CODA: Cohort- and Drift-aware Foundation Model for Multimodal Clinical Reasoning

Changshuo Liu
School of Computing
National University of Singapore

Jiaqi Zhu*
School of Computing
National University of Singapore

Wenqiao Zhang
School of Software Technology
Zhejiang University

Xiaokui Xiao
School of Computing
National University of Singapore

Beng Chin Ooi
School of Software Technology
Zhejiang University

Abstract

Recent advances in multimodal foundation models have demonstrated strong performance across diverse clinical tasks. However, existing approaches predominantly operate at the individual sample level and fail to model the interplay between population-level cohort patterns and patient-specific variation that underlie real-world clinical reasoning. In practice, clinicians routinely reason through patient cohorts sharing similar demographic, physiological, or pathological characteristics, while accounting for patient-specific deviations from the group prototype. Motivated by this gap, we propose CODA, a cohort- and drift-aware foundation model for multimodal clinical reasoning. CODA explicitly integrates cohort structure into the modeling pipeline through two complementary mechanisms: cohort-aware population stratification and drift-aware individual adaptation. In particular, we leverage reinforcement learning to discover clinically meaningful patient cohorts, and further model each patient's deviation from cohort prototypes as a structured drift signal encoding fine-grained individual heterogeneity. These cohort- and drift-aware representations are incorporated into the foundation model to enable reasoning that is simultaneously informed by population-level patterns and sensitive to individual-level variation. Extensive experiments on real-world electronic health record (EHR) benchmarks demonstrate that CODA achieves consistent effectiveness and strong generalizability across diverse clinical tasks, including closed-QA, open-QA, and report generation.

1 Introduction

Generative AI has rapidly evolved from text-only large language models (LLMs) [1, 2, 3] to large vision–language models (LVLMs) [4, 5, 6, 7], enabling unified reasoning over both textual and visual modalities. This paradigm shift is particularly consequential in the medical domain, where clinical decision-making inherently relies on heterogeneous data sources, including medical images, EHR data, and free-text reports [8, 9, 10]. By integrating these modalities within a single framework, recent

*Corresponding author: Jiaqi Zhu (jiaqi77@nus.edu.sg)

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

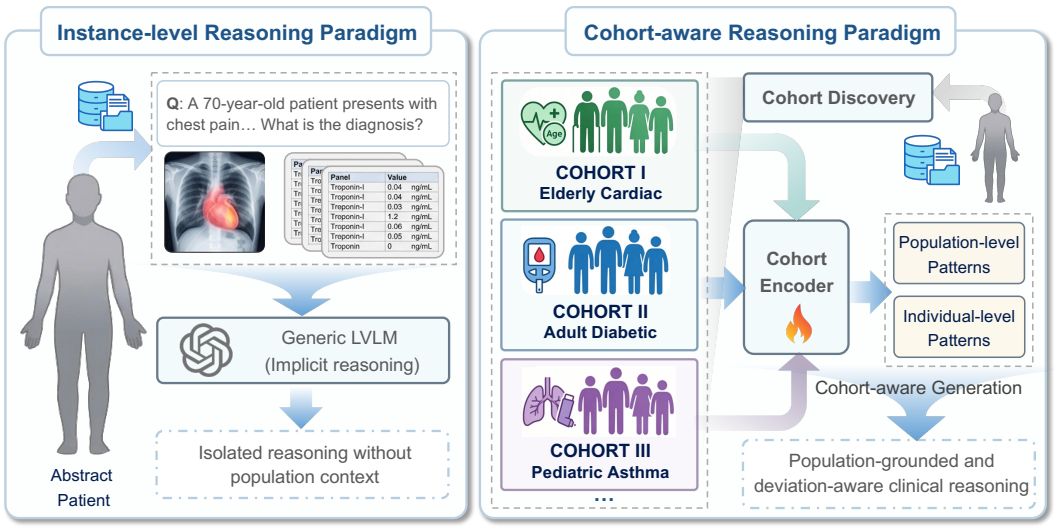


Figure 1: Rethinking Clinical Reasoning: From instance-level predictions to cohort-aware modeling.

approaches have demonstrated strong capabilities in medical question answering [11], multimodal understanding [12], and medical decision support [13]. These advances have driven notable progress in applications such as medical knowledge acquisition and education, treatment planning, and patient outcome prediction [14, 15, 16].

Despite these advances, existing medical foundation models [17, 18, 19] are developed in a task-centric manner, where reasoning is performed on individual queries with limited consideration of broader clinical context. However, clinicians rarely reason about a patient in isolation. Instead, they implicitly compare the current case with previously encountered patients, drawing on experience with recurring patterns across similar populations [20]. Without this kind of contextual grounding, medical foundation models struggle to capture meaningful clinical variation, often leading to fragmented or overly simplified reasoning.

A natural way to incorporate such context is through patient cohorts [21, 22]. By grouping patients with similar clinical profiles, such as demographics or specific health conditions, cohorts help guide how individual cases are interpreted and managed. Figure 1 illustrates the transition from conventional instance-level reasoning to cohort-aware clinical reasoning. Take the age cohorts as an example, a sudden episode of chest pain may be treated as a potential medical emergency for elderly patients with a history of heart disease, often requiring immediate hospital evaluation, whereas the same symptoms for young and healthy individuals may be attributed to benign causes such as muscle strain or stress. This type of cohort-based reasoning allows clinical decisions to be grounded in population-level patterns.

Furthermore, relying on cohort structure alone is insufficient to support accurate clinical decisions. Even within the same cohort, patients may differ in ways that lead to different clinical judgments. For example, two elderly patients with similar cardiovascular risk factors may both have chest pain, yet one may show stable vital signs and mild symptoms that allow for conservative monitoring, while the other may exhibit severe pain that demands immediate intervention. Such differences in symptoms and clinical signals often determine the final intervention, yet are not well captured by coarse cohort patterns. As a result, effective clinical decision-making requires jointly modeling cohort structure and individual heterogeneity, a capability that remains underexplored in existing approaches.

To this end, we propose CODA[2], a **CO**hort- and **D**rift-**A**ware foundation model that explicitly characterizes both cohort structure and individual heterogeneity. Our framework consists of two key components. First, we formulate cohort construction as a reinforcement learning process that automatically partitions patients into clinically meaningful groups. This process builds a cohort tree, where each internal node corresponds to a feature-based status and action, and each leaf represents a cohort. A cohort encoder is then employed to encode these cohorts, and the resulting embeddings are injected into the model input to provide cohort-level context. Second, we move beyond static

[2]Please refer to Appendix B for a discussion of related work.

cohorts and introduce a continuously evolving cohort prototype space, where cohort representations are dynamically updated. For each patient, we further compute a cohort-drift signal that captures how the individual deviates from the corresponding cohorts. This signal reflects patient-specific characteristics that are not explained by shared cohort patterns, and is incorporated into the model to refine its reasoning at the individual level. These cohort-aware and drift-aware representations enable more precise and context-aware clinical reasoning. We summarize our main contributions as follows:

- We propose a cohort- and drift-aware foundation model for multimodal clinical reasoning that jointly incorporates population-level cohort structure and patient-specific heterogeneity.
- We present a reinforcement learning-based cohort discovery and dynamic prototype updates strategy, enabling adaptive and data-driven cohort modeling.
- We explicitly model the individual deviations from cohort prototypes, and incorporate such signals into the backbone through drift-conditioned adapters.
- Extensive experiments demonstrate that CODA consistently outperforms both general-purpose and medical-specific baselines across closed-QA, open-QA, and report generation tasks on diverse real-world EHR datasets, highlighting its effectiveness in enhancing clinical reasoning.

2 Preliminary

2.1 Multimodal Clinical Data Representation in Foundation Models

Modern foundation models encode heterogeneous clinical data by projecting multimodal inputs into a unified token embedding space. Given the input $\mathbf{x} = \{\mathcal{I}, \mathcal{X}\}$ and ground truth output sequence $\mathbf{y}$, where $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ represents a medical image and $\mathcal{X} = \{x_1, x_2, \ldots, x_n, \mathbf{q}\}$ denotes n serialized clinical records and the question $\mathbf{q}$, existing LVLMs first encode the visual and textual inputs separately, and then fuse them into a shared embedding space as

$$\mathbf{Z}^{(0)} = [f_{I\to t}(\mathcal{E}_I(\mathcal{I}))\,;\, \mathcal{E}_t(\mathcal{X})] \in \mathbb{R}^{(L_I+L_t)\times d_t} \tag{1}$$

where $\mathcal{E}_I(\cdot)$ is a vision encoder and $\mathcal{E}_t(\cdot)$ is the textual encoder. L_I and L_t denote the numbers of visual and textual tokens, respectively. $f_{I\to t}$ maps the visual embeddings into the shared embedding space, and d_t denotes the dimensions of the shared embedding space.

The fused embedding is then processed by L_b Transformer blocks $\mathcal{F}$, $\mathbf{Z}^{(\ell)} = \mathcal{F}^{(\ell)}\big(\mathbf{Z}^{(\ell-1)}\big)$, $l = 1, \ldots, L_b$ yielding the final contextualized representation $\mathbf{Z}^{(L_b)}$. Conditioned on $\mathbf{Z}^{(L_b)}$, the model generates the output sequence autoregressively with learnable parameters θ as

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} p_\theta\Big(y_i \mid y_{<i}, \mathbf{Z}^{(L_b)}\Big) \tag{2}$$

2.2 Reinforcement Learning

Reinforcement learning (RL) is a general framework designed for sequential decision-making, typically formulated as a Markov Decision Process (MDP) defined as

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma), \quad s \in \mathcal{S},\ a \in \mathcal{A} \tag{3}$$

where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, P denotes the transition dynamics. r is the reward function, and $\gamma \in [0, 1]$ is the discount factor. Given a state s, a policy $\pi_\phi(a \mid s)$ specifies a distribution over the action space. The policy induces a trajectory τ with probability $p(\tau)$:

$$\tau = (s_1, a_1, s_2, a_2, \ldots, s_T), \quad p(\tau) = p(s_1) \prod_{k=1}^{T} \pi_\phi(a_k \mid s_k)\, P(s_{k+1} \mid s_k, a_k) \tag{4}$$

The objective of RL is to learn a policy that maximizes the expected cumulative reward as

$$\max_{\pi_\phi}\ \mathbb{E}_{\tau\sim p_\tau(\cdot|\pi_\phi)}\left[\sum_{k=1}^{T} \gamma^{k-1} r(s_k, a_k)\right], \tag{5}$$

where $p_\tau(\cdot \mid \pi_\phi)$ denotes the trajectory distribution induced by policy π_ϕ.

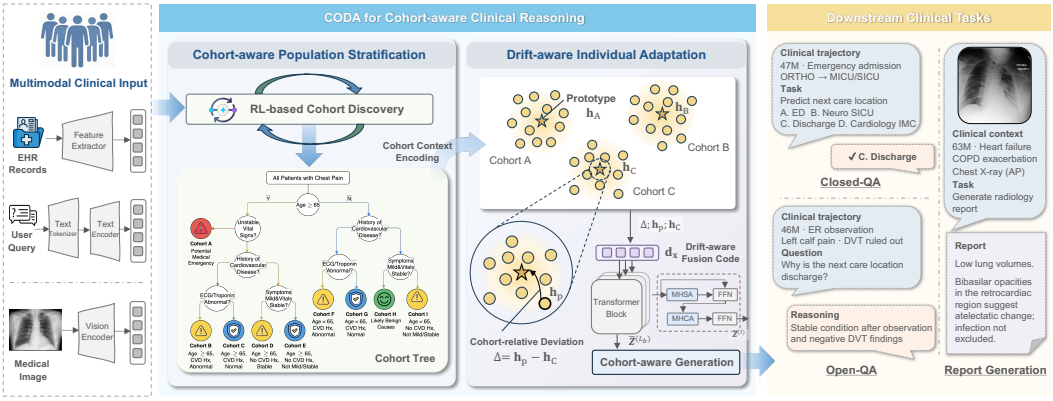


Figure 2: Overview of CODA for cohort- and drift-aware multimodal clinical reasoning.

3 Methodology

3.1 CODA Overview

As illustrated in Figure 2, CODA adopts a two-stage formulation to enable cohort-aware clinical reasoning. In the first stage, we learn a hierarchical cohort tree $\mathcal{T}$ that explicitly partitions the patient population $\mathcal{P}$ into a clinically meaningful cohort set $\mathcal{C}$. In the cohort tree, each internal node parameterizes a feature-based state s and partition action a, while each leaf node represents a cohort $c \in \mathcal{C}$. These partition actions are optimized with reinforcement learning, allowing the cohort structure to be shaped not only by clinical heterogeneity but also by its utility for downstream generation. With the learned cohorts in $\mathcal{C}$, we derive a cohort-aware representation for each sample and inject it into the backbone, thereby providing explicit population-level context rather than relying on the model to infer such structure implicitly from its parameters alone. In the second stage, building upon this cohort-level grounding, we maintain dynamically updated cohort prototypes $\mathbf{h}_c$ during training and characterize each sample by its cohort-relative deviation $\Delta(\mathbf{x}, c)$. This cohort-relative deviation captures fine-grained patient-specific variability beyond shared cohort characteristics, enabling the model to reason jointly over stable population patterns and individualized clinical departures. The cohort- and drift-aware framework is encapsulated in the following integrated objective:

$$\mathcal{J}(\theta, \pi_\phi) = \sum_{\mathbf{x},\mathbf{y}} \Big[\underbrace{\mathbb{E}_{\tau \sim p_\tau(\cdot|\pi_\phi)}[\mathcal{R}(\mathbf{x}, \tau)] + \mathcal{H}(p_{\pi_\phi}(\cdot|\mathbf{x}))}_{\text{Cohort-aware Population Stratification}} + \underbrace{\log p_\theta\big(\mathbf{y} \mid \mathbf{x}, \Delta(\mathbf{x}; c)\big)}_{\text{Drift-aware Individual Adaptation}} \Big] \tag{6}$$

where θ denotes the learnable model parameters and π_ϕ is the cohort-partitioning policy, $\mathcal{R}(\mathbf{x}, \tau)$ is the trajectory-level cohort discovery reward, $\mathcal{H}(p_{\pi_\phi}(\cdot \mid \mathbf{x}))$ is the policy entropy for cohort partition exploration.

3.2 Cohort-aware Population Stratification

To formalize the context-dependency inherent in clinical heuristics, CODA eschews the conventional assumption of a homogeneous patient distribution. Instead, CODA introduces a hierarchical stratification process that partitions the patient population space $\mathcal{P}$ into a set of clinically distinct cohorts $\mathcal{C}$. This process is governed by two synergistic components, i.e., an RL-based cohort discovery mechanism and a cohort context injection schema that injects cohort-level context into the model.

RL-based Cohort Discovery. We formulate cohort discovery as a hierarchical policy learning problem, where a binary-tree-structured policy π_ϕ recursively partitions the patient population $\mathcal{P}$ into a set of cohorts $\mathcal{C}$. Formally, we define a *cohort tree* as a hierarchical binary tree $\mathcal{T} = (V, E)$, where $V = \{v_1, \cdots, v_N\}$ is the set of nodes and $E \subseteq V \times V$ is the set of directed edges representing parent–child relationships. Each node $v \in V$ corresponds to a subset of patients $\mathcal{P}_v \subseteq \mathcal{P}$, with the root node representing the entire population. Internal nodes correspond to decision points that further partition the patient subset, while leaf nodes represent terminal cohorts corresponding to clinically homogeneous patient groups. Each node v is further associated with (i) a state representation s_v that summarizes the statistical and structural properties of the corresponding patient subset, and (ii) a partition action a_v induced by the policy, which determines how the node is expanded or terminated.

The cohort tree thereby induces a recursive partition of the population space, where each root-to-leaf path encodes a sequence of feature-based decisions that uniquely determines a cohort.

Specifically, for each sample $\mathbf{x}$, we extract a clinical feature $\mathbf{u} = g(\mathbf{x})$, where $g(\cdot)$ maps serialized EHR records into a vector of statistical features capturing patient demographics, clinical conditions, and laboratory signals. Each node v is represented by a compact state vector:

$$s_v = [\mu(\mathbf{U}_{\mathcal{P}_v}); \sigma(\mathbf{U}_{\mathcal{P}_v}); |\mathcal{P}_v|; d_v], \tag{7}$$

where $\mathbf{U}_{\mathcal{P}_v} = \{\mathbf{u}_i\}_{i \in \mathcal{P}_v}$ denotes the collection of feature vectors of patients in node v. $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, respectively. $|\mathcal{P}_v|$ is the number of patients and d_v is the depth of node v.

At each node v, the cohort policy π_ϕ induces a stochastic action a_v over the patient subset $\mathcal{P}_v$ to specify whether to stop, the selected medical feature, and the corresponding split threshold, where action a_v is defined as:

$$a_v = (z_v, j_v, \mathrm{t}_v) \sim \pi_\phi(\cdot \mid s_v). \tag{8}$$

where $z_v \in \{0, 1\}$ is a binary stopping variable, with $z_v = 1$ indicating that v becomes a terminal cohort. j_v is the selected feature, and t_v is the split threshold. When $z_v = 0$, the node is expanded by partitioning $\mathcal{P}_v$ into two disjoint subsets:

$$\mathcal{P}_{v_L} = \{i \in \mathcal{P}_v \mid u_{i,j_v} \leq \mathrm{t}_v\}, \quad \mathcal{P}_{v_R} = \mathcal{P}_v \setminus \mathcal{P}_{v_L}, \tag{9}$$

where v_L and v_R denote the left and right child nodes, respectively.

The reward function is designed to encourage each split to produce more compact and balanced child cohorts while penalizing unnecessary tree depth:

$$r_v = \left(\sigma(\mathbf{U}_{\mathcal{P}_v}) - \sum_{k \in \{L,R\}} \frac{|\mathcal{P}_{v_k}|}{|\mathcal{P}_v|} \sigma(\mathbf{U}_{\mathcal{P}_{v_k}})\right) + \left(1 - \frac{\big||\mathcal{P}_{v_L}| - |\mathcal{P}_{v_R}|\big|}{|\mathcal{P}_v|}\right) - \frac{d_v}{d_{\max}}, \tag{10}$$

where $d_{\max}$ is a normalization constant denoting the maximum depth. Here, the first term measures the reduction in intra-cohort variance after splitting, the second term discourages highly imbalanced partitions, and the last term penalizes overly deep trees.

Cohort Context Encoding. For each cohort c, we derive a representative embedding via a similarity-aware aggregation over its members, which is then mapped to a cohort embedding $\mathbf{e}_c$ and injected into the multimodal input sequence for the patients $\mathcal{P}_c$ within cohort c to condition generation:

$$\mathbf{Z}^{(0)} = [f_{I \to t}(\mathcal{E}_I(\mathcal{I})); \mathbf{e}_c; \mathcal{E}_t(\mathcal{X})], \quad \mathbf{e}_c = f_c\left(\sum_{i \in \mathcal{P}_c} \frac{\exp(\cos(\mathbf{h}_i, \bar{\mathbf{h}}_c))}{\sum_{j \in \mathcal{P}_c} \exp(\cos(\mathbf{h}_j, \bar{\mathbf{h}}_c))} \mathbf{h}_i\right), \tag{11}$$

where f_c is a learnable mapping function, $\mathbf{h}_i = f_{\mathbf{u} \to \mathbf{h}}(\mathbf{u}_i)$ is the learned patient representation based on statistical feature $\mathbf{u}_i$ via function $f_{\mathbf{u} \to \mathbf{h}}$, and $\bar{\mathbf{h}}_c$ is the centroid of cohort c in the feature space.

To jointly optimize cohort discovery and cohort-aware generation, we define a unified objective that integrates policy learning and downstream generation:

$$\mathcal{L}_{\text{cohort}} = \mathbb{E}_{\tau \sim p_\tau(\cdot|\pi_\phi)} \left[\sum_{v \in V} \left(-\log \pi_\phi(a_v \mid \mathbf{s}_v)\, R_v - \mathcal{H}(\pi_\phi(\cdot \mid \mathbf{s}_v))\right)\right] + \lambda_g \mathcal{L}_{\text{gen}}, \tag{12}$$

where $R_v = r_v - \mathcal{L}_{\text{gen}}$ denotes the generation-aware reward and $\mathcal{H}(\pi_\phi(\cdot))$ is the entropy of the policy that encourages exploration. $\mathcal{L}_{\text{gen}}$ is the downstream generation loss with λ_g balancing policy learning and generation performance.

In this way, a cohort tree $\mathcal{T}$ and cohort embeddings $\mathbf{e}_c$ are jointly learned. The resulting structure captures population-level heterogeneity via hierarchical partitioning, while the learned embeddings encode cohort-specific context for clinical reasoning. Through the proposed generation-aware objective, cohort discovery is directly aligned with multimodal reasoning, enabling the model to move beyond purely structure-driven partitioning toward reasoning-aware population stratification.

3.3 Drift-aware Individual Adaptation

While the cohort tree $\mathcal{T}$ provides population-level structure, clinical reasoning further requires capturing patient-specific deviations from cohort prototypes. To this end, we introduce a drift-aware individual adaptation mechanism that explicitly models the cohort-relative deviation of each sample and integrates it into the generation process.

Cohort Prototype Representation. For each cohort c, we maintain a learnable prototype representation $\mathbf{h}_c$, which summarizes the central clinical characteristics of the cohort and is initialized using the previously learned cohort embedding $\mathbf{e}_c$. During training, these prototypes are dynamically updated in an online manner based on the incoming samples. Specifically, for a sample $\mathbf{x}$ of patient p assigned to cohort c with feature representation $\mathbf{h}_p$, we update the corresponding prototype using an exponential moving average mechanism:

$$\mathbf{h}_c \leftarrow (1 - \beta\, w(p,c))\mathbf{h}_c + \beta\, w(p,c)\, \frac{\mathbf{h}_p}{\|\mathbf{h}_p\|_2}, \tag{13}$$

where $w(p,c) \in [0,1]$ denotes the assignment confidence of patient p to cohort c, and the feature is ℓ_2-normalized to stabilize prototype updates. $\mathbf{h}_c$ replaces $\mathbf{e}_c$ in $\mathbf{Z}^{(0)}$ and serves as a conditioning signal in the multimodal input sequence, analogous to $\mathbf{e}_c$.

Drift-aware Deviation Modeling. While the cohort tree $\mathcal{T}$ captures population-level structure, clinical reasoning further requires modeling patient-specific deviations from cohort prototypes. To this end, we introduce a unified drift-aware deviation mechanism that explicitly characterizes individual departures from cohort-level patterns and injects them into the generation process.

Given a sample $\mathbf{x}$ with feature representation $\mathbf{h}_p$ and assigned cohort c, we compute its cohort-relative deviation as $\Delta(\mathbf{x}, c) = \mathbf{h}_p - \mathbf{h}_c$, which captures patient-specific variation beyond the cohort prototype. We first transform this deviation into a compact drift-aware fusion code:

$$\mathbf{d}_{\mathbf{x}} = \mathrm{LN}(\mathbf{W}_d\,[\Delta(\mathbf{x},c); \mathbf{h}_p; \mathbf{h}_c; \Delta(\mathbf{x},c) \odot \mathbf{h}_c] + \mathbf{b}_d)\,, \tag{14}$$

where $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise interaction, $\mathrm{LN}(\cdot)$ denotes layer normalization, $\mathbf{W}_d$ is a learnable weight matrix and $\mathbf{b}_d$ is a learnable bias.

The drift code is then used to modulate the hidden representation at transformer block ℓ through a residual adapter. Compared to a simple additive bias, which applies a uniform shift, our formulation introduces a multiplicative interaction between the deviation and the hidden representation. This enables context-dependent, feature-wise modulation, allowing the model to adapt differently across features and tokens.

$$\tilde{\mathbf{Z}}^{(\ell)} = \mathbf{Z}^{(\ell)} + \mathbf{W}_\ell^o \left((1 + \tanh(\mathbf{W}_\ell^\gamma \mathbf{d}_{\mathbf{x}})) \odot \mathrm{GELU}(\mathbf{W}_\ell^\alpha \mathbf{Z}^{(\ell)}) + \mathbf{W}_\ell^\beta \mathbf{d}_{\mathbf{x}} \right), \tag{15}$$

where $\mathbf{W}_\ell^o$ is a learnable output projection that maps the adapted signal back to the hidden space. The term $\mathrm{GELU}(\mathbf{W}_\ell^\alpha \mathbf{Z}^{(\ell)})$ corresponds to a nonlinear transformation of the transformer output, capturing content-dependent features. The modulation term $1 + \tanh(\mathbf{W}_\ell^\gamma \mathbf{d}_{\mathbf{x}})$ acts as a feature-wise scaling factor conditioned on the cohort-relative deviation, where $\tanh(\cdot)$ ensures bounded and stable modulation. Finally, $\mathbf{W}_\ell^\beta \mathbf{d}_{\mathbf{x}}$ introduces an additive bias term that shifts the representation based on the deviation signal.

Cohort-aware Generation. The adapted hidden representation $\tilde{\mathbf{Z}}^{(L_b)}$ at the final transformer block L_b is then used for autoregressive generation:

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} p_\theta\left(y_i \mid y_{<i}, \tilde{\mathbf{Z}}^{(L_b)}\right). \tag{16}$$

To facilitate individual deviation modeling and enforce inter-cohort separability, we introduce a drift loss that explicitly encodes both intra-cohort consistency and inter-cohort discrimination by minimizing the deviation from the assigned cohort prototype while imposing a margin-based constraint against alternative cohorts.

$$\mathcal{L}_{\mathrm{drift}} = \mathcal{L}_{\mathrm{gen}} + \lambda_d(\|\Delta(\mathbf{x},c)\|_2^2 + \max(0, m - \frac{\mathbf{h}_p^\top \mathbf{h}_c}{\|\mathbf{h}_p\|\|\mathbf{h}_c\|} + \frac{\mathbf{h}_p^\top \mathbf{h}_{c'}}{\|\mathbf{h}_p\|\|\mathbf{h}_{c'}\|})) \tag{17}$$

where $m > 0$ is the cohort margin, c' denotes the cohort with the second highest confidence to p, and λ_d controls the trade-off between generation performance and cohort-relative deviation modeling.

Unlike standard generation, where the model relies solely on input features, our formulation conditions the generation process on both the cohort prototype $\mathbf{h}_c$ and the cohort-relative deviation $\Delta(\mathbf{x}, c)$. This allows the model to first anchor predictions to population-level patterns, and then refine them based on patient-specific variations. Overall, this design enables a unified integration of cohort-level regularities and individual-level deviations, leading to context-aware and fine-grained generation.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct extensive experiments on four real-world EHR benchmarks, i.e., EHR-Bench [23], DIYHealth-Bench [5], MIMIC-CXR [24], and PadChest [25]. Specifically, the admissions, services, ingredient events (IE), and provider order entry (POE) tasks from EHR-Bench, together with MIMIC-CXR, are treated as in-distribution (ID) datasets, while the Transfers task from EHR-Bench, Obesity from DIYHealth-Bench, and PadChest are used as out-of-distribution (OOD) datasets. The model is trained on the ID training datasets and evaluated on both ID and OOD settings. Through these datasets, we aim to provide a comprehensive assessment of the effectiveness and generalizability of CODA in answering practical medical questions and generating clinical reports.

Baselines. We evaluate CODA against a broad set of baselines, including state-of-the-art generalist models, (e.g., LLaVA 1.5 [26], Llama 3 [27], InternVL 3 [28], Qwen3 VL [29], Gemma 3 [30], Grok [31], Claude 4.5 Haiku [32] and GPT 5.4 Mini [33]), as well as medical-specific models (e.g., LLaVA-Med v1.5 [11], Med-Flamingo [17], HuatuoGPT-Vision [18], HealthGPT [12], Lingshu [34], and Hulu-Med [35]. All models are evaluated under closed-form multiple-choice question answering (Closed-QA), open-ended question reasoning (Open-QA), and radiology report generation.

Training Strategy. Our training pipeline consists of three stages. (i) *Cross-Modal Alignment.* To adapt the visual and textual alignment to the medical domain, we train the vision projector of the backbone using a mixture of medical VQA and QA datasets, including MedQuAD [36], PubMedVision [37], SLAKE [38], and VQA-RAD [39]. (ii) *Cohort-aware Population Stratification.* With the backbone frozen, we train only the RL-related modules on a subset (15%) of the in-distribution data to learn a hierarchical cohort structure. (iii) *Drift-aware Individual Adaptation.* We fine-tune only the LoRA modules together with deviation-aware components to model patient-specific variations conditioned on cohort context.

Evaluation Metrics. For Closed-QA, we adopt Accuracy (ACC) and Matthews Correlation Coefficient (MCC) [40]. ACC quantifies the proportion of correctly predicted answers, reflecting exact match performance. MCC provides a balanced evaluation by capturing the agreement between predicted and true labels across all classes. For Open-QA, we use BLEU-1 [41] and F1-BioBERT (F1-Bio) [42]. BLEU-1 evaluates lexical overlap based on unigram precision, serving as a signal of surface-level correctness. F1-Bio measures semantic consistency by computing similarity in a contextual embedding space derived from BioBERT [43], enabling robust evaluation beyond surface-level matching. For report generation, we employ F1-RadGraph (F1-Rad) [44] and F1-Bio. F1-Rad assesses the correctness of clinical entities and their relations by comparing structured representations extracted from RadGraph [44], capturing clinical fidelity at a fine-grained level. Please refer to Appendix C for more details about the datasets, model implementation and evaluation metrics.

4.2 Main Results

The experimental results for Closed-QA, Open-QA, and report generation are presented in Tables 1 and 2. We consider three variants: (i) CODA with a Qwen3 backbone, (ii) CODA with an LLaMA3 backbone, and (iii) CODA with a Qwen3 backbone but without LoRA. We summarize the key findings as follows. **(i) State-of-the-art performance.** Three variants of CODA consistently outperform all baselines across all tasks, yielding average gains of 42.18% in ACC for Closed-QA, 6.01% in F1-Bio for Open-QA, and 63.82% in F1-Rad for report generation. These results demonstrate the effectiveness of CODA and suggest that explicitly modeling cohort structure and cohort drift provides a principled mechanism for capturing both population-level patterns and individual-level variations, resulting in consistently superior performance. **(ii) Out-of-Distribution Generalization.** Despite being trained solely on ID data, CODA generalizes well to OOD settings, as evidenced by

Table 1: Comparison of CODA with baselines under *Closed-QA* settings.

Domain	Model	ID								OOD				Avg.	
		Admissions		IE		POE		Services		Transfers		Obesity			
		ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
General Domain	LLaVA 1.5-7B	40.00	20.06	19.44	-7.48	57.00	42.82	63.59	51.45	60.60	47.61	61.90	47.97	50.42	33.74
	LLaMA 3-11B	32.80	10.74	26.52	4.85	35.40	14.55	43.64	25.32	37.00	16.29	54.12	39.58	38.25	18.56
	InternVL 3-8B	28.20	5.06	18.94	-6.52	46.80	29.13	42.39	23.35	27.00	2.58	70.92	59.61	39.04	18.87
	Qwen 3-VL-8B	36.40	15.35	38.38	17.64	40.40	20.49	60.60	48.07	53.80	38.63	69.67	58.10	49.88	33.05
	Gemma 3-12B	18.80	-8.06	19.95	-6.83	35.20	13.55	58.85	45.20	50.00	33.42	71.07	60.24	42.31	22.92
	Grok 4.2	23.60	-1.58	31.57	8.56	35.20	13.43	25.19	0.17	40.60	20.45	56.92	39.69	35.51	13.45
	Claude 4.5 Haiku	35.60	14.09	29.29	6.51	40.40	20.78	42.64	23.54	42.60	23.51	47.90	32.61	39.74	20.17
	GPT 5.4 Mini	34.60	12.60	42.39	23.08	44.40	26.02	43.14	24.17	48.40	31.14	59.72	45.92	45.44	27.15
Medical Domain	LLaVA-Med v1.5-7B	21.00	3.83	25.00	2.21	58.00	45.01	44.89	27.44	34.40	17.29	58.48	45.85	40.29	23.60
	Med-Flamingo-7B	17.40	-3.86	8.08	-3.11	21.80	3.18	25.69	4.28	27.80	1.84	17.42	3.60	19.70	0.99
	HuatuoGPT-Vision-7B	44.00	26.00	40.66	20.78	48.40	32.11	60.10	46.94	57.80	44.60	67.96	56.04	53.15	37.74
	HealthGPT-14B	39.60	19.53	25.51	1.53	38.20	18.04	33.42	12.10	42.00	23.55	55.68	42.71	39.07	19.58
	Lingshu-7B	44.40	26.26	45.45	28.06	50.60	34.10	78.55	71.56	57.00	44.26	70.63	60.52	57.77	44.13
	Hulu-Med-7B	43.60	25.12	26.01	2.39	53.20	37.75	60.60	47.84	36.00	14.56	65.47	53.56	47.48	30.20
	CODA-Qwen3-8B	**76.60**	**68.77**	92.42	89.93	**82.80**	**77.05**	**93.52**	**91.39**	**75.20**	**66.89**	72.32	61.28	**82.14**	**75.88**
	CODA-Llama3-8B	**76.60**	**68.77**	**92.93**	**90.59**	82.00	76.00	92.27	89.71	74.80	66.31	**72.90**	**61.71**	81.92	75.52
	CODA-Qwen3-8B w/o LoRA	72.80	63.75	92.17	89.59	82.40	76.53	85.29	80.55	70.20	60.61	72.16	60.98	79.17	72.00

Table 2: Comparison of CODA with baselines under *Open-QA* and *report generation* settings.

Domain	Model	ID										OOD						Avg.			
		Admissions		IE		POE		Services		MIMIC-CXR		Transfers		Obesity		PadChest		Open-QA		Report	
		F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Rad	F1-Bio
General Domain	LLaVA 1.5-7B	75.95	36.91	75.16	29.21	75.66	33.51	75.08	34.12	1.79	68.12	75.79	41.27	72.23	9.99	11.75	68.78	74.98	30.83	6.77	68.45
	LLaMA 3-11B	74.68	35.65	72.73	24.38	74.83	36.18	74.97	34.34	4.38	65.01	74.88	38.01	70.51	8.68	7.54	62.28	73.77	29.54	5.96	63.64
	InternVL 3-8B	72.20	24.43	73.91	25.01	72.61	23.90	72.39	23.83	6.38	61.62	73.96	26.16	67.90	5.34	9.32	61.45	72.16	21.45	7.85	61.54
	Qwen 3-VL-8B	68.82	18.31	68.77	17.14	68.47	16.82	68.07	17.09	8.23	63.63	72.23	26.96	68.95	6.40	8.15	66.42	69.22	17.12	8.19	65.03
	Gemma 3-12B	72.33	26.56	71.86	25.30	73.57	30.07	74.57	33.98	10.27	68.75	74.72	25.58	66.93	4.07	12.31	65.35	72.33	24.26	11.29	67.05
	Grok 4.2	69.61	19.71	68.97	19.71	68.61	17.82	69.65	19.86	13.52	72.62	71.21	25.32	67.63	5.29	12.74	70.31	69.28	17.95	13.13	71.46
	Claude 4.5 Haiku	65.12	15.36	66.50	17.24	65.31	15.38	65.57	15.18	10.54	67.75	67.78	20.84	64.80	7.67	8.54	61.97	65.85	15.28	9.54	64.86
	GPT 5.4 Mini	68.86	23.15	65.55	26.80	69.59	23.77	68.84	22.84	15.88	72.30	71.24	30.65	65.78	7.07	10.35	68.39	68.31	22.38	13.11	70.34
Medical Domain	LLaVA-Med v1.5-7B	73.85	8.71	75.27	6.30	75.59	10.09	74.44	7.72	0.80	67.11	76.85	11.76	74.94	15.14	3.85	66.79	75.16	9.95	2.33	66.95
	Med-Flamingo-7B	58.64	11.86	55.51	11.36	58.68	13.04	58.61	11.56	0.14	40.54	62.24	9.51	57.95	4.91	2.63	52.08	58.61	10.37	1.38	46.31
	HuatuoGPT-Vision-7B	74.33	33.93	74.51	33.42	74.01	32.45	74.09	32.50	7.77	70.45	74.69	33.66	72.56	10.46	7.60	64.60	74.03	29.40	7.68	67.53
	HealthGPT-14B	69.91	29.27	70.67	27.16	71.03	26.87	72.15	28.22	9.05	73.04	66.58	27.69	66.49	7.49	2.03	68.48	69.47	24.45	5.54	70.76
	Lingshu-7B	76.17	38.17	74.68	35.01	75.88	33.47	76.26	33.23	16.18	78.19	76.11	35.19	72.95	11.16	1.23	68.85	75.34	31.04	8.71	73.52
	Hulu-Med-7B	76.20	34.79	74.43	14.55	75.47	30.74	75.62	27.64	16.31	75.84	76.06	26.06	73.79	12.45	5.52	69.15	75.26	24.37	10.91	72.49
	CODA-Qwen3-8B	**80.82**	**52.45**	**82.21**	**51.69**	**81.02**	**51.97**	**80.21**	**49.19**	**17.49**	**79.03**	**79.57**	**46.42**	**75.37**	**15.65**	**25.52**	**73.93**	**79.87**	**44.56**	**21.51**	**76.48**
	CODA-Llama3-8B	78.49	46.38	78.74	46.11	77.30	46.05	76.35	44.51	17.31	78.56	77.49	46.19	74.95	14.52	24.09	71.77	77.22	40.63	20.70	75.17
	CODA-Qwen3-8B w/o LoRA	80.29	50.50	81.06	48.43	80.47	51.64	79.54	48.37	16.76	78.08	78.58	44.45	74.15	14.82	23.32	71.64	79.02	43.04	20.04	74.86

its strong performance on the Obesity dataset, which represents a home-care scenario with limited EHR availability, as well as on PadChest, where the scale of EHR data is substantially smaller. These results indicate that CODA is robust to both data scarcity and domain shift, maintaining reliable performance across diverse clinical environments. **(iii) Robustness across backbones.** CODA consistently improves performance across diverse backbone architectures (e.g., Qwen3-VL and LLaMA-based models), demonstrating that the proposed cohort-aware mechanism is model-agnostic and broadly applicable. **(iv) Limitations of existing Med-LVLMs.** Existing Med-LVLMs are typically trained on large-scale corpora and a diverse set of tasks, largely mirroring general LVLM training paradigms with domain-specific data substitution. However, such models often overlook clinically meaningful reasoning processes and fail to incorporate domain-specific inductive biases or mechanisms. As a result, when data scale is no longer the primary bottleneck, their performance remains constrained by a lack of clinically grounded modeling.

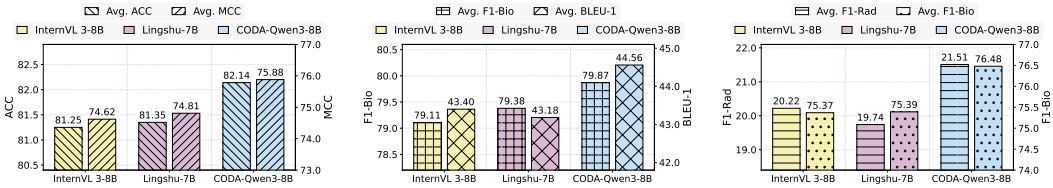
4.3 Ablation Study and In-Depth Analysis

Component-wise Ablation Study. To assess the contributions of key components in CODA, we conduct an ablation study, with results reported in Table 3 and full results in Appendix E.1. Removing the Cohort Context Encoding leads to consistent performance degradation in both in-distribution and out-of-distribution settings, particularly for Closed-QA and report generation tasks. This suggests that explicitly modeling cohort structure is essential for capturing population-level regularities. Removing deviation modeling results in additional performance drops, especially on Open-QA tasks, indicating that cohort-level representations alone are insufficient to capture fine-grained, patient-specific variations required for accurate clinical reasoning. Overall, the full model (CODA) achieves the best performance across all settings, demonstrating that the joint modeling of cohort structure and individual-level deviations is critical for robust generalization.

Comparison with Fine-tuned Baselines. To fully validate the effectiveness of CODA, we compare CODA-Qwen3-8B with two representative baselines, Lingshu-7B and InternVL 3-8B, with results presented in Figure 3 and full results in Appendix E.2. CODA consistently achieves the best performance across all tasks and metrics. These results demonstrate that CODA not only improves predictive accuracy but also enhances semantic consistency and clinical fidelity, highlighting the effectiveness of its cohort-aware and deviation-aware modeling in complex medical reasoning tasks.

Table 3: Ablation study results of CODA. "w/o Cohort" removes the Cohort Context Encoding, while "w/o Deviation" removes Drift-aware Deviation Modeling.

Model	ID						OOD					
	Closed-QA		Open-QA		Report		Closed-QA		Open-QA		Report	
	ACC	MCC	F1-Bio	BLEU-1	F1-Rad	F1-Bio	ACC	MCC	F1-Bio	BLEU-1	F1-Rad	F1-Bio
Vanilla Backbone	43.95	25.39	68.53	17.34	8.23	63.63	61.74	48.36	70.59	16.68	8.15	66.42
w/o Cohort	85.58	80.69	80.11	49.18	15.73	78.58	72.56	62.19	76.82	28.83	24.60	72.51
w/o Deviation	85.56	80.73	80.10	48.86	16.33	78.94	73.05	62.95	76.65	28.82	25.19	72.41
CODA	**86.34**	**81.78**	**81.06**	**51.32**	**17.49**	**79.03**	**73.76**	**64.08**	**77.47**	**31.03**	**25.52**	**73.93**



(a) Results of *Closed-QA*. (b) Results of *Open-QA*. (c) Results of *Report Generation*.

Figure 3: Comparisons with fine-tuned baselines.

Efficiency Study. To evaluate the computational efficiency of CODA, we compare it with representative general and medical baselines. The results are summarized in Table 4. Despite introducing cohort-aware population stratification and drift-aware individual adaptation, CODA maintains competitive inference efficiency while achieving substantially stronger clinical reasoning performance. Specifically, CODA achieves 7.95 s/sample and 20.19 tokens/s throughput, remaining comparable to widely used LLaVA-based baselines while exhibiting substantially lower sample latency than models with higher output complexity, such as InternVL3-8B and Qwen3-VL-8B. In addition, CODA maintains a practical memory footprint for standard single-GPU deployment, while the additional FLOPs/token introduced by cohort-aware conditioning remain modest. Overall, the results demonstrate that CODA achieves a favorable effectiveness–efficiency trade-off, enabling enhanced population-grounded clinical reasoning without sacrificing practical inference efficiency.

Interpretation Study. We visualize the learned patient representations of CODA on OOD tasks using t-SNE [45]. For clarity, we present results from the top-10 largest cohorts. As shown in Figure 4, samples from the same cohort form compact clusters, while different cohorts are largely well separated, indicating intra-cohort consistency and inter-cohort separability. We also observe mild overlap among cohorts 27, 28, and 29, which can be attributed to their shared ancestor nodes in the cohort hierarchy, leading to partially aligned representations. Overall, the learned embeddings exhibit coherent and discriminative geometric structures, where samples are grouped according to shared population-level patterns, suggesting that cohort semantics are effectively captured in the embedding space.

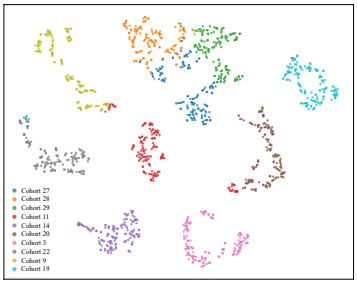
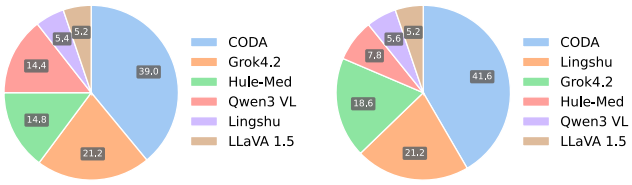


Figure 4: Visualization of embeddings within different cohorts.

4.4 Expert Review

To assess the practical significance of CODA, we conduct an expert review involving both GPT-5.4 and human experts. Specifically, we randomly sample 250 Open-QA instances from the OOD Transfers task and 250 report generation instances from the OOD PadChest task. We compare against five strong baselines, including Grok4.2, Hulu-Med, Qwen3 VL, Lingshu, and LLaVA 1.5. For each sample, experts are asked to select their first preference, i.e., the response that is most accurate and clinically reliable, from six candidate answers. The results are presented in Figure 5. CODA is consistently chosen as the first preference by both human experts and



(a) Results of human review on the first preference. (b) Results of GPT-5.4 review on the first preference.

Figure 5: Expert review results on the first preference.

Table 4: Efficiency comparison of CODA against different baselines.

Model	Tokens (tokens/sample)	Latency (s/sample)	Token Throughput (tokens/s)	Peak GPU Mem. (GB)	FLOPs (G/token)
LLaVA 1.5-7B	187.31	8.62	21.73	13.27	13.60
LLaVA-Med v1.5-7B	69.56	3.21	21.67	14.21	14.60
Hulu-Med-7B	72.75	5.79	28.95	15.65	14.40
Lingshu-7B	167.62	5.79	25.18	15.65	14.40
InternVL 3-8B	417.55	19.53	21.38	15.00	14.80
Qwen 3-VL-8B	545.65	23.15	23.57	16.47	15.03
CODA-8B	160.51	7.95	20.19	17.25	15.69

GPT-5.4, demonstrating its superior clinical utility and robustness. Among the baselines, Grok-4.2 and Lingshu are ranked as the second-best models by human experts and GPT-5.4, respectively. Despite these variations, all baselines remain substantially behind CODA, indicating that the proposed cohort- and drift-aware mechanisms provide consistent and reliable improvements.

We provide additional in-depth analyses on few-shot settings, hyperparameter sensitivity, and case studies in Appendices E.3, E.4, and F, further supporting the generality and reliability of CODA.

5 Conclusions

Generative AI has significantly advanced AI for healthcare. However, current medical foundation models largely overlook the interplay between population-level cohort patterns and patient-specific variation that underlies real-world clinical reasoning. In this paper, we propose CODA, a cohort- and drift-aware medical foundation model that explicitly discovers patient cohorts and integrates them into the reasoning pipeline. CODA introduces cohort-aware population stratification and drift-aware individual adaptation, providing a principled and effective framework for enhancing medical reasoning. Extensive experiments demonstrate that CODA consistently outperforms both general-purpose and medical-specific baselines across diverse clinical tasks, including closed-QA, open-QA, and report generation. Despite its effectiveness, CODA currently leverages structured feature representations and summary statistics for cohort construction, which may introduce additional preprocessing overhead in certain scenarios. Developing more automated and end-to-end approaches for cohort discovery and integration remains an important direction for future work.

Acknowledgements

We dedicate this work to the memory of Professor Beng Chin Ooi, whose wisdom, generosity, and vision deeply influenced our research and growth. His guidance and legacy will continue to inspire us.

This research is supported by the National Natural Science Foundation of China (624B2027), the Ministry of Education, Singapore, under an AcRF Tier 1 Grant (A-8002594-00-00) and the Key Research and Development Projects in Zhejiang Province (No. 2025C01128).

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] OpenAI. Gpt-4 technical report, 2023.

[3] Jiaqi Zhu, Naili Xing, Hexiang Pan, Haotian Gao, Jianwei Yin, Xiaokui Xiao, and Beng Chin Ooi. LogicTree-RAG: Logic tree-guided retrieval-augmented generation for long-form patent drafting. *arXiv preprint arXiv:2609.30943*, 2026.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[5] Changshuo Liu, Wu Junran, Zhongle Xie, Wenqiao Zhang, Kaiping Zheng, Jiaqi Zhu, Qingpeng Cai, Ooi Gene Anne, Marcus Chun Jin Tan, Jianwei Yin, James Wei Luen Yip, and Beng Chin Ooi. DIYHealth suite: Dataset, model, and benchmark for health management at home. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 77028–77076. PMLR, 06–11 Jul 2026.

[6] Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu. Do LLMs understand visual anomalies? uncovering LLM's capabilities in zero-shot anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 48–57, 2024.

[7] Nan Zhou, Ke Zou, Meng Liu, Linchao He, Jiaqi Zhu, Yi Zhang, Hu Chen, and Huazhu Fu. Hallucination detection and correction in medical vlms via counter-evidence verification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–194. Springer, 2026.

[8] Gang Chen, Changshuo Liu, Gene Anne Ooi, Marcus Tan, Zhongle Xie, Jianwei Yin, James Wei Luen Yip, Wenqiao Zhang, Jiaqi Zhu, and Beng Chin Ooi. Generative ai for healthcare: Fundamentals, challenges, and perspectives. *arXiv preprint arXiv:2510.24551*, 2025.

[9] Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey. *arXiv preprint arXiv:2504.21051*, 2025.

[10] Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025.

[11] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[12] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of Forty-second International Conference on Machine Learning (ICML)*, 2025.

[13] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.

[14] 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.

[15] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024.

[16] Jiaqi Zhu, Yunfeng Fan, Geng Han, Xiang Shi, Fang Deng, and Jie Chen. Generative anomaly detection: a comprehensive review of modeling principles, advances, and future opportunities. *Artificial Intelligence Review*, 2026.

[17] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[18] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

[19] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[20] Eta S Berner. *Clinical decision support systems*, volume 233. Springer, 2007.

[21] Xiaofeng Wang and Michael W Kattan. Cohort studies: design, analysis, and reporting. *Chest*, 158(1):S72–S78, 2020.

[22] Changshuo Liu, Lingze Zeng, Kaiping Zheng, Shaofeng Cai, Beng Chin Ooi, and James Wei Luen Yip. Neuralcohort: Cohort-aware neural representation learning for healthcare analytics. In *Forty-second International Conference on Machine Learning*, 2025.

[23] Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025.

[24] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[25] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

[27] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[28] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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

[30] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[31] xAI. Grok 4 technical overview. `https://x.ai/blog`, 2025. Technical report.

[32] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4, 2024.

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[34] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[35] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[36] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511, 2019.

[37] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, 2024.

[38] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[39] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

[40] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

[43] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[44] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[46] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[47] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[48] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.

[49] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

[51] 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.

[52] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[53] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781, 2020.

[54] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[55] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[57] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[58] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[59] 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.

[60] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885, 2023.

[61] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[62] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Appendix

- **A. Notations** 16
- **B. Related Work** 17
- **C. Implementation Details** 17
  - *C.1. Datasets* 17
  - *C.2. Model Implementation Details* 18
  - *C.3. Experimental Environment* 18
  - *C.4. Evaluation Metrics* 19
- **D. Cohort Structure** 20
- **E. Supplementary Experimental Results** 21
  - *E.1. Supplementary Ablation Study* 21
  - *E.2. Supplementary Comparisons with Fine-tuned Baselines* 21
  - *E.3. Comparison with Few-shot Baselines* 22
  - *E.4. Hyperparameter Study* 22
- **F. Case Study** 24
- **G. Broader Impacts** 26
- **NeurIPS Paper Checklist** 27

A Notations

To provide a comprehensive overview of the notations used throughout the paper, we present a summary of notations in Table 5 as a quick reference to facilitate the understanding and recall of each symbol.

Table 5: Notations.

Notation	Description
$\mathbf{x}$	Input instance, consisting of clinical records, a question and an image.
$\mathbf{y}$	The ground truth output sequence.
x_i	The i-th clinical records in $\mathbf{x}$.
$\mathbf{q}$	The question in $\mathbf{x}$.
$\mathcal{I}$	Input image of height H, width W, and three RGB channels.
$\mathcal{X}$	The textual input, including clinical records and a question.
$\mathcal{Z}^{(0)}$	Initial multimodal embedding, formed by visual and text embeddings.
$\mathcal{Z}^{(\ell)}$	Output of ℓ-th transformer block.
L_b	The number of transformer blocks in the backbone.
$f_{I\to t}$	The function that maps the visual embeddings into the shared textual embedding space.
$\mathcal{E}_I(\cdot)$	The visual encoder that maps images to embeddings.
$\mathcal{E}_t(\cdot)$	The textual encoder that maps words to embeddings.
L_I	The number of visual tokens in image $\mathcal{I}$, similar to L_t.
d_t	The dimensions of shared textual embedding space.
p_θ	The conditional distribution parameterized by a large language model with parameters θ.
$\mathcal{M}$	Markov decision process (MDP).
$\mathcal{S}$, $\mathcal{A}$, P, r and γ	State space, action space, transition dynamics, reward function and discount factor, respectively.
π_ϕ	Policy, $\pi_\phi(a \mid s)$ parameterized by ϕ, which in this work corresponds to the cohort-partitioning policy.
τ	Trajectory generated by the policy, i.e., a sequence of states and actions.
$\mathcal{F}^{(\ell)}$	The ℓ-th Transformer block.
$\mathcal{T}$	The learned hierarchical cohort tree.
$\mathcal{P}$	The set of patients.
$\mathcal{C}$	The set of cohorts.
$\mathbf{h}_\mathcal{C}$	The cohort prototypes of cohorts.
$\Delta(\mathbf{x}, c)$	The deviation between input $\mathbf{x}$ and cohort c.
θ	The learnable model parameters.
$p_\tau(\cdot \mid \pi_\phi)$	The trajectory distribution induced by policy π_ϕ
$\mathcal{R}(\mathbf{x}, \tau)$	The trajectory-level cohort discovery reward.
$\mathcal{H}(p)$	Entropy of a distribution p.
$p_{\pi_\phi}(\cdot \mid \mathbf{x})$	Policy distribution conditioned on input $\mathbf{x}$.
$g(\cdot)$	The function that maps serialized EHR records into a vector of statistical features.
V	The set of nodes in $\mathcal{T}$, similar to E. which is the set of edges in $\mathcal{T}$.
$\mu(\cdot)$	The mean function.
$\sigma(\cdot)$	The standard deviation function.
d_v	Depth of node v.
$\mathbf{U}_{\mathcal{P}_v}$	Set of statistical features of patients $\mathcal{P}_v$.
z_v	Binary stopping variable for node v.
j_v	The selected feature for node v.
t_v	The split threshold for node v.
$d_{\max}$	The maximum depth.
$\mathbf{e}_c$	Cohort embedding that will be injected into the multimodal input.
$\mathcal{P}_c$	The patients within cohort c.
$\mathcal{L}_{\text{gen}}$	Downstream generation loss.
R_v	The generation-aware reward.
$\mathbf{h}_p$	Learned representation of patient p.
$\mathbf{W}_\ell^\alpha$	Learnable weight matrix, similar to $\mathbf{W}_\ell^\beta$, $\mathbf{W}_\ell^\gamma$, $\mathbf{W}_\ell^o$ and $\mathbf{W}_d$.

B Related Work

Foundation Models in General Domain. The rapid development of LLMs has significantly reshaped modern AI, enabling strong general-purpose reasoning and coherent, human-like interactions. Beginning with generative pre-trained Transformers (GPT) [1, 2], it has been widely proven that scaling model size and training data leads to substantial improvements in general-purpose question answering across a broad range of tasks [46, 47, 48]. Subsequent advances further enhanced vision language alignment and multimodal understanding through instruction tuning [49, 26, 50] and reinforcement learning from human feedback (RLHF) [51]. Beyond generative modeling, representation learning approaches, including embedding models such as E5 [52] and dense retrieval frameworks like DPR [53], have supported retrieval augmented generation (RAG) [54] by providing high-quality semantic representations. These approaches allow LLMs to incorporate external knowledge, thereby improving factual consistency and mitigating hallucinations. More recently, advanced LLMs, such as Claude [32], GPT-5 [55] and Gemini [56], have demonstrated strong performance in complicated multimodal understanding and generation tasks. Yet, the general-purpose LLMs are not readily applicable to specialized domains such as healthcare. In such high-stakes settings, effective reasoning requires not only general linguistic competence but also a deep understanding of domain-specific knowledge and complex clinical data.

Foundation Models in Medical Domain. Following the success of general-purpose LLMs, extensive efforts have been made to adapt these models to the medical domain. Early efforts focus on domain-specific pretraining, where models such as BioBERT [43], PubMedBERT [57] and ClinicalBERT [58] are trained on large-scale biomedical corpora to learn medical domain knowledge and preserve medical semantics. More recently, Med-PaLM [59] and HuatuoGPT [60] are proposed to extend LLMs to clinical applications, including medical question answering, report generation, and medical decision support. At the same time, multimodal medical LVLMs are developed to integrate heterogeneous clinical data. For example, LLaVA-Med [11] and Med-Flamingo [17] enables joint reasoning over radiology images and medical reports. Building upon these advances, subsequent models, exemplified by HealthGPT [12] and Hulu-Med [35], further scale medical corpora and task diversity to enhance the medical understanding of complex clinical tasks involving both structured and unstructured data. Despite these improvements, existing models mainly operate at the individual instance level, treating each patient independently and relying on implicit medical knowledge learned by model parameters. However, real-world clinical reasoning is inherently cohort-based, where clinicians often compare patients within similar populations and reason about deviations from the typical patterns. The lack of explicit modeling of cohort structure and patient-level derivation limits the ability of current foundation models to capture clinically meaningful heterogeneity.

C Implementation Details

C.1 Datasets

The experiments are conducted on four real-world EHR benchmarks: EHR-Bench [23], DIYHealth-Bench [5], MIMIC-CXR [24], and PadChest [25]. The task description for each dataset is introduced in Table 6. EHR-Bench and DIYHealth-Bench provide raw QA pairs. For the MIMIC-CXR and PadChest datasets, we further preprocess the EHR records and radiology reports into unified QA-pair formats to ensure consistency with EHR-Bench and DIYHealth-Bench. To balance evaluation reliability and computational efficiency, datasets containing more than 2,500 samples are split by reserving 500 samples for testing and using the remaining samples for training. For smaller datasets, we adopt a 4:1 training-to-test split ratio.

Table 6: Task descriptions across datasets.

Dataset	Task Description
Admissions	Predict the next admission type from EHR records (Closed-QA) and provide clinical reasoning (Open-QA).
Services	Predict the next clinical department from EHR records (Closed-QA) and provide clinical reasoning (Open-QA).
Ingredient Events (IE)	Predict the next administered medication or substance from EHR records (Closed-QA) and provide clinical reasoning (Open-QA).
Provider Order Entry (POE)	Predict the next clinical order type from EHR records (Closed-QA) and provide clinical reasoning (Open-QA).
Transfers	Predict the next care location from EHR records (Closed-QA) and provide clinical reasoning (Open-QA).
Obesity	Assess obesity risk from self-reported EHR records (Closed-QA) and provide supporting analysis (Open-QA).
MIMIC-CXR	Generate radiology reports from EHR records and chest X-ray images.
PadChest	Generate radiology reports from a few EHR records and chest X-ray images.

C.2 Model Implementation Details

We adopt Qwen3-8B and LLaMA3-8B as backbone models. The projector $f_{I \to t}$ is implemented as a two-layer MLP. Both $g(\cdot)$ and f_c are instantiated as MLPs, and $f_{\mathbf{u} \to \mathbf{h}}$ is an MLP-based encoder. The maximum depth $d_{\max}$ is set to 100. We set $\lambda_g = 0.3$ and $\lambda_d = 10^{-4}$ across all experiments. The coefficient β is set to 0.01 following prior work [61, 62], and the margin m is set to 0.2 following prior study [63]. The assignment confidence of patient p to cohort c is computed based on the similarity between $\mathbf{h}_p$ and $\mathbf{h}_c$.

We present the detailed configurations for the training process. The specific settings are listed in Table 7. These hyperparameters are set up following prior studies [12, 26, 11, 5]. The architectural parameters are listed in Table 8.

Table 7: Summary of configurations for the training process.

Configuration	Stage (i)	Stage (ii)	Stage (iii)
Optimizer	AdamW	AdamW	AdamW
Adapter LR	/	5e-5	1e-4
Learning Rate	2e-5	1e-4	2e-5
Weight Decay	/	0.01	0.01
Dropout Rate	0.05	0.1	0.1
LR Scheduler	Warmup	Constant	Cosine w/ Warmup
Max Sequence Length	32k	32k	32k

Table 8: Architectural parameters of CODA.

Module	**Parameter**	**Value / Shape**
Tokenizer	Vocabulary size	151,670
Backbone (Qwen3-VL)	Hidden size	4096
	# Transformer layers	36
	Embedding matrix	(151670, 4096)
	Self-attn projections	q: (4096, 4096), k/v: (4096, 1024), o: (4096, 4096)
	MLP	gate/up: (4096, 12288), down: (12288, 4096)
LoRA	Rank r	16
	Matrix $\mathbf{A}$	q/k/v/o/gate/up: (16, 4096), down: (16, 12288)
	Matrix $\mathbf{B}$	q/o: (4096,16), k/v: (1024,16), gate/up: (12288,16), down: (4096,16)
Stage ii	Cohort embedding dim	1024
	Prefix length	8
	Prefix projector	(4096, 4096) $\to$ (4096, 8×4096)
	Cohort encoder hidden dim	4096
Stage iii	Drift dim	1024
	Adapter bottleneck dim	256
	Adapter (down / up)	(4096, 256), (256, 4096)
	Drift generator	(1024, 512)
	Gating module	(1024, 4096)
Cohort Memory	# Cohorts	34
	Prototype matrix	(34, 1024)

C.3 Experimental Environment

All experiments are conducted on a server equipped with an AMD EPYC 9334 CPU (32 cores, 2.7 GHz), 128 GB RAM, and an NVIDIA H100 NVL GPU with CUDA 12.9. The system runs Ubuntu 24.04 with Linux kernel 6.8.0-79-generic.

C.4 Evaluation Metrics

C.4.1 Evaluation Metrics for Closed-QA

- **Accuracy.** Accuracy (ACC) evaluates the overall prediction performance by computing the fraction of correctly classified instances among all samples. It provides a coarse-grained measure of classification effectiveness, especially under balanced label distributions.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

  where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

- **Matthews Correlation Coefficient.** The Matthews Correlation Coefficient (MCC) provides a holistic evaluation of classification performance by capturing the correlation between predicted and true labels through the full confusion matrix. In contrast to accuracy, MCC remains robust under class imbalance, as it simultaneously considers all types of prediction outcomes and their interactions across classes. As a normalized correlation measure, MCC ranges from -1 to 1, where 1 indicates perfect alignment, 0 corresponds to chance-level prediction, and -1 reflects complete disagreement.

$$\text{MCC} = \frac{N \sum_k C_{kk} - \sum_k T_k P_k}{\sqrt{\left(N^2 - \sum_k P_k^2\right)\left(N^2 - \sum_k T_k^2\right)}} \tag{19}$$

  where $C \in \mathbb{R}^{K \times K}$ denotes the confusion matrix, and C_{kl} represents the number of samples whose ground-truth class is k and predicted class is l. $T_k = \sum_l C_{kl}$ and $P_k = \sum_l C_{lk}$ denote the total number of true and predicted instances for class k, respectively. $N = \sum_k \sum_l C_{kl}$ is the total number of samples, and K is the number of classes.

C.4.2 Evaluation Metrics for Open-QA

- **F1-BioBERT.** F1-BioBERT evaluates semantic consistency between generated answers and ground-truth responses by leveraging contextual embeddings from BioBERT [43]. Instead of relying on exact lexical matching, it computes similarity at the representation level, enabling robust assessment of paraphrased or semantically equivalent expressions in medical text.

  Specifically, precision and recall are defined based on semantic matching between predicted and reference tokens (or spans) in the embedding space, and the final score is computed as their harmonic mean:

$$\text{F1-BioBERT} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{20}$$

  Here, matches are determined according to similarity scores derived from BioBERT representations, with higher scores indicating stronger semantic alignment.

- **BLEU-1.** BLEU-1 evaluates lexical fidelity by measuring the modified unigram precision between the generated text and the reference. Compared with higher-order BLEU variants, BLEU-1 focuses on word-level overlap and is therefore suitable for assessing whether key lexical content is preserved.

$$\text{BLEU-1} = \text{BP} \cdot \exp\left(\log p_1\right), \tag{21}$$

  where p_1 is the modified unigram precision:

$$p_1 = \frac{\sum_{w \in \mathcal{V}} \min\left(\text{Count}_{\text{pred}}(w), \text{Count}_{\text{ref}}(w)\right)}{\sum_{w \in \mathcal{V}} \text{Count}_{\text{pred}}(w)}. \tag{22}$$

  Here, $\mathcal{V}$ denotes the unigram vocabulary, while $\text{Count}_{\text{pred}}(w)$ and $\text{Count}_{\text{ref}}(w)$ denote the frequency of unigram w in the generated and reference texts, respectively. BP is the brevity penalty that discourages overly short generations.

C.4.3 Evaluation Metrics for Report Generation

- **F1-RadGraph.** F1-RadGraph evaluates the quality of generated radiology reports by measuring the overlap between predicted and reference clinical graphs constructed using RadGraph [44]. Specifically, it assesses the correctness of extracted medical entities and their associated relations, thereby capturing both factual accuracy and clinical consistency at the structured level. The metric is computed as the harmonic mean of precision and recall over matched graph elements:

$$\text{F1-RadGraph} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

  where precision and recall are defined based on the matching between predicted and ground-truth RadGraph annotations, with true positives corresponding to correctly identified entities and relations, and false positives/negatives representing mismatched or missing clinical concepts.

D Cohort Structure

The learned cohort tree provides an interpretable hierarchical decomposition of the patient space, where each internal node corresponds to a clinically meaningful decision criterion, and each leaf node defines a distinct cohort, as shown in Figure 6. At the top levels, the model prioritizes global clinical signals such as cardiac-related markers, care context (e.g., discharge disposition and observation status), and multimorbidity, indicating that population stratification is first driven by overall severity and healthcare utilization. As the hierarchy progresses, the tree further branches along complementary axes, including chronic disease patterns (e.g., cardiometabolic conditions) and imaging-derived phenotypes (e.g., pleural effusion and pneumothorax), demonstrating that the model disentangles heterogeneous patient populations using semantically coherent, multimodal criteria. At deeper levels, the structure shifts toward fine-grained refinement, where demographic attributes (e.g., age and gender) and laboratory-derived signals (e.g., hematologic and metabolic abnormalities) are used to differentiate patients with subtle physiological variations. Notably, certain features recur across different branches, reflecting context-dependent decision boundaries rather than fixed global rules. In addition, the tree exhibits adaptive depth, with simpler cohorts identified early and more complex populations refined through deeper branches.

Overall, this hierarchy reveals that the proposed method learns clinically grounded and interpretable cohort structures, organizing patients into meaningful subgroups that support coherent and reliable downstream reasoning.

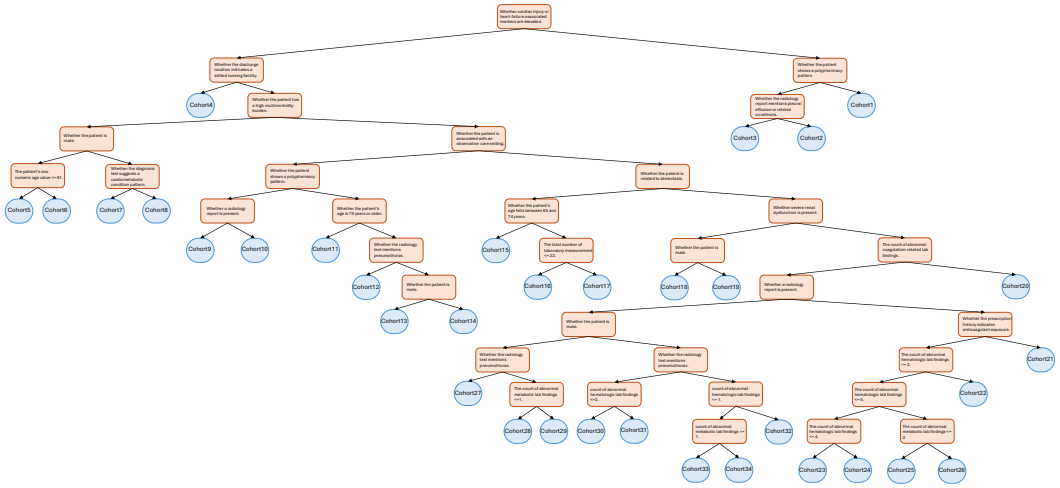


Figure 6: Illustration of Cohort Structure.

E Supplementary Experimental Results

E.1 Supplementary Ablation Study

Tables 9 and 10 provide extended ablation results across closed-QA, open-QA, and report generation settings. Consistent with the main paper, removing either cohort-aware population stratification or drift-aware individual adaptation leads to performance degradation across both in-distribution and out-of-distribution tasks, demonstrating that the two components are complementary and jointly contribute to robust multimodal clinical reasoning.

Under closed-QA settings, CODA achieves the best overall performance across all tasks and evaluation metrics. Similar trends are observed in open-QA and report generation settings. CODA consistently achieves the strongest performance across both language-level and clinical evaluation metrics. Overall, the results demonstrate that population-level cohort structure and patient-specific deviation modeling are both critical for reliable multimodal clinical reasoning.

Table 9: Ablation Study of CODA under *Closed-QA* settings.

Model	ID								OOD				Avg.	
	Admissions		IE		POE		Services		Transfers		Obesity			
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
LoRA Only	74.60	65.78	91.92	89.04	81.20	74.83	92.02	89.42	73.60	64.65	70.30	58.48	80.61	73.70
w/o Cohort	75.20	66.57	92.17	89.57	82.20	76.25	92.77	90.38	74.20	65.16	70.92	59.21	81.24	74.52
w/o Deviation	75.80	67.37	92.17	89.51	82.00	76.00	92.27	90.06	74.40	65.70	71.70	60.20	81.39	74.81
CODA	**76.60**	**68.77**	**92.42**	**89.93**	**82.80**	**77.05**	**93.52**	**91.39**	**75.20**	**66.89**	**72.32**	**61.28**	**82.14**	**75.88**

Table 10: Ablation Study of CODA under *Open-QA* and *report generation* settings.

Model	ID										OOD						Avg.			
	Admissions		IE		POE		Services		MIMIC-CXR		Transfers		Obesity		PadChest		Open-QA		Report	
	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Rad	F1-Bio
LoRA Only	79.06	46.09	80.94	47.20	80.16	50.27	78.87	46.24	15.46	78.74	78.61	44.18	74.61	13.44	24.68	72.24	78.71	41.24	20.07	75.49
w/o Cohort	79.71	48.02	81.10	49.08	80.44	52.03	79.20	47.60	15.73	78.58	78.58	43.66	75.07	14.01	24.60	72.51	79.02	42.40	20.17	75.55
w/o Deviation	79.80	48.64	81.18	48.99	80.06	50.50	79.34	47.32	16.33	78.94	78.55	43.68	74.76	13.96	25.19	72.41	78.95	42.18	20.76	75.67
CODA	**80.82**	**52.45**	**82.21**	**51.69**	**81.02**	**51.97**	**80.21**	**49.19**	**17.49**	**79.03**	**79.57**	**46.42**	**75.37**	**15.65**	**25.52**	**73.93**	**79.87**	**44.56**	**21.51**	**76.48**

E.2 Supplementary Comparisons with Fine-tuned Baselines

Tables 11 and 12 provide the full comparisons between CODA and representative LoRA fine-tuned multimodal medical foundation models under closed-QA, open-QA, and report generation settings. CODA consistently achieves the best overall performance across both in-distribution and out-of-distribution tasks.

Table 11: Comparison of CODA with LoRA fine-tuned baselines under *Closed-QA* settings.

Model	ID								OOD				Avg.	
	Admissions		IE		POE		Services		Transfers		Obesity			
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
InternVL 3-8B	76.00	67.98	91.18	88.51	81.20	74.93	93.27	91.02	74.60	65.75	71.23	59.52	81.25	74.62
Lingshu-7B	76.20	68.24	91.67	88.85	81.20	74.96	93.27	91.07	74.40	65.52	71.38	60.24	81.35	74.81
CODA-Qwen3-8B	**76.60**	**68.77**	**92.42**	**89.93**	**82.80**	**77.05**	**93.52**	**91.39**	**75.20**	**66.89**	**72.32**	**61.28**	**82.14**	**75.88**

Table 12: Comparison of CODA with LoRA fine-tuned baselines under *Open-QA* and *report generation* settings.

Model	ID										OOD						Avg.			
	Admissions		IE		POE		Services		MIMIC-CXR		Transfers		Obesity		PadChest		Open-QA		Report	
	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Rad	F1-Bio
InternVL 3-8B	80.73	52.10	81.54	50.22	80.91	51.51	80.18	**49.39**	16.59	78.69	79.14	46.00	72.17	11.21	23.85	72.05	79.11	43.74	20.22	75.37
Lingshu-7B	80.60	48.92	82.16	50.76	80.99	51.65	80.01	49.03	16.89	78.88	78.82	**46.64**	73.72	12.07	22.58	71.90	79.38	43.44	19.74	75.39
CODA	**80.82**	**52.45**	**82.21**	**51.69**	**81.02**	**51.97**	**80.21**	49.19	**17.49**	**79.03**	**79.57**	46.42	**75.37**	**15.65**	**25.52**	**73.93**	**79.87**	**44.56**	**21.51**	**76.48**

E.3 Comparison with Few-shot Baselines

To further evaluate CODA against strong baselines, we augment several representative models with few-shot demonstrations. The results are reported in Tables 13 and 14. Due to the rich and heterogeneous nature of EHR data, each sample is relatively long, posing challenges for models with limited context length constraints. While InternVL 3 and Hulu-Med benefit from few-shot demonstrations, LLaVA 1.5 and Lingshu exhibit performance degradation, suggesting insufficient capability to effectively process long-context inputs. In particular, the extreme degradation observed in LLaVA 1.5 under the 2-shot setting can also be attributed to its 4K context window, which is insufficient to accommodate both demonstrations and the query. Overall, these results further highlight the effectiveness of CODA.

Table 13: Comparison of CODA with few-shot baselines under *Closed-QA* settings.

Model	ID								OOD				Avg.	
	Admissions		IE		POE		Services		Transfers		Obesity			
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
LLaVA 1.5-7B 1-shot	35.60	17.73	20.20	0.94	41.80	26.70	55.36	41.83	54.60	39.33	58.79	44.27	44.39	28.46
LLaVA 1.5-7B 2-shot	10.60	5.68	8.59	5.72	14.20	11.59	18.20	16.18	46.20	34.57	49.61	35.99	24.57	18.29
InternVL 3-8B 1-shot	38.00	17.64	29.04	5.55	45.40	27.36	55.36	40.53	38.20	17.96	53.65	37.40	43.28	24.41
InternVL 3-8B 2-shot	42.60	23.74	30.30	6.82	43.40	24.72	56.86	42.46	42.60	23.92	53.50	36.94	44.88	26.43
lingshu-7B 1-shot	38.80	19.01	34.85	14.62	49.80	33.25	69.08	58.83	49.20	33.08	65.01	51.82	51.12	35.10
lingshu-7B 2-shot	45.60	27.99	38.13	17.71	48.40	31.26	70.32	60.57	51.00	34.97	68.58	56.05	53.67	38.09
Hulu-Med-7B 1-shot	43.40	24.78	30.81	7.86	57.60	43.66	65.84	54.53	49.20	32.28	60.19	47.34	51.17	35.07
Hulu-Med-7B 2-shot	49.60	33.02	34.09	12.33	59.60	46.31	68.58	58.59	49.40	32.60	63.61	50.78	54.15	38.94
CODA-Qwen3-8B	**76.60**	**68.77**	**92.42**	**89.93**	**82.80**	**77.05**	**93.52**	**91.39**	**75.20**	**66.89**	**72.32**	**61.28**	**82.14**	**75.88**

Table 14: Comparison of CODA with few-shot baselines under *Open-QA* and *report generation* settings.

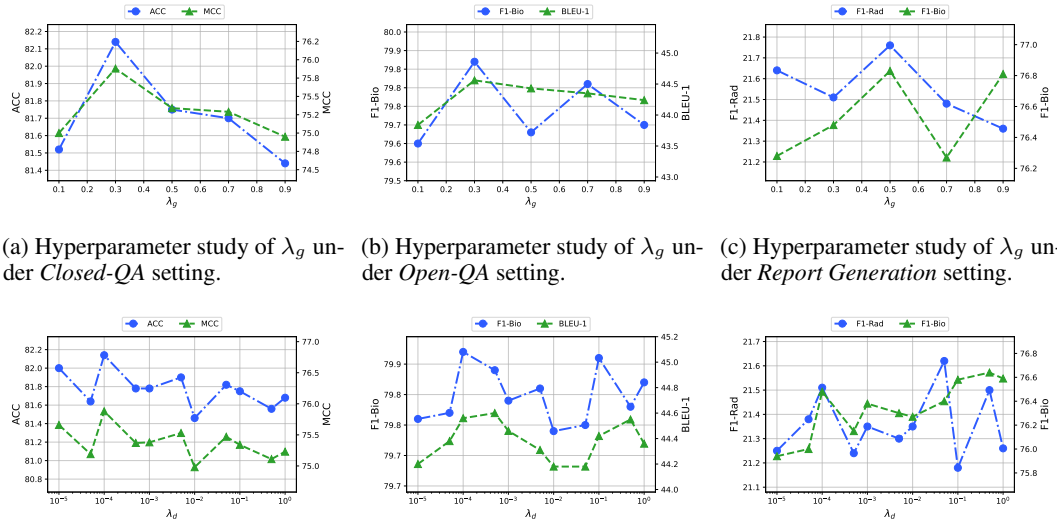
Model	In-distribution Task										Out-of-distribution Task						Avg.			
	Admissions		IE		POE		Services		MIMIC-CXR		Transfers		Obesity		PadChest		Open-QA		Report	
	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Bio	BLEU-1	F1-Rad	F1-Bio	F1-Bio	BLEU-1	F1-Rad	F1-Bio
LLaVA 1.5-7B 1-shot	72.42	23.91	68.45	17.87	71.84	22.52	72.93	22.99	4.17	68.19	76.07	28.96	73.11	11.48	9.25	67.83	72.47	21.29	6.71	68.01
LLaVA 1.5-7B 2-shot	51.98	6.14	50.39	3.98	51.28	5.24	52.83	6.56	2.44	54.40	68.85	23.58	67.65	9.44	4.69	45.58	57.16	9.16	3.57	49.99
InternVL 3-8B 1-shot	74.40	27.94	75.70	28.50	74.89	28.06	74.92	28.32	9.18	67.77	75.44	30.32	69.46	7.04	8.63	63.21	74.13	25.03	8.91	65.49
InternVL 3-8B 2-shot	75.48	30.00	76.24	30.14	75.69	30.34	75.61	30.27	9.32	68.25	76.11	31.96	69.45	6.98	8.31	63.76	74.76	26.62	8.82	66.01
lingshu-7B 1-shot	77.09	38.61	75.65	30.27	76.54	33.60	76.76	34.11	11.98	75.19	76.97	32.17	73.04	10.67	7.85	67.36	76.01	29.91	9.92	71.27
lingshu-7B 2-shot	77.55	39.27	76.18	30.57	76.74	34.79	77.21	34.89	11.53	74.36	77.25	33.63	72.88	10.54	7.67	66.75	76.30	30.61	9.60	70.55
Hulu-Med-7B 1-shot	76.93	31.16	75.42	24.08	76.10	29.40	76.56	29.89	13.86	75.77	76.91	26.78	72.87	10.58	8.86	64.90	75.80	25.32	11.36	70.34
Hulu-Med-7B 2-shot	77.04	32.19	75.83	23.07	76.25	30.06	76.76	29.04	14.50	76.32	77.01	29.95	72.85	10.36	8.29	65.52	75.96	25.78	11.40	70.92
CODA-Qwen3-8B	**80.82**	**52.45**	**82.21**	**51.69**	**81.02**	**51.97**	**80.21**	**49.19**	**17.49**	**79.03**	**79.57**	**46.42**	**75.37**	**15.65**	**25.52**	**73.93**	**79.87**	**44.56**	**21.51**	**76.48**

E.4 Hyperparameter Study

We conduct a sensitivity study on two key hyperparameters, λ_g and λ_d, with results shown in Figure 7. As observed in Figure 7(a–c), performance consistently peaks at moderate values of λ_g (around 0.3–0.5), highlighting the importance of balancing the generation objective and cohort-aware signals. Both smaller and larger values lead to performance degradation, corresponding to the under- and over-dominance of the generation objective, respectively.

In contrast, Figure 7(d–f) shows that the model is relatively insensitive to λ_d, with stable performance across several orders of magnitude. The best results are typically achieved around 10^{-4} or 10^{-1}, suggesting that the drift loss primarily acts as a regularizer.

Overall, these results demonstrate that our method achieves robust performance without requiring precise hyperparameter tuning.



(a) Hyperparameter study of λ_g under *Closed-QA* setting.

(b) Hyperparameter study of λ_g under *Open-QA* setting.

(c) Hyperparameter study of λ_g under *Report Generation* setting.

(d) Hyperparameter study of λ_d under *Closed-QA* setting.

(e) Hyperparameter study of λ_d under *Open-QA* setting.

(f) Hyperparameter study of λ_d under *Report Generation* setting.

Figure 7: Hyperparameter study on λ_g and λ_d.

F Case Study

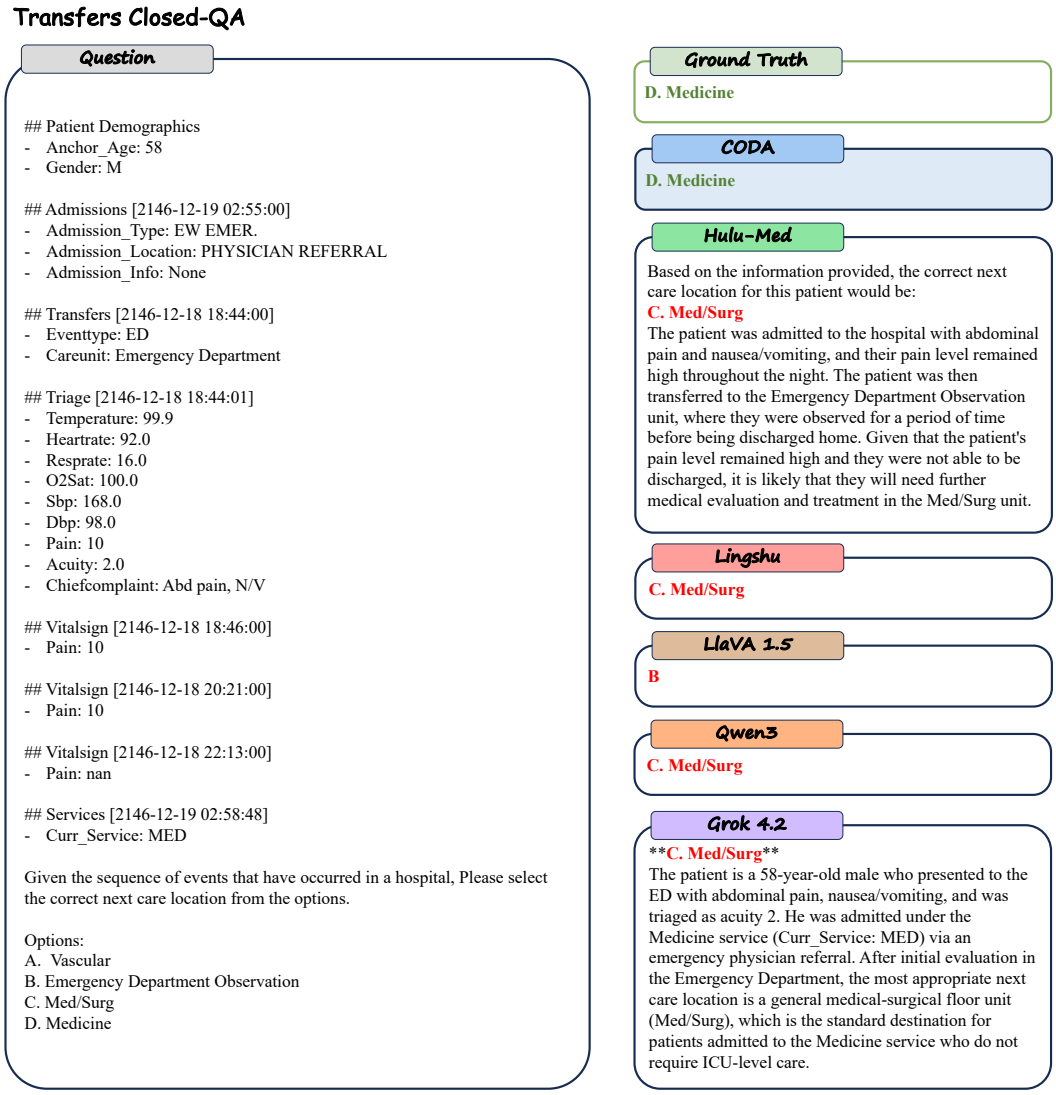


Figure 8: A case study of Transfers task under *Closed-QA* setting.

PadChest Report Generation



Question

- Subject_ID: 19673087215800739904802193294981955729 9
- Anchor_Age: 86
- Gender: M
- PatientBirth: 1923
- Pediatric: No

Current Imaging Study [20091211]
- Study_ID: 216840111366964012558082906712009345092636845
- Image_View: AP_horizontal
- MethodProjection: Manual review of DICOM fields
- Localizations: loc upper lobe, loc left, loc pleural, loc central, loc lobar, loc hemithorax, loc right, loc tracheal, loc superior cave vein, loc bilateral, loc right upper lobe, loc bronchi, loc retrocardiac

Generate the chest X-ray radiology report based on the above information.

Qwen3

Chest X-ray Radiology Report (Research Test Purpose Only)
Clinical Findings:The chest X-ray demonstrates an elderly male patient with significant age-related changes. The cardiac silhouette appears enlarged, and there is evidence of volume loss in the right upper lobe. **The left lung field shows increased density**, **possibly indicating consolidation or atelectasis**. The **right lung is relatively clear**, but there is subtle blunting of the right costophrenic angle. The trachea is midline, and the mediastinum is slightly shifted to the left. The superior vena cava is visible and appears dilated. There is a faint linear opacity in the left lower lobe, possibly representing a pleural effusion or fibrotic change. The right hilum is slightly prominent. The diaphragm appears flattened, and there is mild scoliosis of the thoracic spine. The ribs are thin and show signs of osteoporosis.
Localization Observations:
- **Upper lobe involvement on both sides**
- Left-sided predominance of findings
- Central and mediastinal structures are displaced
- Bilateral pleural involvement
- Right upper lobe volume loss
- Bronchi appear patent
- Retrocardiac region shows increased density\n- Tracheal deviation to the left
- **Superior vena cava is dilated**
- Bilateral lung fields are asymmetric

Impression:
This chest X-ray reveals findings consistent with **chronic pulmonary disease**, possible volume loss in the right upper lobe, and **possible pleural effusion** or **fibrosis** on the left. The enlarged cardiac silhouette and dilated superior vena cava suggest **possible chronic heart failure** or superior vena cava syndrome. The patient's advanced age and **osteoporotic changes** are consistent with his age. Further clinical correlation and possibly CT imaging are recommended for definitive diagnosis.

LlaVA 1.5

The image is a black and white chest X-ray of an adult male patient. The patient was born in 1923. The radiology report indicates that the patient has a **large right upper lobe mass**, which is a potential concern. The image also shows a **large right pleural effusion**, which is another area of concern. The **patient's overall condition appears to be stable**, but further evaluation and follow-up are recommended to determine the cause and severity of these findings.

Ground Truth

Marked radiologic worsening. There is bilateral interstitial-alveolar involvement, predominantly in the upper lobes, chiefly the right upper lobe. Increased retrocardiac density with a small left pleural effusion component. Findings compatible with bronchopneumonia, pulmonary edema, and respiratory distress. Endotracheal tube in correct position, as well as nasogastric tube. Pleural drainage tube in the upper third of the left hemithorax. Right jugular central line with distal tip in the superior vena cava.

CODA

Endotracheal tube in correct position. **Nasogastric tube** with distal tip in the stomach. **Central venous catheter with distal tip in the superior vena cava**. **Right upper lobe infiltrate** with bronchiectasis and pleural thickening. Bilateral pleural effusions. **Retrocardiac infiltrate**. Right hemithorax volume loss.

Hulu-Med

The chest X-ray shows a **large mass in the right upper lobe**. The mass is well-defined and appears to be centrally located. There is **no evidence of pleural effusion** or pneumothorax. The trachea is midline and there are no signs of mediastinal shift. The heart size is normal. The left lung appears clear.

Lingshu

Findings: The heart is normal in size and contour. The **lungs are clear without evidence of infiltrate**. There is **no pneumothorax or effusion**. Impression: **No acute cardiopulmonary disease**.

Grok 4.2

Findings:The chest radiograph demonstrates **a right internal jugular central venous catheter** with the distal tip terminating at the cavoatrial junction. An **endotracheal tube is in place**, with the tip positioned approximately 4 cm above the carina. **A nasogastric tube is also seen coursing below the diaphragm**. There is a **moderate-to-large right pleural effusion** with associated right hemithorax opacification. A left pleural effusion is also present, though smaller in size. **bibasilar** and **retrocardiac airspace opacities** are noted, consistent with **compressive atelectasis or aspiration** in the setting of the pleural fluid. **No focal lobar consolidation is identified** to suggest pneumonia.
The cardiomediastinal silhouette is partially obscured by the bilateral effusions and overlying support apparatus. The visualized osseous structures are intact for the patient's age

Impression:
1. Moderate-to-large right and small left pleural effusions with associated basilar atelectasis.
2. Appropriately positioned right internal jugular central venous catheter, endotracheal tube, and nasogastric tube.
3. No definitive focal pneumonia identified on this single portable view.

Figure 9: A case study of PadChest dataset under *Report Generation* setting.

G Broader Impacts

Current medical foundation models overlook patient-specific characteristics, which are critical for accurate clinical decision-making. By explicitly modeling cohort structure and individual-level deviations, CODA has the potential to support clinicians in delivering more personalized care, particularly in settings with heterogeneous patient populations, and to enhance transparency by exposing cohort-level patterns and patient-specific factors that influence model predictions.

At the same time, several risks should be considered. First, cohort construction may inadvertently encode biases present in the underlying data, potentially leading to disparate performance across demographic groups. Second, reliance on structured features and summary statistics may limit applicability in scenarios where such information is incomplete or noisy. Third, as with other generative models, there remains a risk of producing plausible but incorrect outputs, which could be misinterpreted in high-stakes clinical contexts.

To mitigate these risks, future work should incorporate fairness-aware cohort construction, uncertainty estimation, and human-in-the-loop validation to ensure safe deployment. In addition, rigorous evaluation across diverse populations and clinical settings is necessary to assess generalizability and prevent unintended harms. Overall, while CODA offers a step toward more interpretable and patient-aware medical AI, careful consideration of these societal implications is essential for responsible use.

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: We have clearly claimed our contribution and scope.

   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 have discussed the limitations in Sec 5.

   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: [N/A]

   Justification: Our work 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 have uploaded the codes to supplementary materials 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: We have uploaded the codes to supplementary materials for reproducibility.

   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: We have specified all the training and test details.

   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: We use expert review to validate the significance of our method in real world.

   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 have claimed the information of the computer resources.

   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 obey 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: We have discussed the societal impacts in the 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: This work does not pose such risks.

    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 cited the corresponding assets.

    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: [Yes]

    Justification: The new assets are introduced with documentation.

    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: Our work does not involve crowdsourcing nor research with human subjects.

    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: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    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: [N/A]

    Justification: Core method development in this research does not involve LLMs as any important, original, or non-standard components.

    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.