distillation on positions where privileged information induces salient and stable changes. Finally, OPHD closes the loop with a task-grounded scorer that maps the prediction-time EHR and generated rationale to the outcome label. The scorer provides supervised feedback on whether the rationale helps distinguish disease outcomes and feeds this task-specific signal back into rationale generation. In this way, OPHD trains rationale generation to incorporate reliable evidence from follow-up EHR while remaining aligned with the downstream risk discrimination.

Our contributions are threefold. First, we study the clinically important yet challenging early prediction problem and propose OPHD, an on-policy hindsight distillation framework that utilizes follow-up EHR as training-time privileged evidence to refine prediction-time rationales, without changing the inference-time input. Second, we develop robust multi-view token-level distillation that capitalizes on informative hindsight signals and mitigates EHR noise by comparing a prediction-time reference view with perturbed privileged views. Third, we introduce a task-grounded scorer that feeds supervised prediction feedback back into rationale generation, and show that OPHD achieves the best long-horizon prediction performance across multiple real-world neurodegenerative disease cohorts.

2 Related Work

EHR-based Clinical Risk Prediction. Clinical risk prediction from EHR has been widely studied. Existing methods span engineered-feature models [11, 54, 39], sequential models for longitudinal events [3, 9, 31], graph-based models over patients and medical concepts [4, 30, 17], and EHR foundation models [26, 57, 49]. These approaches generally learn risk functions from historical records available up to the prediction time. Long-horizon early prediction remains especially challenging

because disease-relevant signals in prediction-time EHR can be sparse, indirect, and confounded by healthcare utilization and documentation practices [16, 10].

LLMs for Clinical Reasoning and Prediction. LLMs have been increasingly used in clinical applications, including medical question answering, report interpretation, patient-trial matching, and decision support [45, 46, 19]. Recent studies further use LLMs for disease prediction from EHR through direct prompting and fine-tuning [1, 6, 22], rationale generation and multi-agent reasoning [6, 25], or LLM-assisted representation construction [2, 56, 35, 18]. However, pretrained medical knowledge and prediction-time prompting alone may produce rationales that are clinically plausible but weakly tied to downstream risk discrimination. Such rationales can remain generic, overemphasize common comorbidities, or miss subtle early signals whose relevance is difficult to infer from prediction-time records alone. This creates a reasoning-prediction gap between plausible clinical explanations and task-useful rationales for long-horizon risk prediction.

LLM Reasoning and Distillation. General LLM reasoning methods improve model outputs through chain-of-thought prompting, self-consistency, and verifier-guided selection, which elicit or aggregate multiple reasoning paths at inference time [53, 51, 27]. Feedback-based learning and distillation methods further optimize model behavior using human preferences, reward models, teacher outputs, generated rationales, or pseudo-targets [36, 38, 12, 20, 32, 59]. These methods provide useful mechanisms for eliciting, selecting, and refining reasoning, but they do not directly provide supervision for identifying which weak early clinical signals are relevant to future outcomes. Inference-time sampling can generate multiple candidate rationales, but it does not teach the model which early clinical signals are outcome-relevant; outcome-level rewards provide sparse sequence-level feedback; and directly imitating follow-up-conditioned rationales can introduce train-test mismatch by teaching the model to rely on evidence unavailable at inference.

OPHD targets a supervision gap in long-horizon EHR prediction: early records contain weak and indirect evidence, whereas follow-up EHR reveal how these signals evolve before the outcome. Instead of using follow-up EHR as an inference-time input or directly imitating follow-up-conditioned rationales, OPHD uses it only during training to derive hindsight feedback for prediction-time rationales. A task-grounded scorer further weights this feedback, aligning rationale generation with downstream risk discrimination rather than clinical plausibility alone. This distinguishes OPHD from conventional EHR risk models, prompt-based clinical LLM methods, and generic RL or distillation approaches, which respectively rely on direct record-to-outcome learning, inference-time reasoning from available records, or sparse rewards and fixed teacher targets.

3 Methodology

3.1 Problem Formulation

We follow a common EHR-based early prediction setup and construct the case-control cohort for disease risk prediction. For each patient i, let $Y_i \in \{0, 1\}$ denote the target outcome, where $Y_i = 1$ for cases and $Y_i = 0$ otherwise. Let I_i denote the *index time*, defined as the recorded onset time for cases and an analogous reference time for controls. The goal is to predict Y_i using EHR observed at an earlier prediction time $t_i^0 < I_i$ before the index time. We denote by $\mathcal{X}_i$ the prediction-time EHR of patient i, consisting of all records observed up to t_i^0. During training, we additionally observe follow-up EHR $\Delta\mathcal{X}_i$ collected over $(t_i^0, t_i^1]$, where t_i^1 is a follow-up cutoff time and $t_i^0 < t_i^1 < I_i$. At inference, only $\mathcal{X}_i$ is available, and $\Delta\mathcal{X}_i$ is never used as direct prediction input. The inference-time objective is therefore to estimate $P(Y_i \mid \mathcal{X}_i)$. In our framework, an LLM first generates a rationale R_i from prediction-time EHR $\mathcal{X}_i$, and a task-grounded scorer $g_\phi(\cdot)$ estimates the outcome label from $\mathcal{X}_i$ and R_i, denoted as $\hat{Y}_i = g_\phi(\mathcal{X}_i, R_i)$. The training objective is to use $\Delta\mathcal{X}_i$ as privileged evidence to improve rationale generation, while ensuring that both rationale generation and final prediction depend only on $\mathcal{X}_i$ at inference.

3.2 Overview of OPHD

OPHD comprises three coordinated components. First, LLM π_θ generates an on-policy rationale from patient's prediction-time EHR alone, ensuring that the reasoning trajectory is purely based on information available before prediction. Second, OPHD re-scores the same rationale under the prediction-time view and multiple perturbed privileged views constructed from follow-up EHR,

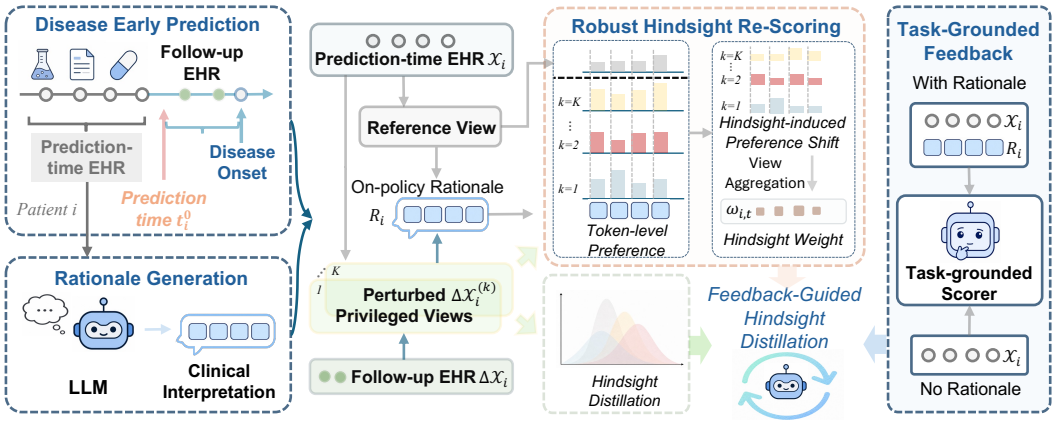


Figure 1: Overview of OPHD. OPHD uses follow-up EHR only during training as privileged hindsight. An on-policy rationale is generated from prediction-time EHR, and the same rationale is then re-scored under reference and perturbed privileged views to obtain robust hindsight signals. These signals are modulated with task-grounded feedback from the scorer to help distill hindsight to the rationale generator. At inference time, OPHD leverages only prediction-time EHR to generate a rationale and produce the final outcome label with the scorer.

identifying token-level hindsight signals that are salient and stable across perturbations. Third, a task-grounded scorer estimates the predictive utility of the generated rationale and feeds this supervised signal back into rationale learning. Importantly, OPHD does not distill from a separately generated follow-up-conditioned rationale. Instead, it preserves the same prediction-time rationale and compares how it is evaluated under different information views. This design transfers hindsight to prediction-time reasoning without training the model to imitate explanations that rely on information unavailable at inference time. Figure 1 illustrates the overall framework.

3.3 Prediction-Time On-Policy Rationale

Given the prediction-time EHR $\mathcal{X}_i$ and an LLM rationale generator π_θ, OPHD first uses π_θ to generate a rationale with only information available at inference time. The rationale is sampled autoregressively as:

$$R_i = r_{i,1:T_i} \sim \pi_\theta(\cdot \mid \mathcal{X}_i, P_R), \tag{1}$$

where P_R denotes the rationale-generation prompt and T_i is the rationale length. It captures the model's initial interpretation of the patient's early clinical history with no access to follow-up EHR $\Delta\mathcal{X}_i$, such as potentially risk-relevant patterns from diagnoses, drug prescriptions, and other recorded events. Since long-horizon prediction often relies on sparse and nonspecific evidence, this rationale may remain generic or fail to emphasize the disease-relevant early signals. The generated rationale serves as the on-policy trajectory for subsequent hindsight distillation. OPHD keeps this trajectory fixed and later re-scores it under different information views, rather than replacing it with a separately generated follow-up-conditioned rationale.

3.4 Robust Hindsight Re-Scoring

After on-policy rational generation, OPHD re-scores the realized token trajectory $R_i = r_{i,1:T_i}$ to elucidate hindsight knowledge from privileged follow-up information during training. In implementation, these re-scoring views are instantiated using a frozen copy of the LLM denoted as F_θ, while the rationale generator π_θ is updated by the distillation process. We use the terms *reference view* and *privileged view* to emphasize that the views differ by available information, not by generating different rationales. The prediction-time reference view observes only $\mathcal{X}_i$ and provides an inference-consistent evaluation of the rationale as:

$$z_{i,t}^0 = F_\theta\left(r_{i,<t}, \mathcal{X}_i, P_T\right), \tag{2}$$

where P_T denotes the re-scoring prompt and $z_{i,t}^0 \in \mathbb{R}^{|\mathcal{V}|}$ denotes the next-token logits over vocabulary $\mathcal{V}$. To incorporate training-time hindsight, we construct K privileged views via perturbing the follow-

up EHR. Let $\mathcal{T}$ denote a stochastic perturbation operator applied to $\Delta\mathcal{X}_i$, such as EHR event dropout, we obtain $\Delta\mathcal{X}_i^{(k)} \sim \mathcal{T}(\Delta\mathcal{X}_i)$ and re-score the same rationale trajectory for each $k = 1, \ldots, K$:

$$z_{i,t}^{P,k} = F_\theta\left(r_{i,<t}, \mathcal{X}_i, \Delta\mathcal{X}_i^{(k)}, P_T\right). \tag{3}$$

Importantly, privileged views do not generate new rationales; instead, they re-evaluate the same prediction-time rationale. As conditioned on train-time follow-up EHR information, they can reveal how hindsight evidence changes the model's token-level preference. Let $P_{i,t}^0$ and $P_{i,t}^{P,k}$ denote the next-token distributions under the prediction-time view and the k-th perturbed view, respectively. We measure the hindsight-induced preference shift for the realized token $r_{i,t}$ by the log-probability difference:

$$\delta_{i,k,t} = \log P_{i,t}^{P,k}(r_{i,t}) - \log P_{i,t}^0(r_{i,t}). \tag{4}$$

A large absolute shift indicates that privileged follow-up information substantially alters the model's preference for the token relative to the prediction-time reference view. To mitigate noise in follow-up EHR, we aggregate the shifts across K perturbed views and compute: $\mu_{i,t} = \frac{1}{K}\sum_{k=1}^{K}\delta_{i,k,t}$ and $\sigma_{i,t} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\delta_{i,k,t} - \mu_{i,t}\right)^2}$. The final hindsight weight at position t is defined as follows:

$$w_{i,t} = |\mu_{i,t}| \cdot \frac{1}{1 + \lambda\sigma_{i,t}}, \tag{5}$$

where $\lambda \geq 0$ controls the penalty on unstable shifts. Thus, $w_{i,t}$ is large when follow-up EHR evidence induces a high, stable token-level preference shift across perturbed privileged views, and small when the hindsight signal is weak or unstable.

3.5 Feedback-Guided Hindsight Distillation

Task-Grounded Feedback Robust hindsight re-scoring identifies token positions where follow-up EHR evidence induces salient and stable preference shifts. However, such shifts alone are insufficient to determine whether the corresponding hindsight signal should be distilled. When the generated rationale offers a useful interpretation of the prediction-time EHR, follow-up evidence can refine it by strengthening or weakening model's preference for specific tokens. Yet, if the rationale provides little support for the target label, it may lack a reliable prediction-time anchor for hindsight correction. Follow-up-induced token shifts may then reflect future-specific events rather than associations that can be inferred from early EHR. Directly distilling such shifts risks teaching the prediction-time policy to imitate information unavailable at inference, instead of learning how weak early signals relate to later disease evolution.

To align token-level hindsight distillation with cohort-specific risk discrimination, we utilize the task-grounded scorer g_ϕ as a source of task-level feedback. Trained with a supervised cross-entropy objective $\mathcal{L}_{\text{score}}$, the scorer evaluates whether a generated rationale provides prediction-time evidence that supports the target label. From a probabilistic perspective, a rationale is useful when it provides more information beyond prediction-time EHR. In other words, a rationale should increase support for the observed label under $p_\phi(Y_i \mid \mathcal{X}_i, R_i)$, relative to the rationale-free prediction $p_\phi(Y_i \mid \mathcal{X}_i)$. Motivated by this, we define a scorer-derived rationale weight as the gain in label support by including the rationale as:

$$a_i = \vartheta\left(\log p_\phi(Y_i \mid \mathcal{X}_i, R_i) - \log p_\phi(Y_i \mid \mathcal{X}_i)\right). \tag{6}$$

where ϑ denotes activation for which we use ReLU. Here, $p_\phi(Y_i \mid \mathcal{X}_i, R_i)$ is the scorer probability assigned to the observed label provided with both prediction-time EHR and the rationale, whereas $p_\phi(Y_i \mid \mathcal{X}_i)$ is computed by applying the same scorer with a null rationale input. When a_i is used to weight the distillation loss, we detach it from the computational graph so that the scorer is optimized only through the supervised prediction loss.

Thus, a_i increases when the rationale improves scorer's support for the observed label, and decreases when the rationale provides little discriminative information. This scorer-derived feedback links rationale generation and downstream prediction, allowing OPHD to upweight rollouts whose rationales improve discrimination between future cases and controls. Next, OPHD combines this rationale-level signal with the token-level robust hindsight distillation.

Hindsight Distillation The robust hindsight weight $w_{i,t}$ captures whether follow-up evidence stably changes the model's preference for a reasoning token, while the scorer-derived rationale weight a_i captures whether the full rationale increases support for the observed outcome. We combine them into a final relevance weight $\tilde{w}_{i,t} = a_i w_{i,t}$, which emphasizes reasoning components that are both hindsight-informed and prediction-relevant.

To conduct token-level distillation, we re-evaluate the same on-policy rationale trajectory using the trainable rationale generator π_θ with prediction-time EHR. Specifically, we compare the resulting next-token distribution $P^{\theta}_{i,t}$ from π_θ with each perturbed privileged distribution separately. For each view, we form the mixture distribution as $M_{i,k,t} = \frac{1}{2}\left(P^{\theta}_{i,t} + P^{P,k}_{i,t}\right)$ and compute the Jensen–Shannon divergence as:

$$d_{i,k,t} = \frac{1}{2} D_{\mathrm{KL}}\left(P^{\theta}_{i,t} \| M_{i,k,t}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(P^{P,k}_{i,t} \| M_{i,k,t}\right). \tag{7}$$

Because perturbed privileged views may introduce view-specific noise, we use JS divergence as a bounded, mixture-based discrepancy rather than directly minimizing a reverse-KL objective. This provides a more conservative distributional distillation signal. We aggregate the resulting view-specific discrepancies by $\bar{d}_{i,t} = \frac{1}{K}\sum_{k=1}^{K} d_{i,k,t}$. With a valid-token mask $m_{i,t} \in \{0,1\}$ that selects rollout tokens and excludes prompt and padding tokens, the feedback-guided hindsight distillation loss is then defined as:

$$\mathcal{L}_{\mathrm{OPHD}} = \frac{\sum_i \sum_{t=1}^{T_i} m_{i,t} \tilde{w}_{i,t} \bar{d}_{i,t}}{\sum_i \sum_{t=1}^{T_i} m_{i,t} \tilde{w}_{i,t} + \epsilon}, \tag{8}$$

where $\epsilon > 0$ is a small constant for numerical stability. To avoid degenerate batches in which all feedback-guided weights are close to zero, we normalize $\tilde{w}_{i,t}$ over valid rollout tokens within each mini-batch. This keeps the distillation signal active while still emphasizing tokens with stronger hindsight and task-grounded feedback.

3.6 Optimization and Inference

During training, we first warm up the scorer g_ϕ using prediction-time EHR $\mathcal{X}_i$ to optimize the supervised prediction loss $\mathcal{L}_{\mathrm{score}}$. This yields a stable rationale-free reference for estimating $p_\phi(Y_i \mid \mathcal{X}_i)$. The scorer is then used to compute the scorer-derived rationale weight a_i, which controls how strongly hindsight distillation should affect each generated rationale. As a_i is detached from the computational graph, distillation gradients update only the rationale generator. After updating the generator, we regenerate prediction-time rationales and fine-tune the scorer with $\mathcal{L}_{\mathrm{score}}$ using $\mathcal{X}_i$ and the updated rationales.

At inference time, follow-up EHR $\Delta\mathcal{X}_i$, perturbed privileged views, and hindsight re-scoring are not used. Given only the prediction-time EHR $\mathcal{X}_i$, OPHD first generates a rationale $R_i \sim \pi_\theta(\cdot \mid \mathcal{X}_i, P_R)$ using the learned rationale generator. The task-grounded scorer then produces the final risk prediction as $\hat{Y}_i = g_\phi(\mathcal{X}_i, R_i)$. Thus, the deployed model follows the same inference-time information boundary as standard early risk prediction models, while benefiting from follow-up-guided rationale learning during training.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate OPHD on three real-world EHR prediction tasks spanning major neurodegenerative outcomes: Alzheimer's disease (AD), Parkinson's disease (PD), and Alzheimer's disease and related dementias (ADRD). The AD and PD cohorts are constructed from the *All of Us* (AoU) Research Program [40], a nationwide U.S. EHR resource harmonized under the OMOP common data model [41]. For ADRD, we utilize the local clinical repository [33] standardized under the PCORnet common data model. For all three tasks, we construct case-control cohorts according to clinically established phenotype definitions and apply matching within the development split. Across all cohorts, model inputs are restricted to EHR records available before the predefined prediction time, while follow-up EHR is leveraged only during training as privileged evidence. Detailed cohort

Table 1: Performance comparison for different models on varying disease cohorts. The best results are highlighted in **bold**, while the strongest baseline scores are underlined.

Cohort	Model	AUROC	AUPRC	Sens@90
AD	GRAPE	0.559 ± 0.023	0.013 ± 0.001	0.190 ± 0.005
	MUSE	0.587 ± 0.009	0.016 ± 0.004	0.209 ± 0.017
	MOTOR	0.564 ± 0.017	0.013 ± 0.001	0.165 ± 0.008
	Zeroshot-LLM	0.608 ± 0.013	0.010 ± 0.007	0.090 ± 0.013
	Fewshot-LLM	0.616 ± 0.014	0.013 ± 0.004	0.123 ± 0.011
	Multi-sample	0.648 ± 0.016	0.018 ± 0.007	0.186 ± 0.015
	Best-of-K	0.633 ± 0.015	0.016 ± 0.003	0.157 ± 0.013
	ReConcile	0.628 ± 0.020	0.015 ± 0.002	0.145 ± 0.010
	DearLLM	0.616 ± 0.014	0.016 ± 0.005	0.158 ± 0.014
	OPHD	$\mathbf{0.681} \pm 0.011$	$\mathbf{0.024} \pm 0.007$	$\mathbf{0.274} \pm 0.012$
PD	GRAPE	0.621 ± 0.009	0.019 ± 0.002	0.247 ± 0.013
	MUSE	0.633 ± 0.024	0.019 ± 0.001	0.215 ± 0.035
	MOTOR	0.586 ± 0.021	0.016 ± 0.002	0.178 ± 0.014
	Zeroshot-LLM	0.572 ± 0.016	0.010 ± 0.001	0.146 ± 0.010
	Fewshot-LLM	0.577 ± 0.019	0.012 ± 0.002	0.159 ± 0.015
	Multi-sample	0.624 ± 0.014	0.017 ± 0.002	0.221 ± 0.015
	Best-of-K	0.616 ± 0.018	0.016 ± 0.004	0.205 ± 0.022
	ReConcile	0.593 ± 0.022	0.013 ± 0.002	0.184 ± 0.016
	DearLLM	0.627 ± 0.025	0.019 ± 0.005	0.195 ± 0.017
	OPHD	$\mathbf{0.664} \pm 0.024$	$\mathbf{0.024} \pm 0.003$	$\mathbf{0.292} \pm 0.020$
ADRD	GRAPE	0.575 ± 0.023	0.026 ± 0.001	0.180 ± 0.022
	MUSE	0.586 ± 0.021	0.024 ± 0.007	0.207 ± 0.034
	MOTOR	0.546 ± 0.017	0.017 ± 0.004	0.123 ± 0.012
	Zeroshot-LLM	0.555 ± 0.021	0.013 ± 0.003	0.070 ± 0.011
	Fewshot-LLM	0.567 ± 0.018	0.015 ± 0.005	0.085 ± 0.024
	Multi-sample	0.593 ± 0.017	0.018 ± 0.002	0.147 ± 0.015
	Best-of-K	0.587 ± 0.016	0.017 ± 0.003	0.123 ± 0.013
	ReConcile	0.559 ± 0.016	0.014 ± 0.005	0.089 ± 0.014
	DearLLM	0.566 ± 0.015	0.016 ± 0.004	0.106 ± 0.015
	OPHD	$\mathbf{0.636} \pm 0.009$	$\mathbf{0.030} \pm 0.005$	$\mathbf{0.234} \pm 0.017$

definitions, time-related criteria, matching procedures, and inclusion/exclusion rules are provided in Appendix A.

Baselines. We compare OPHD with three groups of baselines. Conventional EHR prediction models comprise GRAPE [58], MUSE [55], and MOTOR [49], which represent graph-based, multi-view, and foundation-model approaches for structured EHR prediction. LLM reasoning baselines include Zeroshot-LLM and Fewshot-LLM, which prompt the LLM to generate intermediate clinical rationales, as well as Multi-sample variants [52] and Best-of-K that sample multiple rationales for selection or aggregation. We also evaluate LLM-enhanced methods, including ReConcile [2] and DearLLM [56]. More baseline details are provided in Appendix C.

Experimental Details. We assess model performance using AUROC, AUPRC, and threshold-based metrics at 90% specificity. While AUROC measures overall discrimination, AUPRC is more informative when the outcome is rare, as is typical for neurodegenerative diseases in practice. We further report sensitivity at the 90% specificity level to evaluate performance under a stringent false-positive constraint considering clinical relevance.

The framework is implemented in PyTorch using Qwen3-4B-Instruct-2507 as the rationale generator and Qwen3-0.6B as the scorer. Models are adapted with LoRA [15], with hyperparameters determined by grid search. More experimental settings related to study design are provided in Appendix B and additional implementation details can be found in Appendix D.

4.2 Overall Performance

Table 1 reports the overall predictive performance across AD, PD, and ADRD cohorts. OPHD consistently achieves the best performance across all disease cohorts and evaluation metrics, outperforming both conventional EHR models and LLM-based reasoning baselines. Compared with conventional EHR models, the improvement suggests that directly learning representations from prediction-time records is insufficient for long-horizon prediction, where disease-relevant evidence is sparse and indirect. OPHD also outperforms LLM-based reasoning baselines, including zero-shot and few-shot prompting, multi-sample and best-of-K reasoning, ReConcile, and DearLLM. This indicates that

Table 2: Ablation study of OPHD components. Performance is reported on three evaluation cohorts.

Variant	AD		PD		ADRD	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
OPHD	**0.6814**	**0.0236**	**0.6637**	**0.0238**	**0.6360**	**0.0300**
Scorer w/o Rationale	0.5934	0.0096	0.5328	0.0092	0.5828	0.0177
w/o Follow-up EHR	0.6476	0.0144	0.6195	0.0149	0.5764	0.0162
w/o Perturbation	0.6484	0.0146	0.6415	0.0127	0.6279	0.0195
w/o Scorer Feedback ($a_i = 1$)	0.6596	0.0195	0.6467	0.0174	0.6248	0.0233
w/o Token Weighting ($w_{i,t} = 1$)	0.6250	0.0132	0.5763	0.0128	0.6035	0.0194
Uniform Privileged Distillation	0.6029	0.0117	0.5543	0.0109	0.5926	0.0184

simply eliciting rationales from an LLM or aggregating multiple generated explanations does not fully address the reasoning-prediction gap. In contrast, OPHD uses follow-up-guided re-scoring and task-grounded feedback to selectively refine rationales that are more useful for downstream risk discrimination. In addition, the gains are observed not only in threshold-free discrimination metrics such as AUROC and AUPRC, but also in clinically relevant operating-point metric, i.e., sensitivity at 90% specificity, which is particularly important in early risk prediction with rare positive samples. The consistent improvements under this stringent specificity constraint suggest that OPHD better identifies future cases while maintaining controlled false-positive rates. We perform paired t-tests on per-patient prediction scores to assess whether the observed improvements are statistically reliable. The improvements are found to be statistically significant ($p < 0.05$), confirming that the gains are unlikely to be due to random variation. Overall, the results support the effectiveness of using follow-up EHR as training-time hindsight to improve prediction-time reasoning and risk prediction.

4.3 Ablation Study

We conduct ablation studies to examine how each design choice contributes to OPHD. To assess the role of training-time hindsight, we compare OPHD with a variant that trains without follow-up EHR. To evaluate the robustness mechanism, we replace the perturbed privileged views with a single follow-up-conditioned view, thereby testing whether multi-view perturbation helps suppress unstable hindsight signals. To study the effect of task-grounded feedback, we set $a_i = 1$, so that hindsight distillation is no longer weighted by the scorer-derived rationale feedback. We further set $w_{i,t} = 1$ to test whether token-level hindsight weighting is necessary, forcing all reasoning tokens to receive equal distillation weight. Finally, we include a uniform privileged distillation variant with both $a_i = 1$ and $w_{i,t} = 1$, which tests whether the performance gain comes from the proposed weighting mechanism rather than privileged distillation alone. Table 2 shows that each OPHD component contributes to the final performance. Training without follow-up EHR leads to lower performance, confirming the value of training-time hindsight. Replacing perturbed privileged views with a single follow-up-conditioned view also degrades performance, suggesting that robust multi-view estimation helps suppress unstable hindsight signals. Disabling scorer feedback by setting $a_i = 1$ weakens the model, indicating that follow-up-induced shifts should be grounded in downstream prediction utility. Setting $w_{i,t} = 1$ similarly reduces performance, showing that not all reasoning tokens should receive equal distillation weight. Finally, uniform privileged distillation performs worse than OPHD, demonstrating that the gains come from selectively weighting hindsight signals rather than privileged distillation alone.

4.4 Further Analysis

Robustness to perturbation number K. We vary the number of perturbed privileged views K to examine the robustness of multi-view hindsight estimation. As shown in Figure 2, performance generally improves after introducing perturbed follow-up views compared with using no perturbation or only a single privileged view. The improvement is consistent across AD, PD, and ADRD, suggesting that perturbation helps reduce sensitivity to noisy follow-up events. The best overall performance is achieved around $K = 2$, indicating

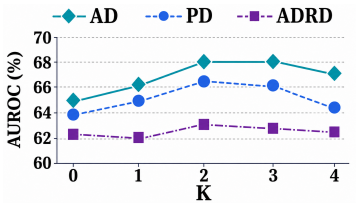


Figure 2: Robustness to the number of perturbed privileged views.

that only a small number of perturbed privileged views is needed to obtain stable hindsight signals. Further increasing K does not provide additional gains and may slightly reduce performance, likely because excessive perturbation produces less informative or overly diluted privileged views. These results support the lightweight multi-view design in OPHD, which improves robustness without substantially increasing computational cost.

Task-grounded feedback analysis. We further analyze whether the scorer-derived rationale weight a_i captures the predictive usefulness of rationales. We rank rationales by a_i and compare OPHD with two variants that remove the lowest and highest one-third groups, respectively. As shown in Figure 3, removing high-a_i updates leads to a larger AUROC drop than removing low-a_i updates, indicating that high-a_i rationales contribute more strongly to downstream discrimination. This supports the role of task-grounded scorer feedback in addressing the reasoning-prediction gap. Together with the ablation results in Table 2, this analysis shows that the scorer is not merely an auxiliary prediction head, but provides a task-level feedback signal for prioritizing rationales.

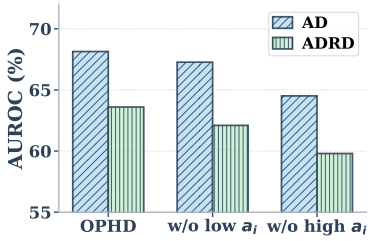


Figure 3: Task-grounded scorer feedback analysis.

Comparison with follow-up conditioned rationale imitation. We compare OPHD with a direct imitation strategy, where a privileged model generates rationales conditioned on both prediction-time EHR and follow-up EHR, and the prediction-time model is trained to imitate them. As shown in Figure 4, direct imitation performs worse than OPHD across all cohorts. Although it uses the same follow-up information during training, it creates a stronger train–test mismatch because teacher rationales may depend on evidence unavailable at inference. In contrast, OPHD keeps the prediction-time rationale fixed and uses follow-up EHR only to re-score the same reasoning trajectory. This makes hindsight supervision more consistent with the inference setting, where only prediction-time EHR is available.

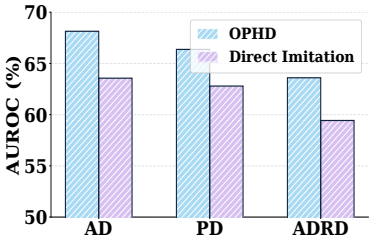


Figure 4: Comparison with follow-up-conditioned rationale imitation.

More discussion and experiments. Additional discussion and extended experimental analyses are provided in Appendix E.

5 Conclusion

We presented OPHD, an on-policy hindsight distillation framework for early risk prediction from EHR data. OPHD uses follow-up EHR as training-time privileged evidence while preserving the inference-time constraint that only prediction-time EHR is available. By re-scoring the same prediction-time rationale under perturbed privileged views and incorporating task-grounded scorer feedback, OPHD transfers stable hindsight signals into rationale learning without directly imitating future-conditioned explanations. Experiments across AD, PD, and ADRD cohorts show that OPHD consistently outperforms conventional EHR models and LLM-based reasoning baselines across both threshold-free metrics and clinically relevant high-specificity operating points. Ablation studies further confirm the contribution of follow-up EHR, perturbed privileged views, token-level hindsight weighting, and scorer-derived feedback. Additional analyses show that OPHD is more effective than direct follow-up-conditioned rationale imitation, remains robust across perturbation settings, and benefits most from rationales with higher task-grounded scorer feedback. These results support the value of using privileged hindsight supervision to improve prediction-time clinical reasoning while maintaining an inference-compatible information boundary. This study has two main limitations: evaluation is limited to neurodegenerative disease cohorts, and generated rationales should be viewed as model-produced intermediate representations rather than definitive clinical explanations or causal evidence. OPHD may support earlier identification of high-risk patients, but clinical use would require careful validation and human oversight.

Acknowledgments and Disclosure of Funding

The authors would like to acknowledge the support from NSF 2212175, NIH RF1AG072449, RF1AG084178, R01AG080991, R01AG080624, R01AG076448, R01AG076234, and R01NS140142.

References

[1] Ofir Ben Shoham and Nadav Rappoport. Cpllm: Clinical prediction with large language models. *PLOS Digital Health*, 2024.

[2] Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *ACL*, 2024.

[3] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[4] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *KDD*, 2017.

[5] Philip Chung, Christine T Fong, Andrew M Walters, Nima Aghaeepour, Meliha Yetisgen, and Vikas N O'Reilly-Shah. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA surgery*, 2024.

[6] Hejie Cui, Zhuocheng Shen, Jieyu Zhang, Hui Shao, Lianhui Qin, Joyce C Ho, and Carl Yang. Llms-based few-shot disease predictions using ehr: A novel approach combining predictive agent reasoning and critical agent instruction. In *AMIA Annual Symposium Proceedings*, 2025.

[7] Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 2016.

[8] Daniel Ferreira, Noémi Nogueira, Joana Guimarães, and Rui Araújo. Anti-dementia drugs: what is the evidence in advanced stages? *Porto Biomedical Journal*, 2024.

[9] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *WWW*, 2020.

[10] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 2018.

[11] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John PA Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198, 2016.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Emma Holler, Christina Ludema, Zina Ben Miled, Molly Rosenberg, Corey Kalbaugh, Malaz Boustani, and Sanjay Mohanty. Development and validation of a routine electronic health record-based delirium prediction model for surgical patients without dementia: Retrospective case-control study. *JMIR Perioperative Medicine*, 2025.

[14] Md Ekramul Hossain, Arif Khan, Mohammad Ali Moni, and Shahadat Uddin. Use of electronic health data for disease prediction: A comprehensive literature review. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[16] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 2012.

[17] Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*, 2023.

[18] Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*, 2024.

[19] Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *Nature communications*, 15(1):9074, 2024.

[20] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *EMNLP*, 2016.

[21] Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harmander Kaur. The use of electronic health records to support population health: a systematic review of the literature. *Journal of medical systems*, 2018.

[22] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, volume 2025, pages 20715–20735, 2025.

[23] Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *AAAI*, 2024.

[24] Qian Li, Xi Yang, Jie Xu, Yi Guo, Xing He, Hui Hu, Tianchen Lyu, David Marra, Amber Miller, Glenn Smith, et al. Early prediction of alzheimer's disease and related dementias using real-world electronic health records. *Alzheimer's & Dementia*, 2023.

[25] Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, 2025.

[26] 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(1):7155, 2020.

[27] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2023.

[28] Bin Liu, Ying Li, Zhaonan Sun, Soumya Ghosh, and Kenney Ng. Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In *AAAI*, 2018.

[29] Haohui Lu and Usman Naseem. Can large language models enhance predictions of disease progression? investigating through disease network link prediction. In *EMNLP*, 2024.

[30] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *KDD*, 2018.

[31] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*, 2020.

[32] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *ACL*, pages 1773–1781, 2023.

[33] Cynthia D Morris, Karen McCracken, Mary Samuels, and Eric Orwoll. Creating an institutional resource for research education and career development: A novel model from oregon clinical and translational research institute. *Clinical and Translational Science*, 2014.

[34] Parvati Naliyatthaliyazchayil, Raajitha Muthyala, Judy Wawira Gichoya, and Saptarshi Purkayastha. Evaluating the reasoning capabilities of large language models for medical coding and hospital readmission risk stratification: zero-shot prompting approach. *Journal of medical Internet research*, 2025.

[35] Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Carer-clinical reasoning-enhanced representation for temporal health risk prediction. In *EMNLP*, pages 10392–10407, 2024.

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[37] You Hyun Park, Yong Wook Kim, Dae Ryong Kang, Sang Chul Lee, and Seo Yeon Yoon. Prediction of all-cause mortality in parkinson's disease with explainable artificial intelligence using administrative healthcare data. *npj Parkinson's Disease*, 2025.

[38] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023.

[39] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[40] Andrea H Ramirez, Lina Sulieman, David J Schlueter, Alese Halvorson, Jun Qian, Francis Ratsimbazafy, Roxana Loperena, Kelsey Mayo, Melissa Basford, Nicole Deflaux, et al. The all of us research program: data quality, utility, and diversity. *Patterns*, 2022.

[41] Ines Reinecke, Michéle Zoch, Christian Reich, Martin Sedlmayr, and Franziska Bathelt. The usage of ohdsi omop–a scoping review. *German Medical Data Sciences 2021: Digital Medicine: Recognize-Understand-Heal*, 2021.

[42] Patrick Schober and Thomas R Vetter. Propensity score matching in observational research. *Anesthesia & Analgesia*, 2020.

[43] Anette Schrag, Zacharias Anastasiou, Gareth Ambler, Alastair Noyce, and Kate Walters. Predicting diagnosis of parkinson's disease: a risk algorithm based on primary care presentations. *Movement Disorders*, 2019.

[44] Anette Schrag, Jens Bohlken, Lotte Dammertz, Stefan Teipel, Wiebke Hermann, Manas K Akmatov, Jörg Bätzing, and Jakob Holstiege. Widening the spectrum of risk factors, comorbidities, and prodromal features of parkinson disease. *JAMA neurology*, 2023.

[45] 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.

[46] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[47] Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P Martin. Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 2021.

[48] Karthik Soman, Charlotte A Nelson, Gabriel Cerono, Samuel M Goldman, Sergio E Baranzini, and Ethan G Brown. Early detection of parkinson's disease through enriching the electronic health record using a biomedical knowledge graph. *Frontiers in Medicine*, 2023.

[49] Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *ICLR*, 2024.

[50] Alice S Tang, Katherine P Rankin, Gabriel Cerono, Silvia Miramontes, Hunter Mills, Jacquelyn Roger, Billy Zeng, Charlotte Nelson, Karthik Soman, Sarah Woldemariam, et al. Leveraging electronic health records and knowledge networks for alzheimer's disease prediction and sex-specific biological insights. *Nature Aging*, 2024.

[51] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[52] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

[54] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

[55] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *ICLR*, 2024.

[56] Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Dearllm: Enhancing personalized healthcare via large language models-deduced feature correlations. In *AAAI*, 2025.

[57] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

[58] Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. In *NeurIPS*, 2020.

[59] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

[60] Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, et al. Large language models for disease diagnosis: A scoping review. *Nature Partner Journals Artificial Intelligence*, 2025.

A More Dataset Details

Our evaluation leverages three retrospective case-control cohorts derived from two EHR resources: the nationwide *All of Us* (AoU) Research Program and the regional clinical repository (LocalEHR). Across all cohorts, outcomes are operationalized using diagnosis- and medication-based criteria as established in prior studies. Structured EHR trajectories as well as demographic information are extracted for the constructed cohorts. Unless otherwise specified, individuals are required to be at least 50 years old at index time to mitigate impacts of genetic factors, aligning the cohorts with late-life neurodegenerative disease prediction [50, 43, 24].

A.1 All of Us Alzheimer's Disease Cohort

As a nationwide U.S. research initiative, All of Us (AoU) collects large-scale, de-identified longitudinal EHR data from diversified participant populations [40], and harmonizes them under the OMOP Common Data Model (OMOP CDM) [41]. From this resource, we construct a case-control cohort for Alzheimer's disease (AD) prediction. AD cases are identified using established diagnostic codes (ICD-9 331.0 and ICD-10 G30), with the index time defined as the earliest record of either an AD diagnosis or an anti-dementia prescription (ATC code N06D) [24, 50, 8]. Controls are required to have no recorded AD-related diagnoses among the G[123]* ICD-10 categories and no exposure to anti-dementia medications [50]. The resulting cohort is randomly divided into 90% for model development and 10% held-out for evaluation. Cases and controls are further matched within the development split by propensity scores based on demographic and visit-related covariates to reduce potential confounding[42].

A.2 All of Us Parkinson's Disease Cohort

We similarly construct an AoU-based cohort for early prediction of Parkinson's disease (PD), with idiopathic PD as the target phenotype. To improve etiologic specificity, individuals with diagnoses or medications suggestive of non-idiopathic parkinsonism or potential PD mimics, including atypical or secondary parkinsonism, primary psychiatric disorders, and functional neurologic disorders are excluded [37, 43, 44]. PD cases are identified with at least one PD diagnosis, with anti-PD medication records further informing case index-time assignment [48]. Controls are defined by the absence of recorded PD diagnoses, PD-related medication exposure, and diagnosis- or medication-based evidence of major neurodegenerative or secondary parkinsonian conditions. Likewise, 10% of initial cases and controls are reserved for held-out evaluation, while the remaining development split is propensity-score matched on demographics and healthcare utilization for model development.

A.3 Alzheimer's Disease and Related Dementias Cohort

We exploit the local hospital's repository for Alzheimer's disease and related dementias (ADRD) prediction, including AD, vascular dementia (VaD), frontotemporal dementa (FD) and Lewy-body dementia (LeD), and mixed-typed conditions. For ADRD, case and control definitions adopt the same criteria as in prior work [24], supplemented by anti-dementia medication prescriptions to achieve index-time adjustment. Controls are required to have no recorded ADRD diagnoses, no conditions associated with or causal to dementia (e.g., mild cognitive impairment), and no exposure to anti-dementia medications. To accommodate dataset scale, 25% are randomly held out for evaluation with the others for development. As demographic variables other than age are largely unavailable in this dataset, matching within the development data is performed using age and comorbidity score.

For all three cohorts, EHR features are processed with medical code transformation and filtering procedures, e.g., from ICD codes to phecodes[24]. Demographic, condition, and medication data that are available within the observation window are extracted and utilized as model features. Summarily, the AD, PD, and ADRD cohorts included 605/12,156, 444/7,772, and 379/9,809 cases/controls, respectively. Mean ages of index time among cases and controls are calculated in all three cohorts (AD: 70.83 vs. 63.30 years; PD: 66.00 vs. 62.51 years; ADRD: 79.69 vs. 75.15 years).

Algorithm 1 On-Policy Hindsight Distillation (OPHD)

Require: Training cohort $\mathcal{D} = \{(\mathcal{X}_i, \Delta\mathcal{X}_i, Y_i)\}_{i=1}^{N}$; rationale generator π_θ; frozen re-scoring model F_θ; scorer g_ϕ; rationale prompt P_R; re-scoring prompt P_T; perturbation operator $\mathcal{T}$; number of views K.

Ensure: Updated rationale generator π_θ and scorer g_ϕ.

for each training iteration **do**
  Sample a mini-batch $\mathcal{B}$.
  for each patient $i \in \mathcal{B}$ **do**
    Generate prediction-time rationale $R_i \sim \pi_\theta(\cdot \mid \mathcal{X}_i, P_R)$.
    Re-score R_i under the prediction-time reference view using $F_\theta(r_{i,<t}, \mathcal{X}_i, P_T)$.
    for $k = 1$ to K **do**
      Sample perturbed follow-up view $\Delta\mathcal{X}_i^{(k)} \sim \mathcal{T}(\Delta\mathcal{X}_i)$.
      Re-score the same rationale under the privileged view using $F_\theta(r_{i,<t}, \mathcal{X}_i, \Delta\mathcal{X}_i^{(k)}, P_T)$.
    end for
    Compute robust hindsight weights $w_{i,t}$ from the magnitude and stability of follow-up-induced token shifts.
    Compute scorer-derived rationale feedback a_i from the label-support gain of R_i.
    Form feedback-guided token weights $\widetilde{w}_{i,t} = a_i\, w_{i,t}$.
  end for
  Compute multi-view token-level distillation loss $\mathcal{L}_{\mathrm{OPHD}}$ using $\widetilde{w}_{i,t}$.
  Compute supervised scorer loss $\mathcal{L}_{\mathrm{score}}$.
end for
return learned rationale generator π_θ and scorer g_ϕ.

B Early Prediction Setup

We follow the common case-control study setup for EHR-based early prediction, where prediction is generally operationalized by anchoring each patient to an index time and restricting model inputs to records observed before a predefined prediction time [24]. For case patients, the index time is defined as the first clinical indication of the disease onset. Analogous index time for controls is assigned one year before the last recorded visit [50]. Following prior EHR-based studies on neurodegenerative disease risk prediction, we use a five-year prediction horizon as a challenging early-risk setting. Model inputs are restricted to records observed before this prediction window, thereby defining an observation window that extends from the beginning of the available EHR history up to the prediction time. We also exclude the medical codes used for case/control eligibility or index-time assignment from the model inputs, ensuring that prediction reflects models' ability to use remaining EHR observations rather than cohort criteria proxies. To ensure minimal longitudinal context, at least one year of recorded history within the observation window is required for the participants.

C Baselines

We compare OPHD with conventional EHR representation learning methods, LLM-based reasoning baselines, and LLM-enhanced clinical prediction methods. All methods are evaluated under the same information constraint: only prediction-time EHR is available at test time and follow-up data are leveraged only during training when applicable.

- **GRAPE** [58] models feature relationships by constructing a graph comprising both patient nodes and feature nodes, with observed feature values represented as edge attributes. The learned patient embeddings are used for downstream disease risk prediction.
- **MUSE** [55] constructs a heterogeneous graph with patient nodes and modality-specific nodes. The adopted modalities contain diagnosis, medication, and demographics, while edges between patients and modality nodes are attributed with corresponding modality-specific EHR feature vectors. A Siamese GNN is trained by contrastive learning on original and augmented graphs to obtain modality-robust patient representations.
- **MOTOR** [49] is a foundation model for longitudinal EHR representation learning. It learns patient representations from temporal clinical event sequences and supports generic

downstream prediction tasks. Output representations from MOTOR are then utilized for outcome prediction.

- **LLM Zero-shot Inference** directly prompts an instruction-tuned language model to predict disease risk without task-specific examples or parameter updates. Historical EHR summaries are converted into task-specific prompts, and the model generates risk assessments based on instruction-following capability with Chain-of-Thought reasoning (CoT) applied.
- **LLM Few-shot Inference** augments the task prompt with a small number of labeled in-context examples with CoT applied. With three examples provided in the prompt, the LLM generates a risk assessment from prediction-time EHR through in-context reasoning, without updating model parameters.
- **Multi-sample** [52] samples multiple rationales from the LLM and aggregates their corresponding predictions. Inspired by self-consistency-style reasoning, multiple reasoning paths are used to reduce generation variance and improve prediction stability.
- **Best-of-K** samples multiple rationales generated with CoT from the LLM for each sample. The rationale with the highest scorer confidence is selected for final prediction. This baseline tests whether simple multi-sample selection can improve over a single generated rationale.
- **ReConcile** [2] adopts an LLM-based reasoning framework that generates multiple candidate clinical rationales and reconciles them through an aggregation step to produce a final prediction. We use two agents in the experiments.
- **DearLLM** [56] uses LLMs to construct a personalized feature-prior graph from temporal EHR history, where correlations between clinical features are estimated through prompt-based reasoning. A graph encoder learns feature-aware patient representations on this LLM-induced graph, which are combined with EHR embeddings for prediction.

D Implementation Details

We implement the proposed OPHD model in PyTorch. The rationale generator π_θ was initialized from Qwen3-4B-Instruct-2507 and adapted with LoRA rather than full fine-tuning, with AdamW optimizer. We perform hyperparameter search over the LoRA rank $r = \{4, 8\}$ and the number of adapted final layers $n = \{2, 4, 8, 10\}$. We utilized two perturbed privileged views in the main experiments, with 10% dropout applied to diagnosis and medication follow-up codes. Other hyperparameters include $\lambda = 1.0$ for hindsight weight calculation and top-k vocabulary truncation with $k = 50$ for JSD computation. Besides, the generator allows for a maximum prompt length of 2048 tokens, and a maximum generation length of 512 tokens.

We adopted a separate task-grounded scorer to estimate the predictive utility of generated rationales as well as for risk prediction, initialized from Qwen3-0.6B with a linear sequence-classification head. The scorer is trained with LoRA using AdamW optimizing cross-entropy classification loss. We select the scorer LoRA configuration by grid search over the rank $r = \{2, 4, 8\}$ and final layer number $n = \{2, 4, 8\}$ for LoRA to apply.

Each cohort is randomly split into training, validation, and testing sets, where the testing set is directly held-out from case and control candidates for each cohort to reflect the natural case prevalence. For a fair comparison, all baseline methods were evaluated using identical data splits with hyperparameters selected via grid search. All LLM-based baselines use the same LLM backbone as the rationale generator in our proposed model. We report averaged results and standard deviation over three runs.

E Additional Experiments

E.1 Effect of Gain-based Scorer Feedback

We evaluate whether scorer feedback should measure the incremental predictive contribution of the generated rationale. In OPHD, the rationale weight a_i is defined by comparing the scorer's label support with and without the rationale:

$$a_i = \sigma\left(\log p_\phi(Y_i \mid \mathcal{X}_i, R_i) - \log p_\phi(Y_i \mid \mathcal{X}_i)\right).$$

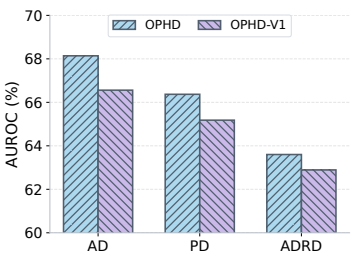


Figure 5: Effect of gain-based scorer feedback.

This design distinguishes rationales that add task-relevant information from patients that are already easy to classify using prediction-time EHR alone.

As shown in Figure 5, replacing the gain-based weight with the scorer probability $p_\phi(Y_i \mid \mathcal{X}_i, R_i)$ leads to worse performance. This suggests that a high scorer probability does not necessarily imply that the rationale itself provides additional predictive value; it may simply reflect that the original EHR is already informative. In contrast, gain-based feedback better isolates the rationale's incremental contribution beyond the original EHR input. These results support using gain-based scorer feedback to emphasize rationales that are not only clinically plausible, but also predictive for the target outcome.

E.2 Sensitivity to Follow-Up Window

We evaluate the sensitivity of OPHD to the temporal cutoff used for training-time follow-up evidence. Specifically, we compare two settings in which the follow-up EHR is truncated at 1 year or 3 years before the outcome/index time. The 1-year setting retains follow-up records closer to the outcome, while the 3-year setting uses only earlier follow-up records that are farther from the outcome.

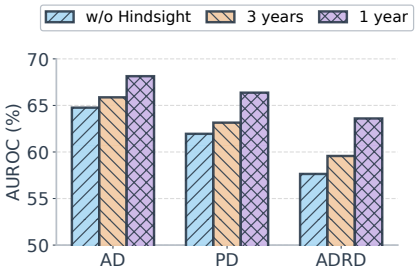


Figure 6: Sensitivity to follow-up window.

As shown in Figure 6, both follow-up settings improve over the no-hindsight baseline, confirming that training-time follow-up evidence provides useful privileged information. The 1-year setting achieves the best performance because records closer to the outcome contain stronger hindsight signals about how prediction-time clinical patterns evolve. The 3-year setting remains beneficial but yields smaller gains, since earlier follow-up records are less directly informative for identifying stable follow-up-induced preference shifts. These results suggest that OPHD can use follow-up evidence under different temporal cutoffs, while closer follow-up records provide more informative hindsight supervision.

F Limitations

OPHD uses generated rationales as intermediate representations for prediction, but these rationales should not be viewed as definitive clinical explanations, causal evidence, or standalone decision justifications. In addition, our evaluation is limited to selected neurodegenerative disease cohorts using structured EHR data. Future work should examine whether privileged-hindsight distillation generalizes to broader clinical prediction tasks, additional healthcare systems, and multimodal patient records such as clinical notes, imaging, and genomics.

G Broader Impacts

OPHD is intended to support research on early risk prediction from structured EHR data. By improving the use of prediction-time signals, the framework may help identify patients who could benefit from earlier monitoring, follow-up assessment, or preventive care planning. At the same time, OPHD should not be interpreted as an autonomous clinical decision system. Its predictions and generated rationales require clinical oversight before any real-world use. The goal of OPHD is to provide a methodological tool for studying training-time hindsight supervision, not to replace clinician judgment or determine patient care on its own.

H LLM Instructions

In this section, we present representative prompts used in our framework. We use Alzheimer's disease (AD) as an example to illustrate the instructions for rationale generation, privileged hindsight modeling, and score-based evaluation.

Rationale Generation Prompt
Alzheimer's Disease (AD)

You are a national-level experienced physician with expertise in Alzheimer's Disease (AD). You are interpreting EHR records to understand patient-level clinical patterns. You are evaluating the patient at the PREDICTION TIME (i.e., baseline time), which is exactly five years before the future outcome window. You have access to the patient's EHR history available up to this time. For training purposes only, you are also shown follow-up information from baseline time to a follow-up cutoff time (e.g., 1 year prior to the outcome window). This follow-up information is provided solely to help you understand how baseline patterns may develop and thus be interpreted well. Your output must be written strictly as if you are at the prediction time with access only to baseline information.

You will review the patient's EHR history at the baseline time: e.g., whether the history reflects cognitive, functional, behavioral, vascular, metabolic, neuropsychiatric, or other systemic patterns, whether such patterns appear patient-specific rather than generic, and whether these associate with AD risk.

Write one concise teaching-style clinical note as if you are explaining your reasoning to a medical student. Use clear, stepwise medical reasoning in full sentences, written as a single paragraph (do not use lists or headings). Explicitly name relevant diseases or clinical concepts, when appropriate, related to specific baseline evidence. Include (i) the most important baseline clinical patterns, (ii) plausible clinical interpretation and why it fits the baseline record.

After completing the reasoning, you must end the paragraph with one final sentence labeled exactly as: Baseline clinical interpretation: ...

Guidance:
- Do not center reasoning on the absence of documented cognitive/functional assessments.
- Focus on interpreting what is present in the baseline record and how salient each pattern is.
- Don not reply on other data modalities.
- Write from a retrospective descriptive perspective.

Content guidelines:
- Describe overarching clinical patterns or meanings inferred from the baseline record.
- Do not state or imply future risk, likelihood, concern, probability, or time horizon.
- Do not state whether the patient will or will not develop AD.
- Use explanatory language that interprets patterns rather than enumerating diagnoses or codes.

Patient-specific input:
Baseline demographics: Age (`{age}`), Sex (`{sex}`).
Baseline diagnoses: `{base_codes_diagnosis}`
Baseline medications: `{base_codes_medication}`

Write only the final clinical reasoning paragraph.

Privileged View Prompt
Alzheimer's Disease (AD)

You are a national-level experienced physician with expertise in Alzheimer's Disease (AD). You are interpreting EHR records to understand patient-level clinical patterns. You are evaluating the patient at the PREDICTION TIME (i.e., baseline time), which is exactly five years before the future outcome window. You have access to the patient's EHR history available up to this time. For training purposes only, you are also shown follow-up information from baseline time to a follow-up cutoff time (e.g., 1 year prior to the outcome window). This follow-up information is provided solely to help you understand how baseline patterns may develop and thus be interpreted well. Your output must be written strictly as if you are at the prediction time with access only to baseline information.

You will review the patient's EHR history at the baseline time: e.g., whether the history reflects cognitive, functional, behavioral, vascular, metabolic, neuropsychiatric, or other systemic patterns,

whether such patterns appear patient-specific rather than generic, and whether these associate with AD risk.

Write one concise teaching-style clinical note as if you are explaining your reasoning to a medical student. Use clear, stepwise medical reasoning in full sentences, written as a single paragraph (do not use lists or headings). Explicitly name relevant diseases or clinical concepts, when appropriate, related to specific baseline evidence. Include (i) the most important baseline clinical patterns, (ii) plausible clinical interpretation and why it fits the baseline record.

After completing the reasoning, you must end the paragraph with one final sentence labeled exactly as: Baseline clinical interpretation: ...

Guidance:
- Do not center reasoning on the absence of documented cognitive/functional assessments.
- Focus on interpreting what is present in the baseline record and how salient each pattern is.
- Don not reply on other data modalities.
- Write from a retrospective descriptive perspective.

Content guidelines:
- Describe overarching clinical patterns or meanings inferred from the baseline record.
- Do not state or imply future risk, likelihood, concern, probability, or time horizon.
- Do not state whether the patient will or will not develop AD.
- Use explanatory language that interprets patterns rather than enumerating diagnoses or codes.

Patient-specific input:
Training-only follow-up (newly emerged during the aforementioned follow-up period):
Follow-up diagnoses: `{follow_up_codes_diagnosis}`
Follow-up medications: `{follow_up_codes_medication}`
Baseline demographics: Age (`{age}`), Sex (`{sex}`).
Baseline diagnoses: `{base_codes_diagnosis}`
Baseline medications: `{base_codes_medication}`

Write only the final clinical reasoning paragraph.

Scorer Instruction
Alzheimer's Disease (AD)

You are an experienced physician with expertise in Alzheimer's disease (AD). You are conducting an early clinical prediction for AD risk using EHR data, with the goal of predicting disease development several years in advance rather than diagnosing current conditions. You are evaluating the patient at the PREDICTION TIME (i.e., baseline time), which is exactly five years before the future outcome window. You have access to the patient's EHR history available up to this time.

Patient-specific input:
Baseline demographics: Age (`{age}`), Sex (`{sex}`).
Baseline diagnoses: `{base_codes_diagnosis}`
Baseline medications: `{base_codes_medication}`

Clinical Interpretation: `{prediction-time rationale}` .

Output a single scalar risk score as the risk of developing into AD at the outcome window.

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 accurately summarizes the paper's contributions, designs, and experimental findings, and are consistent with the detailed content presented in the subsequent sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discussed limitations of the proposed model in the Conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [No]

Justification: The paper does not include theoretical results

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: We provide detailed descriptions for experimental settings in the main text and Appendix, and release the source code for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: The real-world data in the study is accessible for registered researchers. We have discussed dataset details in the main text and Appendix, and release the code at https://github.com/charles-code-for-submission/OPHD.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and the test details can be found in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: Experimental results with their standard deviation are provided in Table 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: We provide the used software environment and hardware in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: We affirm that this research fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: In Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve this.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: We have obtained authorization for the used data.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: No new asset is proposed.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: The paper obtained IRB based on the requirements of the institution.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper proposes a hindsight distillation framework with LLM as one of the coordinated modules.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.