

ATHENA: Knowledge-guided agentic neural architecture search for AutoFormer-based electronic health record modeling

Deyi Li^{a,*,1}, Qi Xu^{a,1}, Lingyao Li^b, Tiansheng Wang^{c,d}, Muxuan Liang^e and Mei Liu^{a,**}

^aDepartment of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA

^bCollege of Information Science, University of Arizona, Tucson, AZ, USA

^cDepartment of Pharmaceutical Health Outcomes and Policy, College of Pharmacy, University of Houston, Houston, TX, USA

^dDepartment of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

^eDepartment of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Keywords:

Electronic health records
Clinical predictive modeling
Neural architecture search
Knowledge transfer
Multi-agent systems

ABSTRACT

Transformer-based models are widely used for clinical prediction from electronic health records (EHRs), yet their architectures still require substantial manual tuning, and the optimal configuration may vary across tasks and hospitals. Neural architecture search (NAS) automates architecture design, but conventional methods are computationally costly for Transformer-based EHR models. Recent large language model (LLM)-guided NAS methods reduce manual search design but typically conduct each search independently, without reusing architecture knowledge across hospitals. In this study, we propose ATHENA (Agentic Transfer across Hospitals for EHR Neural Architecture Search), a knowledge-guided agentic NAS framework for Transformer-based EHR modeling. ATHENA uses a weight-sharing supernet that is pretrained once per hospital, allowing candidate architectures to be instantiated as inherited subnetworks and evaluated through fine-tuning rather than independent pretraining. It also incorporates a two-layer cross-hospital architecture prior. The first layer retrieves high-performing architecture examples from source sites based on task descriptors, while the second estimates the effects of architectural components using SHapley Additive exPlanations (SHAP)-based meta-regression. These priors guide a multi-agent LLM search together with validation feedback from the target hospital. Across six clinical prediction tasks and two independent health systems, ATHENA matches or outperforms four NAS baselines in 9 of 12 hospital–task evaluations at a search budget of 30. It also shows more consistent architecture selection across repeated searches. ATHENA provides a practical approach for reducing manual architecture tuning in Transformer-based EHR modeling. Code is publicly available at <https://github.com/GatorAIM/ATHENA>.

1. Introduction

The increasing availability of large-scale electronic health record (EHR) data has motivated the development of models that can learn predictive patterns directly from routinely collected clinical records [1]. Deep learning has become a standard approach for these tasks, with Transformer-based models increasingly adopted for their ability to capture dependencies among medical events across a patient’s longitudinal record [2]. A common two-stage paradigm consists of pretraining a Transformer backbone on large-scale EHR data to learn generalizable clinical patterns, followed by fine-tuning to adapt the model to task-specific objectives [3–6].

Despite these advances, Transformer architectures for EHR modeling are still largely determined through manual design, relying on prior experience and trial-and-error. The optimal architecture can vary across prediction tasks and health systems, as differences in cohort composition, clinical workflows, and outcome characteristics may favor different architectural configurations [7]. Consequently, a single hand-designed architecture may not perform consistently well across tasks and sites, whereas manually redesigning the

architecture for each new setting is computationally costly and fails to systematically leverage knowledge gained from previous settings.

Neural architecture search (NAS) addresses this problem by treating architecture design as an optimization problem over a predefined search space [8, 9]. Classical NAS approaches, including reinforcement learning [10], evolutionary algorithm (EA) [11], differentiable search [12], and Bayesian optimization [13], have been successfully applied to vision tasks. However, evaluating each candidate architecture can require substantial training. This is particularly problematic for Transformer-based EHR modeling, where pretraining on large longitudinal EHR datasets can dominate the computational cost of model development [3]. Weight-sharing NAS reduces this cost by training a single over-parameterized supernet whose subnetworks share parameters, allowing candidate architectures to be instantiated as inherited subnetworks and evaluated through fine-tuning rather than independently pretrained from scratch [14]. This makes weight-sharing NAS particularly well suited to architecture search for Transformer-based EHR models.

Recent work has also explored using large language models (LLMs) to guide NAS [15–18]. Given descriptions of the task, search space, and computational budget, an LLM can propose new architectural configurations, providing a flexible alternative to manually designed search heuristics. However, existing LLM-guided approaches generally treat

*Corresponding author

**Corresponding author

✉ lideyi@ufl.edu (D. Li); mei.liu@ufl.edu (M. Liu)

ORCID(s): 0009-0004-7039-7303 (D. Li); 0000-0002-8036-2110 (M. Liu)

¹These authors contributed equally to this work.

each search as an independent optimization problem. When architecture search is repeated across clinical tasks or health systems, information from previous searches is not explicitly retained as a reusable prior. Consequently, each new search must rediscover useful architectural patterns through its own evaluations, even when related tasks or sites have already been explored.

In this study, we propose ATHENA (Agentic Transfer across Hospitals for EHR Neural Architecture Search), a knowledge-guided agentic NAS framework for Transformer-based EHR modeling. ATHENA addresses the two limitations identified above through weight-sharing evaluation and a cross-hospital architecture prior. Specifically, we construct an AutoFormer-style supernet [19] that is pretrained once for each hospital and reused throughout architecture search, allowing each candidate architecture to inherit the corresponding subnetwork weights and be evaluated through fine-tuning rather than independent pretraining. ATHENA then combines a multi-agent LLM search with a two-layer prior that transfers architectural knowledge from previously studied hospitals. The first layer retrieves high-performing architectures from source sites based on task descriptors, while the second estimates the effects of architectural components using SHapley Additive exPlanations (SHAP)-based meta-regression. Together, these components allow the search to build on previously observed architectural patterns rather than treating each hospital-task pair as an independent optimization problem.

Across six clinical prediction tasks and two independent health systems, ATHENA achieves up to a 15.3-fold speedup in architecture evaluation compared with conventional independent pretraining and fine-tuning. With a limited search budget, ATHENA matches or outperforms conventional and LLM-guided NAS baselines while exhibiting more consistent architecture selection across repeated searches.

2. Related work

2.1. Transformer-based EHR modeling

The longitudinal and irregular nature of EHR data makes modeling patient trajectories particularly challenging. Transformer architectures [20] have consequently become widely used for EHR modeling because of their ability to capture dependencies across longitudinal clinical events. BEHRT [21] represents medical codes as tokens and models longitudinal patient histories using Transformer encoders, while Med-BERT [3] demonstrates the effectiveness of large-scale pretraining on millions of patient records.

Subsequent work has incorporated richer temporal and clinical information into Transformer-based EHR models. CEHR-BERT [22] introduces temporal-aware embeddings and auxiliary learning objectives, whereas Hi-BEHR [23] employs a hierarchical architecture to capture long-range dependencies in extended patient histories. More recent studies have explored Transformer architectures beyond the conventional encoder-only design. TransformEHR [4] adopts an encoder-decoder architecture with generative pretraining

objectives for longitudinal clinical modeling, while Foresight [24] employs GPT-style autoregressive pretraining to model patient trajectories. CLMBR [25] learns transferable patient representations through autoregressive next-code prediction.

Collectively, these studies demonstrate the value of pretraining and contextualized representation learning for clinical prediction. Despite these advances, the architectures of Transformer-based EHR models remain largely handcrafted, with model configurations typically selected through manual tuning or limited hyperparameter search. This limitation motivates the use of NAS to systematically explore Transformer architectures for downstream clinical prediction tasks.

2.2. NAS for healthcare models

NAS automates architecture design by defining a search space, a search strategy, and a performance-estimation scheme [26]. Recent advances, including gradient-based optimization [12] and weight-sharing supernet methods [14], have substantially reduced the computational cost of architecture search. NAS has also been extended to Transformer architectures through methods such as Evolved Transformer [27], HAT [28], NAS-BERT [29], AutoBERT-Zero [30], and Primer [31]. However, these methods were primarily developed for natural-language and vision tasks and do not explicitly account for the characteristics of longitudinal EHR data.

Within healthcare, NAS has been applied predominantly to medical imaging [18, 32] and multimodal learning [33, 34]. For structured EHR data, existing efforts have focused mainly on modality fusion and multi-task prediction. MUFASA [33] and AutoFM [34] search modality-specific architectures and fusion strategies for integrating longitudinal EHR records with clinical notes, while recent work has also explored NAS for multi-task disease prediction [34]. To our knowledge, existing EHR NAS approaches primarily target fusion or task-sharing structures rather than the internal architecture of Transformer backbones. Consequently, architecture search for Transformer backbones tailored to different clinical prediction tasks remains largely unexplored.

2.3. LLM-driven NAS

Recent work has explored the use of LLMs for NAS, leveraging architectural knowledge encoded during pretraining to generate and refine candidate architectures. Early studies demonstrated that LLMs can serve as architecture generators or search controllers within NAS pipelines. GENIUS [16] employs an LLM as a black-box architecture optimizer that iteratively proposes and refines candidate architectures through natural-language interaction, while EvoPrompting [35] integrates LLM-based mutation and crossover operators into an evolutionary search framework. Subsequent work has further combined LLM reasoning with conventional search strategies. GPT-NAS [36] couples GPT-guided architecture generation with evolutionary optimization, whereas LLMatic [37] integrates LLM-based code generation with quality-diversity search to discover diverse and robust neural architectures.

Applications of LLM-driven NAS to healthcare remain limited and have focused primarily on medical imaging. For example, a recent framework for histopathology diagnosis uses an LLM to iteratively refine the search space of a one-shot NAS pipeline, enabling the discovery of lightweight and transferable models across multiple pathology tasks [18]. Despite these advances, LLM-driven NAS has not been systematically explored for longitudinal EHR modeling. Moreover, existing methods typically treat each architecture search as an independent optimization problem and do not explicitly reuse architectural knowledge from previous searches across related clinical tasks or health systems. ATHENA extends this emerging paradigm by combining multi-agent LLM search with retrieved high-performing architectures and cross-hospital architecture-effect priors for Transformer-based EHR model design.

3. Methods

3.1. Overview

ATHENA is a two-stage framework for knowledge-guided NAS of EHR Transformers (Figure 1). Let $\mathcal{S}$ denote the source hospitals, H a held-out target hospital, and $\mathcal{T}$ the clinical prediction tasks. The target hospital is excluded from all metadata used to construct the cross-hospital prior.

In *Stage 1*, ATHENA pretrains one hospital-specific AutoFormer-style supernet at each $S \in \mathcal{S}$ using masked language modeling (MLM) and evaluates a fixed collection of subnet architectures on the source tasks. Each metadata record has the form

$$(S, t, a, \text{Params}_{S,t}(a), \mathbf{m}_{S,t}^{\text{src}}(a)),$$

where $t \in \mathcal{T}$, a is an architecture, $\text{Params}_{S,t}(a)$ is its parameter count, and $\mathbf{m}_{S,t}^{\text{src}}(a)$ contains its source performance metrics. Pooled source records yield two complementary priors: *Layer 1* retrieves concrete high-performing architectures from a task-matched source hospital, whereas *Layer 2* summarizes cross-hospital associations between architectural choices and performance.

In *Stage 2*, a Proposal Agent generates candidates, a Critic Agent checks and refines them, and an Experiment Agent evaluates accepted subnetworks and chooses whether the next round should emphasize exploration or exploitation. Both prior layers provide soft guidance together with feedback from earlier target-hospital evaluations.

3.2. Problem formulation

Let $\mathcal{C}$ denote the vocabulary of diagnosis, medication, laboratory, and procedure tokens. The chronologically ordered visits of patient p are

$$\mathcal{V}_p = (v_{p,1}, \dots, v_{p,T_p}), \quad v_{p,\tau} = \{c_{p,\tau,1}, \dots, c_{p,\tau,N_{p,\tau}}\},$$

where $c_{p,\tau,i} \in \mathcal{C}$. Visits are chronological, but no within-visit order is assumed because event timestamps within an encounter may not reflect the underlying clinical sequence [3, 38, 39]. For each prediction example, only visits available

by its index time are retained. They are flattened into a token sequence, prepended with [CLS], and annotated with token-type and visit-index embeddings.

For target hospital H and task t , a candidate architecture is

$$a = (d, L, \gamma, h),$$

where d is the embedding dimension, L is the number of encoder layers, γ is the multi-layer perceptron (MLP) expansion ratio, and h is the number of attention heads. Let $\mathcal{D}$, $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{H}$ denote their respective candidate sets. The unfiltered search space is

$$\mathcal{A}_0 = \mathcal{D} \times \mathcal{L} \times \mathcal{R} \times \mathcal{H},$$

in the order (d, L, γ, h) . The target-specific legal set is

$$\mathcal{A}_{H,t} = \left\{ a \in \mathcal{A}_0 : d \bmod h = 0, \text{Params}_{H,t}(a) \leq P_{\max} \right\},$$

where $P_{\max}$ is the parameter-count limit. The divisibility condition ensures an integral attention-head dimension.

During target search, each evaluated architecture produces a validation metric vector over

$$\mathcal{G} = \{\text{Accuracy}, F_1, \text{AUROC}, \text{AUPRC}\}.$$

For binary tasks, F_1 is the positive-class binary F_1 , and AUROC and AUPRC use the positive-class probability. For multilabel tasks, all metrics are macro-averaged over label classes. After b accepted evaluations, let

$$\mathcal{E}_b = ((a_i, \mathbf{m}_{\text{val},t}(a_i)))_{i=1}^b$$

denote the ordered target-validation history. ATHENA ranks each metric in descending performance order, assigns average ranks to ties, and computes

$$R_t(a; \mathcal{E}_b) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \text{rank}_{g,t}^{\downarrow}(a; \mathcal{E}_b), \quad (1)$$

where lower ranks are preferred. If composite ranks tie, the architecture evaluated first is retained, making selection deterministic with respect to the archive order.

Let $b_{\text{end}} \leq B$ be the number of architectures evaluated before budget exhaustion or early termination. ATHENA returns

$$a_{H,t}^* = \arg \min_{a: (a, \mathbf{m}) \in \mathcal{E}_{b_{\text{end}}}} R_t(a; \mathcal{E}_{b_{\text{end}}}).$$

Here B is the maximum evaluation budget.

3.3. Transformer subnet architecture

Given architecture $a = (d, L, \gamma, h)$, token i is represented by the sum of its code, token-type, and visit-index embeddings:

$$\mathbf{h}_i^0 = \mathbf{e}_i^{\text{code}} + \mathbf{e}_i^{\text{type}} + \mathbf{e}_i^{\text{visit}} \in \mathbb{R}^d,$$

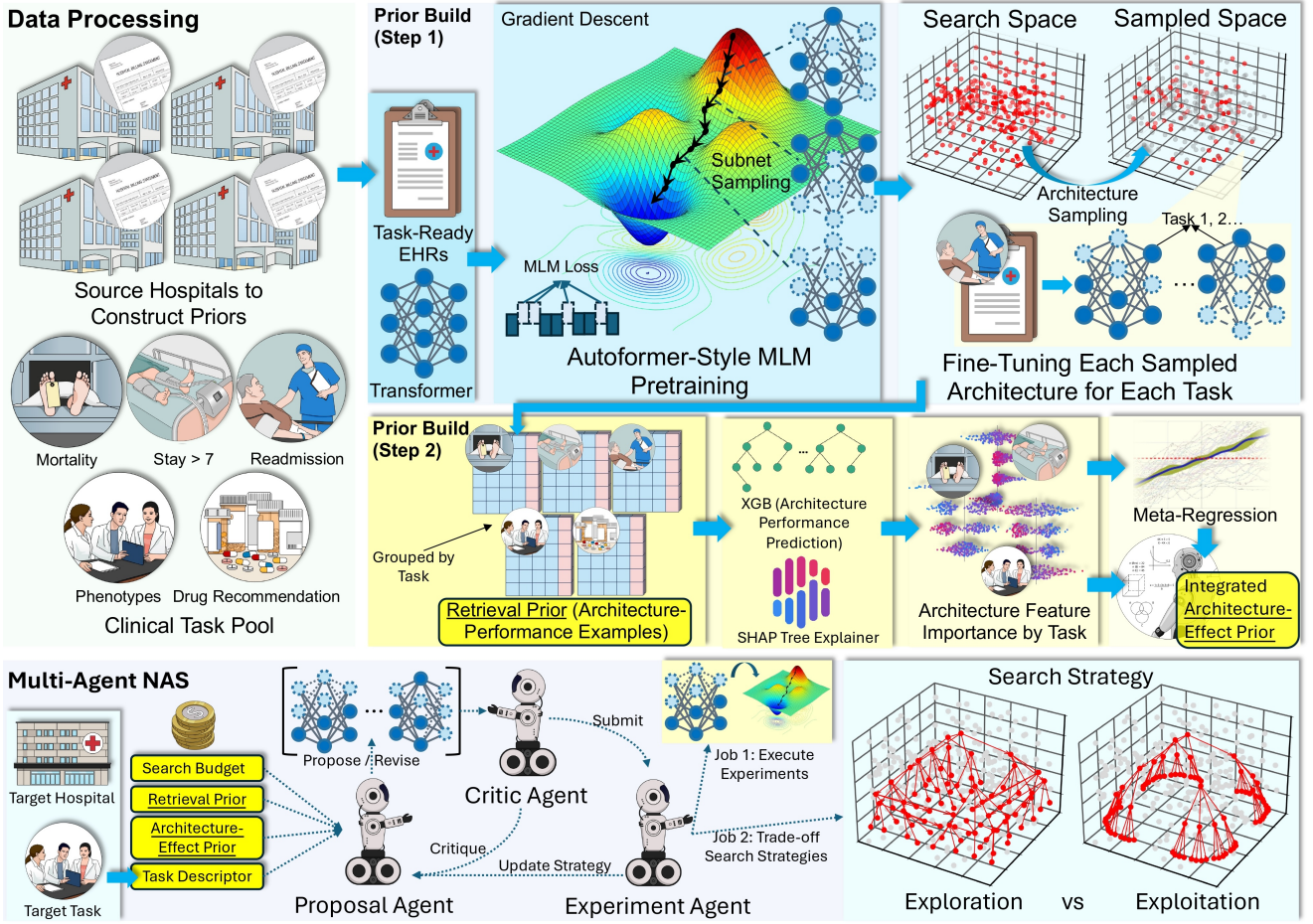


Figure 1: The ATHENA framework. Stage 1 evaluates sampled architectures across source hospitals and tasks to obtain architecture-performance metadata. The metadata yield a retrieval prior of high-performing source architectures and a SHAP-based architecture-effect prior. Stage 2 combines these priors with target-validation feedback to guide agentic NAS while adaptively balancing exploration and exploitation.

where $\mathbf{e}_i^{\text{code}}$, $\mathbf{e}_i^{\text{type}}$, and $\mathbf{e}_i^{\text{visit}}$ denote the corresponding embedding vectors.

The input sequence consists of a prepended [CLS] token followed by the flattened EHR token sequence:

$$\mathbf{H}^0 = [\mathbf{h}_{[\text{CLS}]}^0, \mathbf{h}_1^0, \dots, \mathbf{h}_n^0] \in \mathbb{R}^{(n+1) \times d}.$$

ATHENA uses a pre-normalization Transformer encoder. Multi-head attention (MHA), feed-forward network (FFN), and layer normalization (LN) denote the corresponding Transformer operations below. For layer $\ell = 1, \dots, L$,

$$\begin{aligned} \tilde{\mathbf{H}}^\ell &= \mathbf{H}^{\ell-1} + \text{MHA}_{a,\ell} \left(\text{LN}_{a,\ell}^{\text{attn}} (\mathbf{H}^{\ell-1}) \right), \\ \mathbf{H}^\ell &= \tilde{\mathbf{H}}^\ell + \text{FFN}_{a,\ell} \left(\text{LN}_{a,\ell}^{\text{ffn}} (\tilde{\mathbf{H}}^\ell) \right), \end{aligned}$$

where $\text{MHA}_{a,\ell}$ uses h heads and $\text{FFN}_{a,\ell}$ has hidden width γd . Dropout and drop-path are applied within the residual branches.

For downstream prediction, the hidden representation corresponding to the final [CLS] token is used as the patient-level representation:

$$\mathbf{z}_{p,j,a} = \text{LN}_a \left(\mathbf{h}_{p,j,[\text{CLS}]}^L \right),$$

where $\mathbf{h}_{p,j,[\text{CLS}]}^L$ is the final [CLS] hidden state for target admission j .

3.4. AutoFormer-style supernet training and subnet evaluation

Training every architecture independently would require repeated self-supervised pretraining and downstream fine-tuning. ATHENA therefore uses an AutoFormer-style shared-weight supernet [19] spanning $\mathcal{A}_0$. The supernet takes the componentwise maximal configuration in the search space. A subnet activates the first d embedding channels, the first L encoder blocks, and the first γd FFN units. Its attention module partitions a fixed internal query/key/value representation into h heads, while the input and output projections are sliced to width d . Thus, each candidate architecture is instantiated as a weight-inherited subnetwork of the shared supernet rather

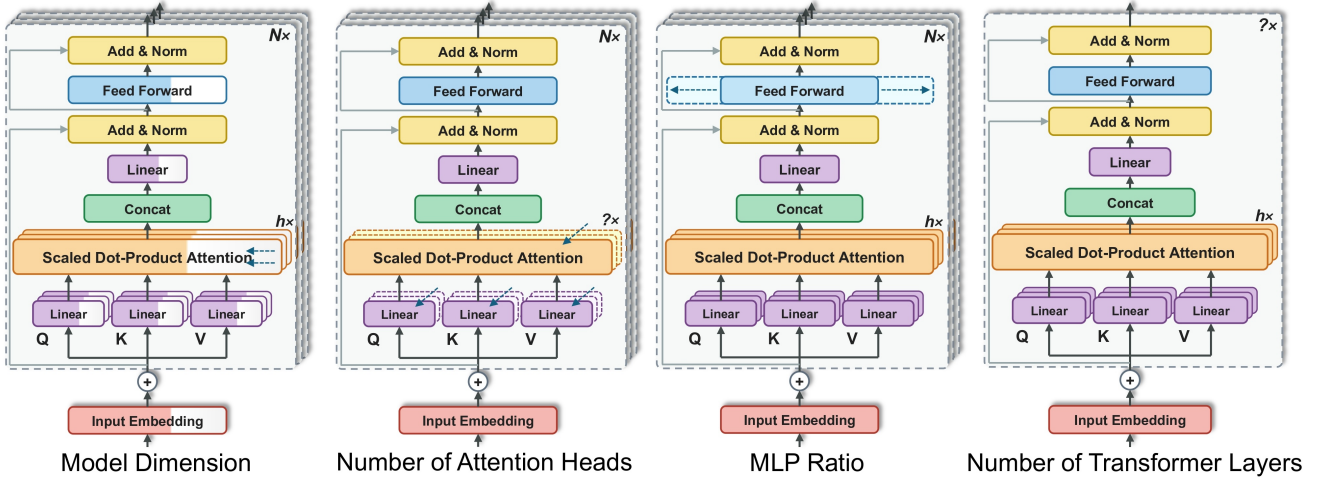


Figure 2: Illustration of the four configurable Transformer architecture parameters in the NAS space.

than as an independently trained model. Figure 2 illustrates how the four configurable parameters alter the resulting subnet.

One supernet is pretrained per hospital and shared by all methods and random seeds at that hospital. Each MLM minibatch samples $a \sim \text{Uniform}(\mathcal{A}_0)$, activates the corresponding parameter slices, and updates only those active weights. During downstream fine-tuning, a is fixed and the same inherited slices are updated at every epoch.

Following BERT-style masking [40], a subset of non-special tokens is selected for MLM and corrupted. Let Ω denote the set of masked positions. The pretraining loss is

$$\mathcal{L}_{\text{MLM}} = - \sum_{i \in \Omega} \log p_{\mathbf{W}_a}(c_i | \tilde{\mathbf{H}}^0),$$

where $\tilde{\mathbf{H}}^0$ is constructed from the corrupted tokens and $\mathbf{W}_a$ denotes the active subnet weights. Selected non-special tokens are replaced by [MASK], replaced by a random token of the same modality, or left unchanged.

Each candidate inherits its encoder weights from this checkpoint, is fine-tuned on the target training split, and is monitored on the target validation split. The resulting validation metrics form one record in the search archive.

3.5. Cross-hospital prior construction

ATHENA uses source-hospital metadata to construct a two-layer prior. *Layer 1* retrieves high-performing architectures from a task-matched source hospital. *Layer 2* pools source hospitals to estimate task-specific architectural preferences and interactions. Both layers are provided to the agents as context rather than imposed as hard constraints.

3.5.1. Task-driven retrieval prior

For task t at hospital H , the task descriptor is

$$\mathbf{g}_{H,t} = \left[\mathbb{I}_{\text{binary}}(t), \mathbb{I}_{\text{multilabel}}(t), \tilde{C}_t, \mathcal{H}_{\text{label}}(H, t), \pi_{H,t}^+, \tilde{\Delta}_t \right],$$

where the first two entries identify the task type, $\tilde{C}_t$ is the normalized number of output classes, $\mathcal{H}_{\text{label}}$ is label entropy,

π^+ is positive-label prevalence, and $\tilde{\Delta}_t$ is the normalized prediction horizon. The data-dependent entries are computed from each hospital's downstream training split; consequently, the target descriptor uses target-training data only.

When exact task t is present at the source hospitals, ATHENA selects

$$S_t^* = \arg \max_{S \in \mathcal{S}} \frac{\mathbf{g}_{H,t}^T \mathbf{g}_{S,t}}{\|\mathbf{g}_{H,t}\|_2 \|\mathbf{g}_{S,t}\|_2}.$$

Here $\|\cdot\|_2$ denotes the Euclidean norm. ATHENA then ranks the legal source architectures for (S_t^*, t) by the composite source-performance rank corresponding to Equation 1 and retrieves the top K . These configurations and their source metrics form the *Layer 1* retrieval set $\mathcal{A}_{\text{ret}}$.

If exact task metadata are unavailable, ATHENA selects the source hospital with the highest cosine similarity between standardized dataset profiles (e.g., sample size, modality counts, and mean encounters per patient). Within the selected hospital, ATHENA chooses the available surrogate task whose descriptor is most similar to $\mathbf{g}_{H,t}$.

3.5.2. Architecture-effect prior

For *Layer 2*, architecture evaluations for task t are pooled across source hospitals. For source record i , ATHENA defines

$$y_{t,i} = -R_t(a_i; \mathcal{E}_{S,t}),$$

where $\mathcal{E}_{S,t}$ is the pooled source metadata table for task t and larger $y_{t,i}$ indicates better performance relative to all source records for that task. Hospital identity is retained for the subsequent mixed-effects analysis. For each task, ATHENA fits the XGBoost surrogate [41]

$$y_{t,i} = f_{\text{XGBoost},t}(a_i) + \varepsilon_{t,i}.$$

The discrete-valued architecture features are supplied to the tree model in their numeric form. SHAP TreeExplainer [42] produces a signed contribution $\phi_{t,i}^{(q)}$ for each architectural feature q and record i .

To separate population-level architectural patterns from hospital variation, ATHENA fits

$$\phi_{i,i}^{(q)} = \mu_{i,\ell_{i,q}}^{(q)} + u_{i,S_i}^{(q)} + \epsilon_{i,i}^{(q)}, \quad u_{i,S}^{(q)} \sim \mathcal{N}(0, \sigma_{u,q}^2),$$

where $\ell_{i,q}$ is the categorical level of q and $u_{i,S}^{(q)}$ is a hospital random intercept. The reported $\mu_{i,\ell}^{(q)}$ values are level-specific mean SHAP contributions reconstructed from the reference-coded model.

Levels whose confidence intervals lie entirely above or below zero are labeled *preferred* or *discouraged*, respectively; the remainder are *inconclusive*. ATHENA applies the same idea to the most influential feature pair to obtain supported interaction rules. Feature importance, level labels, confidence intervals, and interaction rules form the *Layer 2* prior $\mathcal{P}_{\text{meta}}$. They are presented to the agents as non-causal, directional evidence.

3.6. Agentic NAS

ATHENA coordinates a *Proposal Agent* A_p , a *Critic Agent* A_c , and an *Experiment Agent*. The Experiment Agent consists of a deterministic evaluator A_e and an LLM-based strategy module A_s . Proposal, critique, revision, and strategy selection are LLM calls; legality checks, subnet fine-tuning, metric calculation, ranking, and final selection are controller operations.

For each target hospital-task pair (H, t) , ATHENA constructs a fixed context

$$\mathcal{C}_{H,t} = \{\mathbf{g}_{H,t}, \mathcal{A}_{\text{ret}}, \mathcal{P}_{\text{meta}}\},$$

containing the target task descriptor, the K retrieved architectures, and the *Layer 2* prior.

At round r , ATHENA maintains two distinct histories. The *ordered search memory* (or search transcript) $\mathcal{M}^{(r)}$ contains proposals, critiques, revisions, evaluation feedback, and strategy decisions, and is supplied as LLM context. The *structured validation archive*

$$\mathcal{E}^{(r)} = ((a_i, \mathbf{m}_{\text{val},t}(a_i)))_{i=1}^{b_r}$$

contains only evaluated architectures and their validation metrics. It is used by the deterministic controller for ranking and selection. Thus, $\mathcal{M}^{(r)}$ carries accumulated search experience, whereas $\mathcal{E}^{(r)}$ is the authoritative record of empirical evidence. The memory is an ordered transcript, not a separately learned or retrieval-based long-term-memory system.

The round also has remaining budget $B - b_r$, and a strategy

$$s^{(r)} = (z^{(r)}, \eta^{(r)}), \quad z^{(r)} \in \{\text{exploration}, \text{exploitation}\},$$

where $\eta^{(r)}$ is a natural-language rationale. Exploration encourages coverage of under-sampled architectural choices, whereas exploitation focuses proposals around configurations that have performed well on target validation.

Conditioned on $(\mathcal{C}_{H,t}, \mathcal{M}^{(r)}, \mathcal{E}^{(r)}, B - b_r, s^{(r)})$, A_p proposes candidates with rationales. The Critic checks legality and novelty, provides structured feedback, and allows rejected

Proposal Agent Prompt Template

Role: You are an NAS agent for Transformer models applied to longitudinal EHR data. Generate candidate architectures for the target hospital-task pair.

Search space and constraints: Choices for `embed_dim`, `depth`, `mlp_ratio`, and `num_heads`; require `embed_dim % num_heads = 0`; enforce parameter-count constraints; identify infeasible regions.

Task context: Target-task statistics $\mathbf{g}_{H,t}$.

Retrieval prior: Retrieved top- K source architectures $\mathcal{A}_{\text{ret}}$ and their source performance.

Architecture-effect prior: Preferred/discouraged architectural levels, feature-importance rankings, confidence labels, and interaction rules $\mathcal{P}_{\text{meta}}$.

Search state: Ordered transcript $\mathcal{M}_{H,t}^{(r)}$, validation history $\mathcal{E}^{(r)}$, current best architecture, and remaining budget.

Strategy directive: $s^{(r)} = (z^{(r)}, \eta^{(r)})$. Under *exploration*, prioritize diversity and coverage of under-explored regions. Under *exploitation*, refine architectures near the current best region.

Output: JSON array of candidate architectures containing `embed_dim`, `depth`, `mlp_ratio`, `num_heads`, and `rationale`. Return only valid JSON. Test-set metrics are never provided.

Figure 3: The Proposal Agent prompt template. The implementation populates each block with task-specific values before invoking the LLM.

non-duplicate candidates to be revised for at most R passes. Parameter-limit violations and duplicates are hard rejections; disagreement with a *Layer 2* preference is only a soft concern. The proposal–critique–revision records are appended to $\mathcal{M}^{(r)}$ even when no candidate is evaluated. Figure 3 shows the Proposal Agent prompt template.

The evaluator fine-tunes each accepted subnet and appends the resulting records to both $\mathcal{E}^{(r)}$ and $\mathcal{M}^{(r)}$. The controller then recomputes the leader using Equation 1. When budget remains, A_s reads the updated transcript and validation trajectory and selects the next strategy, treating the cross-hospital prior as secondary evidence. The strategy decision is also appended to $\mathcal{M}^{(r)}$.

If a round produces no legal, non-duplicate candidate, no evaluation budget is consumed and a consecutive-failure counter is incremented. The counter resets after any successful evaluation round. Search terminates when the budget B is exhausted or after F_{max} consecutive empty rounds, and ATHENA returns the current validation-selected leader. A run in which no candidate is evaluated is recorded as unsuccessful. Algorithm 1 summarizes this search logic.

4. Experiment configuration

4.1. Data sources and cohort construction

This study uses EHR data from OneFlorida+ [43, 44] and MIMIC-IV [45]. OneFlorida+ is a large clinical research network within PCORnet [46] and contains standardized EHR data across multiple health systems. MIMIC-IV is a public de-identified EHR database from Beth Israel Deaconess Medical Center containing more than 524,000 hospital admissions from over 257,000 patients. Both datasets include longitudinal diagnoses, procedures, medications, and laboratory records used to construct the clinical prediction tasks.

Algorithm 1 ATHENA Agentic NAS Loop

Require: Fixed context $C_{H,t}$, legal set $\mathcal{A}_{H,t}$, budget B , critique-pass limit R , failure limit $F_{\max}$

Ensure: Validation-selected architecture $a_{H,t}^*$, or failure if no candidate is evaluated

```

1: Initialize ordered search memory  $\mathcal{M} \leftarrow ()$ , validation archive  $\mathcal{E} \leftarrow ()$ , strategy  $s \leftarrow \text{exploration}$ , and failure count  $f \leftarrow 0$ 
2: while  $|\mathcal{E}| < B$  and  $f < F_{\max}$  do
3:    $\mathcal{P} \leftarrow A_p(C_{H,t}, \mathcal{M}, \mathcal{E}, B - |\mathcal{E}|, s)$ 
4:    $(\mathcal{U}, \Delta\mathcal{M}) \leftarrow \text{CritiqueAndRevise}(\mathcal{P}, C_{H,t}, \mathcal{M}, \mathcal{E}, s, R)$ 
5:    $\mathcal{M} \leftarrow \mathcal{M} \parallel \Delta\mathcal{M}$ 
6:   Remove illegal and duplicate candidates from  $\mathcal{U}$  in proposal order
7:   Retain at most the first  $B - |\mathcal{E}|$  candidates in  $\mathcal{U}$ 
8:   if  $\mathcal{U} = \emptyset$  then
9:      $f \leftarrow f + 1$ ; continue
10:  end if
11:   $\mathcal{V} \leftarrow A_e(\mathcal{U}^*)$  using target training and validation data
12:   $\mathcal{E} \leftarrow \mathcal{E} \parallel \mathcal{V}$ ;  $\mathcal{M} \leftarrow \mathcal{M} \parallel \mathcal{V}$ ;  $f \leftarrow 0$ 
13:   $a_{\text{best}} \leftarrow \arg \min_{a: (a,m) \in \mathcal{E}} R_i(a; \mathcal{E})$ 
14:  if  $|\mathcal{E}| < B$  then
15:     $s \leftarrow A_s(C_{H,t}, \mathcal{M}, \mathcal{E}, a_{\text{best}}, B - |\mathcal{E}|)$ 
16:     $\mathcal{M} \leftarrow \mathcal{M} \parallel s$ 
17:  end if
18: end while
19: if  $\mathcal{E} = \emptyset$  then
20:   return failure
21: end if
22: return  $a_{H,t}^* \leftarrow a_{\text{best}}$ 

```

For ATHENA, the processed OneFlorida+ data comprise four prior-source sites (Sites A–D) and one internal held-out target (Site E); MIMIC-IV serves as the external held-out target. Records from both target cohorts are excluded from the cross-hospital prior. The OneFlorida+ and MIMIC-IV cohorts follow the same EHR tokenization, task-construction, and split conventions.

Each hospital’s unlabeled EHR corpus is first divided into a pretraining pool of patients without downstream labels and a downstream pool containing the remaining patients. The pretraining pool is further split 90%/10% for masked-language-model supernet pretraining and validation.

We evaluate six downstream clinical prediction tasks: in-hospital mortality (Mortality), Stay > 7d, Readmission (3M), Phenotype (6M), Phenotype (12M), and same-visit Drug Recommendation. The two phenotype tasks are multilabel predictions over 18 commonly benchmarked phenotype classes [47, 48]. Phenotype cohorts include patients with a qualifying next admission. Drug Recommendation uses a separate eligibility cohort consisting of patients with at least one in-vocabulary medication. Cohort inclusion and exclusion criteria and detailed task definitions are provided in Supplementary Method S1 and Supplementary Table S1.

Within the downstream pool, patient-level training, validation, and test splits are defined by task family: 20%/40%/40% for the three binary tasks (Mortality, Stay > 7d, and Readmission (3M)) and 40%/30%/30% for Phenotype (6M) and Phenotype (12M). Drug Recommendation is independently split 40%/30%/30% within its eligibility

cohort. For all tasks, the training set is used for subnet fine-tuning, the validation set guides architecture search and model selection, and the test set is evaluated only once after the search terminates.

4.2. Transformer architecture search space

The NAS search space comprises four configurable Transformer architecture parameters: embedding dimension, depth, number of attention heads, and MLP expansion ratio (Figure 2). Embedding dimension takes values in {32, 64, 128, 256}, while depth, number of attention heads, and MLP expansion ratio each take values in {1, 2, 4, 8}. The Cartesian product of these choices yields $4^4 = 256$ candidate architectures. The same search space is used across all hospitals, tasks, and NAS methods to ensure a consistent comparison. Other implementation details are provided in Supplementary Method S2.

4.3. Compared methods and ablations

We compare ATHENA against representative baselines from both classical and LLM-based NAS:

- **Random Search.** Uniformly samples legal, non-duplicate architectures from $\mathcal{A}_{H,t}$ until the evaluation budget is exhausted.
- **EA.** Following regularized evolution for NAS [11], this baseline maintains a population of candidate architectures and iteratively selects high-performing parents through tournament selection, generates new candidates via mutation, and removes the oldest individuals from the population.
- **GENIUS** [16]. A single-agent LLM-based NAS framework that iteratively proposes candidate architectures based on search history and validation feedback, without the multi-agent collaboration used in ATHENA.
- **CoLLM-NAS** [17]. A collaborative LLM-based NAS framework that uses a stateful Navigator to iteratively refine search strategies from evaluation feedback and historical trajectories, and a stateless Generator to translate these strategies into candidate architectures. A Coordinator manages their interaction, validates generated architectures, and maintains the search archive. Unlike ATHENA, CoLLM-NAS does not leverage cross-site architecture priors.

ATHENA incorporates a two-layer cross-hospital prior consisting of the *Layer 1* retrieval prior and the *Layer 2* architecture-effect prior. To quantify the contribution of each component, we evaluate the following ablated variants:

- **L1-only.** Retains the retrieval-based cross-hospital prior while removing the architecture-effect prior, quantifying the contribution of *Layer 2*.
- **Leave-One-Task-Out (LOTO) Retrieval.** Excludes task-matched source records during retrieval, forcing knowledge transfer from related but non-identical tasks

and evaluating the robustness of retrieval-based priors when exact precedents are unavailable.

- **Cold Start.** Removes both the retrieval prior and the architecture-effect prior, evaluating the effectiveness of the agentic NAS framework in the absence of cross-hospital knowledge.

5. Results

5.1. Study cohorts

The cross-hospital prior was constructed from four de-identified OneFlorida+ source sites (Sites A–D). Their pretraining pools comprised 125,819 patients, representing 49.3% of patients at the four source sites (Table 1). The target cohorts were excluded from prior construction. The pretraining pools comprised 54,410 patients at Site E (64.4% of the cohort) and 42,494 patients in MIMIC-IV (70.0% of the cohort). The source and target cohorts differed substantially in case mix and data density (Supplementary Table S2). Across the six cohorts, Mortality ranged from 0.4% to 4.2%, Stay > 7d from 9.6% to 34.0%, and Readmission (3M) from 12.7% to 23.9%. Median length of stay ranged from 2 to 5 days, and admissions per patient ranged from 1.4 to 2.7. Diagnoses per admission ranged from 5.4 to 19.1 and medications per admission from 3.2 to 13.1. Comorbidity prevalence was also heterogeneous: chronic kidney disease ranged from 2.8% to 16.0% of admissions and coronary atherosclerosis from 1.1% to 23.3%. Together, these differences provided heterogeneous settings in which to evaluate transfer of architecture knowledge.

5.2. Supernet ranking fidelity and search efficiency

Before using supernet-based evaluations to guide architecture search, we first assessed whether weight-inherited subnetworks preserved the relative performance ranking of independently trained architectures. For each MIMIC-IV task, we sampled 150 valid Transformer architectures (58.59% of the search space) and evaluated each under two settings: fine-tuning with weights inherited from the pretrained supernet and independent pretraining followed by fine-tuning. We then computed Spearman rank correlations between the resulting architecture rankings based on AUROC and AUPRC. This analysis examined whether the computationally cheaper supernet-based approach could serve as a reliable proxy for conventional pretrain-then-fine-tune evaluation.

The AutoFormer-style weight-sharing supernet preserved the relative performance of conventionally pretrained and fine-tuned Transformer architectures (Table 2). Spearman correlations were positive for every task and metric. Ranking fidelity was particularly high for the three multilabel tasks: correlations ranged from 0.940 to 0.951 for Phenotype (6M) and Phenotype (12M), and from 0.944 to 0.958 for Drug Recommendation. Stay > 7d also showed strong agreement ($\rho = 0.809$ for AUROC and $\rho = 0.806$ for AUPRC). Agreement was more moderate for Mortality and Readmission (3M), with the lowest correlation observed for Readmission (3M) AUPRC ($\rho = 0.412$). Thus, although fidelity varied

Table 1

Pretraining-pool sizes and patient-level fine-tuning splits. OneFlorida+ contributor sites are de-identified as Site A–E. *Share* denotes the fraction of cohort patients in the pretraining pool. Fine-tuning counts are shown in separate training, validation, and test columns. *Binary* denotes the common split for Mortality, Stay > 7d, and Readmission (3M). Phenotype and Drug Recommendation are abbreviated as *Pheno.* and *Drug Rec.*, respectively. [†]Drug Recommendation uses a separately defined drug-eligible cohort (≥ 1 in-vocabulary medication) and an independent patient-level split; its counts may therefore exceed the pretraining-pool patient count in the *Patients* column.

		Pretraining		Fine-tuning (Binary)		
Cohort		Patients	Share	Train	Val.	Test
Prior source pool						
Site A		22,186	42.0%	6,158	12,234	12,253
Site B		29,222	52.4%	5,300	10,566	10,647
Site C		37,152	42.0%	10,134	20,594	20,647
Site D		37,259	63.9%	4,204	8,409	8,410
Total		125,819	49.3%	25,796	51,803	51,957
Target cohorts						
Site E (internal)		54,410	64.4%	6,013	12,027	12,029
MIMIC-IV (external)		42,494	70.0%	3,642	7,284	7,286
Cohort	Task			Train	Val.	Test
Prior source pool						
Site A	Pheno. 6M		11,083	8,147	8,405	
	Pheno. 12M		12,235	9,124	9,100	
	Drug Rec. [†]		78,999	59,249	59,251	
Site B	Pheno. 6M		9,229	6,892	6,921	
	Pheno. 12M		10,285	7,792	7,643	
	Drug Rec. [†]		73,660	55,245	55,246	
Site C	Pheno. 6M		18,257	13,922	13,699	
	Pheno. 12M		20,082	14,953	14,919	
	Drug Rec. [†]		92,600	69,450	69,450	
Site D	Pheno. 6M		7,103	5,327	5,328	
	Pheno. 12M		7,988	5,991	5,991	
	Drug Rec. [†]		22,958	17,218	17,219	
Total	Pheno. 6M		45,672	34,288	34,353	
	Pheno. 12M		50,590	37,860	37,653	
	Drug Rec. [†]		268,217	201,162	201,166	
Target cohorts						
Site E	Pheno. 6M		9,195	6,896	6,897	
	Pheno. 12M		10,382	7,786	7,787	
	Drug Rec. [†]		33,380	25,035	25,036	
MIMIC-IV	Pheno. 6M		3,866	2,900	2,901	
	Pheno. 12M		4,478	3,359	3,360	
	Drug Rec. [†]		24,017	18,012	18,014	

by task, the supernet provided an informative proxy ranking across the complete search space.

We next measured the GPU time required to evaluate increasing numbers of candidate architectures on one NVIDIA L4. The weight-sharing calculation included one supernet pretraining run followed by subnet fine-tuning for each candidate, whereas conventional search repeated pretraining and fine-tuning independently for every candidate. Weight

Table 2

Supernet ranking fidelity on MIMIC-IV. For each task, 150 architectures are sampled and evaluated both as weight-inherited subnetworks of the pretrained supernet and through independent pretraining followed by fine-tuning. ρ denotes the Spearman rank correlation between the resulting architecture rankings.

Task	ρ (AUROC)	ρ (AUPRC)
Mortality	0.544	0.567
Stay > 7d	0.809	0.806
Readmission (3M)	0.570	0.412
Phenotype (6M)	0.951	0.944
Phenotype (12M)	0.940	0.944
Drug Recommendation	0.944	0.958

Table 3

Search compute under weight sharing, in GPU-minutes on one NVIDIA L4. The shared-weight approach pretrains the supernet *once* and fine-tunes each of N candidates on it ($T_{\text{pre}} + N T_{\text{ft}}$); conventional from-scratch NAS incurs $N(T_{\text{pre}} + T_{\text{ft}})$. Measured times are OneFlorida+ $T_{\text{pre}}=30$ and $T_{\text{ft}}=2$, and MIMIC-IV $T_{\text{pre}}=29$ and $T_{\text{ft}}=1$. All compared NAS methods share one supernet and therefore have the same per-candidate evaluation cost. Values at the maximum budget of 30 evaluations are shown in **bold**.

N	OneFlorida+ held-out			MIMIC-IV		
	Shared	Scratch	Speedup	Shared	Scratch	Speedup
5	40	160	4.0×	34	150	4.4×
10	50	320	6.4×	39	300	7.7×
20	70	640	9.1×	49	600	12.2×
30	90	960	10.7×	59	900	15.3×

sharing yielded larger reductions in search cost as more architectures were evaluated (Table 3). At a budget of five evaluations, the measured speedup over conventional pretrain-then-fine-tune evaluation was 4.0× on OneFlorida+ and 4.4× on MIMIC-IV. At 30 evaluations, the corresponding speedups increased to 10.7× and 15.3×, reducing the estimated cost from 960 to 90 GPU-minutes on OneFlorida+ and from 900 to 59 GPU-minutes on MIMIC-IV.

5.3. Task-specific architecture-effect prior

The cross-hospital architecture-effect prior captured both shared and task-specific Transformer design preferences across the six clinical tasks (Supplementary Figure S1). The most consistent signal was model width. Embedding dimensions of 128 and 256 were reliably preferred for Mortality, while dimensions of 64, 128, and 256 were preferred for the other five tasks. Increasing the number of attention heads was also generally favorable: configurations with 2, 4, or 8 heads had positive effects for Mortality, Stay > 7d, Readmission (3M), Phenotype (6M), and Phenotype (12M). For Drug Recommendation, 4 and 8 heads had positive effects relative to the one-head reference level, whereas the effect of 2 heads was inconclusive.

Depth showed a more heterogeneous pattern. Mortality and Stay > 7d favored depths of 2, 4, and 8 over the single-layer reference configuration. Readmission (3M) did not show the same trend: depths of 2 and 8 were discouraged, while the effect of a depth of 4 was inconclusive. Among the multilabel tasks, Phenotype (6M) favored a depth of 2 but discouraged depths of 4 and 8; Phenotype (12M) and Drug Recommendation discouraged all depths greater than one. The MLP expansion ratio showed a similarly conservative pattern. Ratios greater than one were consistently discouraged for Mortality, Stay > 7d, Readmission (3M), and Phenotype (6M). For Phenotype (12M), ratios of 4 and 8 were discouraged and a ratio of 2 was inconclusive; for Drug Recommendation, a ratio of 4 was discouraged while ratios of 2 and 8 were inconclusive. Thus, the prior identified wider embeddings and multi-head attention as relatively transferable signals, while retaining task-dependent guidance for depth and MLP expansion ratio rather than prescribing one architecture for all outcomes.

Pairwise rules derived from each task's two most influential features were also supplied to the search prompt (Supplementary Table S3). For five of the six tasks, the direction of the identified preference remained unchanged across all levels of the second feature. Mortality was the exception, where the preference for an embedding dimension of 256 emerged when the number of attention heads was at least two.

5.4. Predictive performance under different search budgets

ATHENA achieved the best overall test AUPRC ranking across all search budgets (Table 4). With only five architecture evaluations, it ranked first in 10 of 12 hospital-task comparisons and second in the remaining two, yielding an average rank of 1.17. It ranked first across all six OneFlorida+ tasks and four of six MIMIC-IV tasks, including all multilabel tasks at both targets. The largest gains over the strongest baseline were observed for OneFlorida+ Phenotype (12M), OneFlorida+ Phenotype (6M), and MIMIC-IV Drug Recommendation.

This advantage persisted at larger budgets. At both 20 and 30 evaluations, ATHENA ranked first in 9 of 12 comparisons, with average ranks of 1.29 and 1.42, respectively, compared with 3.17 and 2.50 for the next-best baseline. At budget 30, the AUROC analysis (Supplementary Table S4) showed a similar pattern, with ATHENA ranking first in 7 of 12 comparisons (including one tie with EA) and achieving the best average rank (1.71). Overall, the advantage was consistent across both target health systems.

Validation trajectories further showed that ATHENA's gains generally emerged within the first few evaluations rather than only after most of the search budget had been consumed (Figure 4; Supplementary Figure S2). This pattern was most evident for Phenotype (6M), Phenotype (12M), and Drug Recommendation, for which ATHENA rapidly attained strong validation AUPRC and remained leading or competitive through most of the search. The margins were

Table 4

Test AUPRC across target hospitals and prediction tasks under increasing NAS budgets of 5, 20, and 30 architecture evaluations (mean $\pm$ standard deviation over five random seeds). For each row, the best mean is in **bold** and the second-best mean is underlined; tied values share the mark. *Avg. Rank* is the mean within-row rank over the 12 hospital-task comparisons in each panel (1 = best; ties receive the average rank), with the number of best-in-row comparisons in parentheses. All methods search the same pretrained supernet and therefore differ only in search strategy.

Target	Task	Random Search	EA	GENIUS	CoLLM-NAS	ATHENA
<i>(a) Search budget = 5 evaluations</i>						
OneFlorida+	Mortality	57.78 $\pm$ 0.90	57.78 $\pm$ 0.90	58.04 $\pm$ 0.34	57.72 $\pm$ 1.36	58.23 $\pm$ 1.72
OneFlorida+	Stay > 7d	73.25 $\pm$ 0.33	73.25 $\pm$ 0.33	<u>74.00 $\pm$ 0.34</u>	73.59 $\pm$ 0.45	74.03 $\pm$ 0.37
OneFlorida+	Readmission (3M)	52.09 $\pm$ 0.42	52.09 $\pm$ 0.42	52.06 $\pm$ 0.58	<u>52.16 $\pm$ 0.40</u>	52.26 $\pm$ 0.27
OneFlorida+	Phenotype (6M)	<u>29.65 $\pm$ 0.59</u>	29.53 $\pm$ 0.40	29.28 $\pm$ 0.87	29.29 $\pm$ 1.50	30.42 $\pm$ 0.51
OneFlorida+	Phenotype (12M)	<u>30.86 $\pm$ 0.71</u>	30.74 $\pm$ 0.69	30.39 $\pm$ 0.57	30.23 $\pm$ 0.60	32.17 $\pm$ 0.42
OneFlorida+	Drug Recommendation	19.09 $\pm$ 0.24	19.09 $\pm$ 0.24	19.11 $\pm$ 0.27	<u>19.16 $\pm$ 0.22</u>	19.32 $\pm$ 0.08
MIMIC-IV	Mortality	79.25 $\pm$ 0.64	79.25 $\pm$ 0.64	79.29 $\pm$ 0.62	79.52 $\pm$ 0.43	79.31 $\pm$ 0.40
MIMIC-IV	Stay > 7d	74.67 $\pm$ 0.41	74.80 $\pm$ 0.59	<u>75.22 $\pm$ 0.66</u>	75.02 $\pm$ 0.60	75.66 $\pm$ 0.71
MIMIC-IV	Readmission (3M)	43.64 $\pm$ 0.47	43.64 $\pm$ 0.47	44.29 $\pm$ 0.26	44.55 $\pm$ 0.55	<u>44.47 $\pm$ 0.44</u>
MIMIC-IV	Phenotype (6M)	48.96 $\pm$ 0.76	<u>48.97 $\pm$ 0.76</u>	48.19 $\pm$ 0.66	48.63 $\pm$ 0.67	49.25 $\pm$ 0.57
MIMIC-IV	Phenotype (12M)	48.64 $\pm$ 0.41	48.63 $\pm$ 0.39	47.98 $\pm$ 0.41	<u>48.79 $\pm$ 0.64</u>	49.26 $\pm$ 0.09
MIMIC-IV	Drug Recommendation	<u>20.63 $\pm$ 0.43</u>	<u>20.63 $\pm$ 0.43</u>	20.36 $\pm$ 0.36	20.24 $\pm$ 0.12	21.16 $\pm$ 0.17
<i>Avg. Rank (best-in-row)</i>		3.54 (0)	3.62 (0)	3.58 (0)	<u>3.08 (2)</u>	1.17 (10)
<i>(b) Search budget = 20 evaluations</i>						
OneFlorida+	Mortality	57.46 $\pm$ 0.50	58.10 $\pm$ 0.48	<u>58.18 $\pm$ 0.61</u>	57.88 $\pm$ 0.92	58.85 $\pm$ 0.84
OneFlorida+	Stay > 7d	73.28 $\pm$ 0.52	73.42 $\pm$ 0.23	73.60 $\pm$ 0.73	<u>73.65 $\pm$ 0.51</u>	73.90 $\pm$ 0.28
OneFlorida+	Readmission (3M)	52.21 $\pm$ 0.22	52.34 $\pm$ 0.18	52.40 $\pm$ 0.16	52.49 $\pm$ 0.12	52.44 $\pm$ 0.28
OneFlorida+	Phenotype (6M)	<u>30.57 $\pm$ 0.20</u>	30.23 $\pm$ 1.00	30.02 $\pm$ 0.52	29.46 $\pm$ 1.57	31.06 $\pm$ 0.55
OneFlorida+	Phenotype (12M)	<u>31.53 $\pm$ 0.66</u>	<u>32.14 $\pm$ 0.85</u>	31.37 $\pm$ 0.76	30.54 $\pm$ 1.06	32.62 $\pm$ 0.29
OneFlorida+	Drug Recommendation	19.28 $\pm$ 0.12	19.27 $\pm$ 0.21	19.44 $\pm$ 0.06	19.29 $\pm$ 0.19	<u>19.32 $\pm$ 0.04</u>
MIMIC-IV	Mortality	79.16 $\pm$ 0.41	79.46 $\pm$ 0.24	79.21 $\pm$ 0.45	<u>79.48 $\pm$ 0.63</u>	79.75 $\pm$ 0.37
MIMIC-IV	Stay > 7d	75.18 $\pm$ 0.64	75.67 $\pm$ 0.26	75.39 $\pm$ 0.45	75.20 $\pm$ 0.58	<u>75.64 $\pm$ 0.48</u>
MIMIC-IV	Readmission (3M)	44.25 $\pm$ 0.35	<u>44.88 $\pm$ 0.97</u>	44.61 $\pm$ 0.46	44.94 $\pm$ 0.57	44.94 $\pm$ 0.55
MIMIC-IV	Phenotype (6M)	<u>49.43 $\pm$ 0.53</u>	49.22 $\pm$ 0.69	49.26 $\pm$ 0.69	49.01 $\pm$ 0.73	49.84 $\pm$ 0.35
MIMIC-IV	Phenotype (12M)	<u>49.10 $\pm$ 0.16</u>	49.04 $\pm$ 0.28	49.05 $\pm$ 0.24	48.66 $\pm$ 0.59	49.42 $\pm$ 0.20
MIMIC-IV	Drug Recommendation	<u>21.08 $\pm$ 0.09</u>	20.98 $\pm$ 0.33	20.75 $\pm$ 0.35	20.39 $\pm$ 0.32	21.15 $\pm$ 0.13
<i>Avg. Rank (best-in-row)</i>		3.75 (0)	3.25 (1)	<u>3.17 (1)</u>	3.54 (2)	1.29 (9)
<i>(c) Search budget = 30 evaluations</i>						
OneFlorida+	Mortality	57.91 $\pm$ 0.40	58.07 $\pm$ 0.49	<u>58.18 $\pm$ 0.61</u>	57.88 $\pm$ 0.92	58.85 $\pm$ 0.84
OneFlorida+	Stay > 7d	73.52 $\pm$ 0.78	<u>73.97 $\pm$ 0.40</u>	74.10 $\pm$ 0.13	73.65 $\pm$ 0.51	73.90 $\pm$ 0.27
OneFlorida+	Readmission (3M)	52.27 $\pm$ 0.24	52.41 $\pm$ 0.20	<u>52.54 $\pm$ 0.23</u>	52.49 $\pm$ 0.12	52.56 $\pm$ 0.24
OneFlorida+	Phenotype (6M)	30.57 $\pm$ 0.20	<u>31.01 $\pm$ 0.30</u>	30.02 $\pm$ 0.52	29.46 $\pm$ 1.57	31.06 $\pm$ 0.55
OneFlorida+	Phenotype (12M)	31.71 $\pm$ 0.49	<u>32.34 $\pm$ 0.75</u>	31.44 $\pm$ 0.71	30.54 $\pm$ 1.06	32.54 $\pm$ 0.40
OneFlorida+	Drug Recommendation	19.32 $\pm$ 0.11	<u>19.43 $\pm$ 0.08</u>	19.44 $\pm$ 0.06	19.29 $\pm$ 0.19	19.34 $\pm$ 0.07
MIMIC-IV	Mortality	79.16 $\pm$ 0.42	79.30 $\pm$ 0.54	79.12 $\pm$ 0.57	<u>79.48 $\pm$ 0.63</u>	79.75 $\pm$ 0.37
MIMIC-IV	Stay > 7d	75.01 $\pm$ 0.50	<u>75.69 $\pm$ 0.41</u>	75.56 $\pm$ 0.42	75.20 $\pm$ 0.58	75.88 $\pm$ 0.43
MIMIC-IV	Readmission (3M)	44.76 $\pm$ 0.16	44.78 $\pm$ 1.02	44.79 $\pm$ 0.33	<u>44.94 $\pm$ 0.57</u>	45.13 $\pm$ 0.53
MIMIC-IV	Phenotype (6M)	<u>49.61 $\pm$ 0.51</u>	49.56 $\pm$ 0.48	49.22 $\pm$ 0.68	49.01 $\pm$ 0.73	49.84 $\pm$ 0.35
MIMIC-IV	Phenotype (12M)	<u>49.11 $\pm$ 0.14</u>	<u>49.14 $\pm$ 0.29</u>	49.05 $\pm$ 0.24	48.66 $\pm$ 0.59	49.42 $\pm$ 0.20
MIMIC-IV	Drug Recommendation	21.03 $\pm$ 0.12	21.15 $\pm$ 0.06	20.78 $\pm$ 0.39	20.39 $\pm$ 0.32	<u>21.08 $\pm$ 0.08</u>
<i>Avg. Rank (best-in-row)</i>		3.83 (0)	<u>2.50 (1)</u>	3.08 (2)	4.17 (0)	1.42 (9)

Table 5

Search behavior at a budget of 30 evaluations, averaged over the six tasks and five random seeds. *Params* is the size of the selected architecture; *Evals* is the number of distinct architectures actually evaluated.

Method	OneFlorida+		MIMIC-IV	
	Params (M)	Evals	Params (M)	Evals
Random Search	1.28	30.0	2.34	30.0
EA	1.33	30.0	2.11	30.0
GENIUS	1.90	23.6	2.42	20.3
CoLLM-NAS	1.28	10.7	1.93	9.1
ATHENA	1.55	22.5	2.58	19.6

most persistent for Phenotype (12M) on both targets, whereas competing methods narrowed or occasionally closed the gap on the other tasks as the budget increased. Mortality, Stay > 7d, and Readmission (3M) showed closer convergence among methods. Some LLM-based searches terminated before 30 evaluations when they could no longer generate new legal, non-duplicate candidates. Overall, ATHENA retained most of its early gains, with smaller and task-dependent improvements from later evaluations, indicating that much of its advantage was established under a relatively small search budget.

5.5. Performance–complexity trade-offs and selection behavior

The validation Pareto analyses showed that performance was not a monotonic function of model size (Figure 5; Supplementary Figure S3). In the binary tasks, relatively small architectures could be competitive with substantially larger models, while architectures with similar parameter counts often had visibly different AUPRC values. The multilabel tasks exhibited stepwise improvements across parameter bands, but still showed considerable within-band variation. ATHENA evaluated candidates across these bands and repeatedly placed architectures near the empirical Pareto frontier rather than concentrating exclusively at the largest end of the search space. This pattern was present on both the internal and external targets.

ATHENA’s predictive performance was not explained by systematically selecting the largest models or by using more candidate evaluations (Table 5). On OneFlorida+, the architectures selected by ATHENA averaged 1.55 million parameters, compared with 1.90 million for GENIUS, despite ATHENA’s substantially better average performance rank. ATHENA evaluated an average of 22.5 architectures on OneFlorida+ and 19.6 on MIMIC-IV, below the maximum budget of 30 and slightly fewer than GENIUS on both targets.

Architecture selection was also more reproducible across repeated searches (Table 6). For each task t , we calculated the modal architecture selection rate as the proportion of the five random seeds selecting the most frequently chosen

architecture, and averaged this rate across the six tasks:

$$\text{Mean modal selection (\%)} = \frac{1}{6} \sum_{t=1}^6 100 \times \frac{\max_a(n_{t,a})}{5},$$

where $n_{t,a}$ is the number of random seeds selecting architecture a for task t . On OneFlorida+, ATHENA and EA achieved the highest mean modal architecture selection rate of 40.0%, compared with 26.7% for Random Search and 33.3% for both GENIUS and CoLLM-NAS. On MIMIC-IV, ATHENA achieved the highest rate at 43.3%, compared with 20.0%–33.3% for the baselines. Higher rates indicate greater convergence toward the same architecture across repeated searches and reflect selection consistency rather than architecture quality.

5.6. Robustness of the cross-hospital prior

The ablation results separated the contributions of retrieval and architecture-effect guidance (Figure 6; Supplementary Figure S4). Removing *Layer 2 (L1-only)* generally retained competitive performance but reduced AUPRC on several tasks, particularly Phenotype (6M) and Phenotype (12M) on OneFlorida+ and Mortality and Readmission (3M) on MIMIC-IV.

Removing both prior layers (Cold Start) produced the clearest degradation for Phenotype (6M) and Phenotype (12M) on both targets.

When exact task-matched retrieval records were withheld, LOTO Retrieval remained comparable to exact retrieval and was occasionally higher, showing that related-task source records could provide useful fallback guidance.

6. Discussion

In this study, we developed ATHENA, a knowledge-guided agentic NAS framework that combines weight-sharing evaluation with reusable cross-hospital architecture knowledge for Transformer-based EHR modeling. Across six clinical prediction tasks and two target health systems, ATHENA achieved the strongest overall AUPRC ranking across search budgets while requiring substantially less computation than independently pretraining each candidate architecture. ATHENA also produced more consistent architecture selections across repeated searches, and its performance gains were not explained simply by evaluating more candidates or selecting larger models. These findings suggest that architecture-search experience accumulated from prior hospital–task settings can be reused to make model development more efficient and reproducible when adapting EHR models across clinical prediction settings.

Our findings also highlight the complementary roles of the two cross-hospital priors. The retrieval prior provides architecture-level examples from related source settings, whereas the architecture-effect prior summarizes task-specific effects of individual design choices across hospitals. The ablation results support this distinction: removing the architecture-effect prior reduced performance on several tasks, while removing both priors produced larger declines, particularly

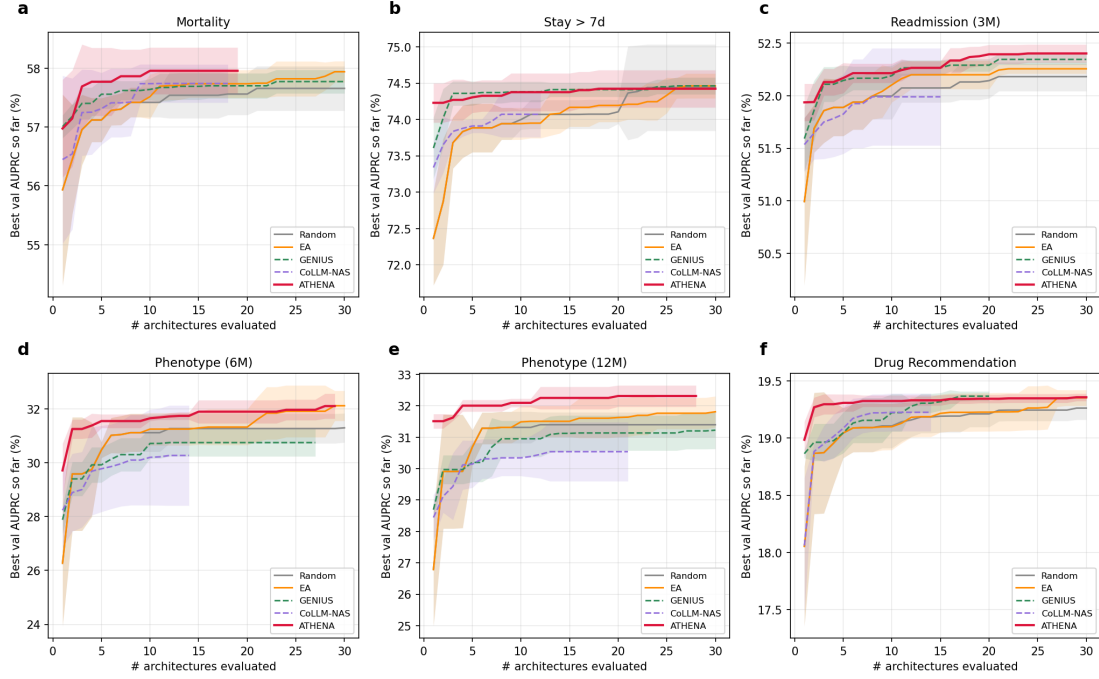


Figure 4: Search trajectories on the held-out OneFlorida+ target for (a) Mortality, (b) Stay > 7d, (c) Readmission (3M), (d) Phenotype (6M), (e) Phenotype (12M), and (f) Drug Recommendation. Lines and shaded bands show the mean and SD, respectively, of the best validation AUPRC attained across five random seeds as a function of the number of evaluated architectures. Curves may end before the full budget when proposal saturation yields no new legal architecture.

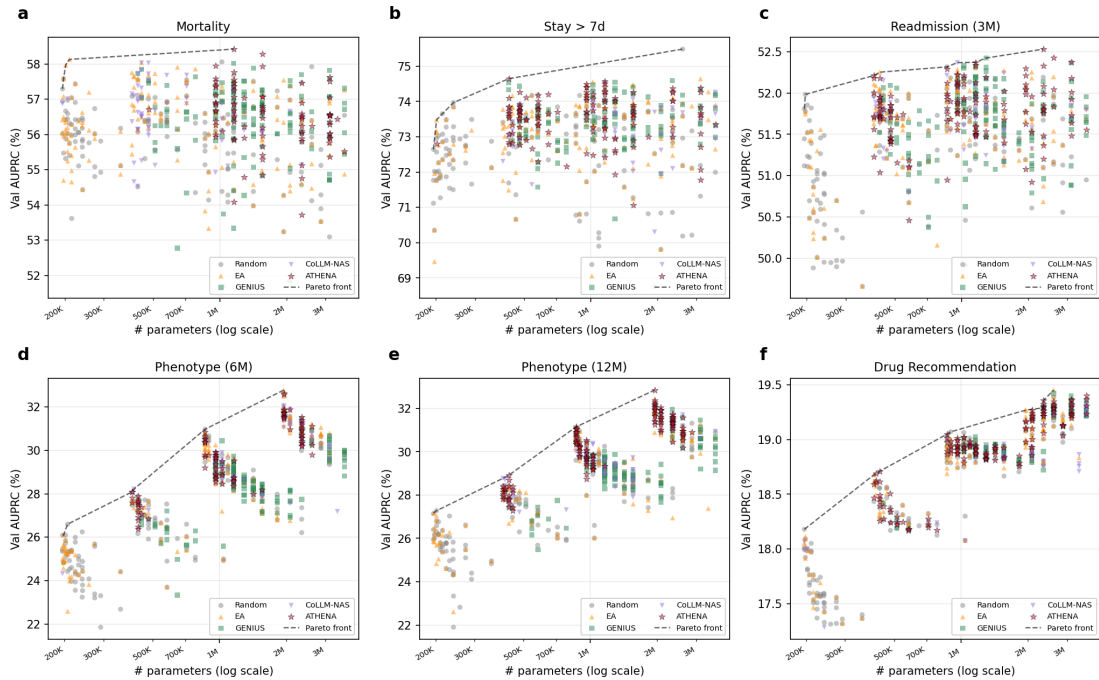


Figure 5: Validation AUPRC versus parameter count on the held-out OneFlorida+ target. Markers identify the search method and the dashed line denotes the empirical Pareto frontier across evaluated architectures.

Table 6

Modal architecture selected by each method, reported as embed_dim/depth/heads/mlp_ratio, with the number of random seeds selecting that architecture (out of five) in parentheses. For task t , let $n_{t,a}$ denote the number of random seeds selecting architecture a ; the task-level modal selection rate is $100 \times \max_a(n_{t,a})/5$. *Mean modal architecture selection (%)* is the average of these rates across the six tasks, i.e. $\frac{1}{6} \sum_{t=1}^6 100 \times \max_a(n_{t,a})/5$. A dash (—) indicates that all five random seeds selected distinct architectures, corresponding to a task-level rate of 20%. Higher values indicate greater convergence toward the same architecture across random seeds and reflect selection consistency, not architecture quality.

Target	Task	Random Search	EA	GENIUS	CoLLM-NAS	ATHENA
OneFlorida+	Mortality	—	64/2/2/2 (2)	256/2/8/1 (2)	—	—
OneFlorida+	Stay > 7d	—	—	—	—	128/4/4/1 (2)
OneFlorida+	Readmission (3M)	128/1/8/1 (2)	—	—	128/4/4/2 (2)	128/2/8/2 (2)
OneFlorida+	Phenotype (6M)	—	256/1/8/2 (2)	256/2/4/1 (2)	—	256/1/4/1 (2)
OneFlorida+	Phenotype (12M)	—	256/1/8/1 (4)	—	128/2/4/2 (2)	256/1/8/1 (4)
OneFlorida+	Drug Recommendation	256/2/8/2 (2)	256/2/4/2 (2)	256/2/8/4 (3)	256/4/8/2 (3)	—
<i>Mean modal architecture selection (%)</i>		26.7	40.0	33.3	33.3	40.0
MIMIC-IV	Mortality	—	—	—	—	256/4/8/1 (3)
MIMIC-IV	Stay > 7d	—	—	256/2/4/2 (2)	—	256/4/8/1 (3)
MIMIC-IV	Readmission (3M)	—	—	—	128/4/4/4 (2)	—
MIMIC-IV	Phenotype (6M)	—	256/1/8/1 (2)	256/2/4/2 (2)	128/4/2/2 (2)	256/1/8/1 (2)
MIMIC-IV	Phenotype (12M)	—	—	256/2/8/4 (2)	—	256/1/4/1 (2)
MIMIC-IV	Drug Recommendation	—	256/1/8/1 (2)	256/2/8/4 (2)	128/4/8/8 (2)	256/1/4/2 (2)
<i>Mean modal architecture selection (%)</i>		20.0	26.7	33.3	30.0	43.3

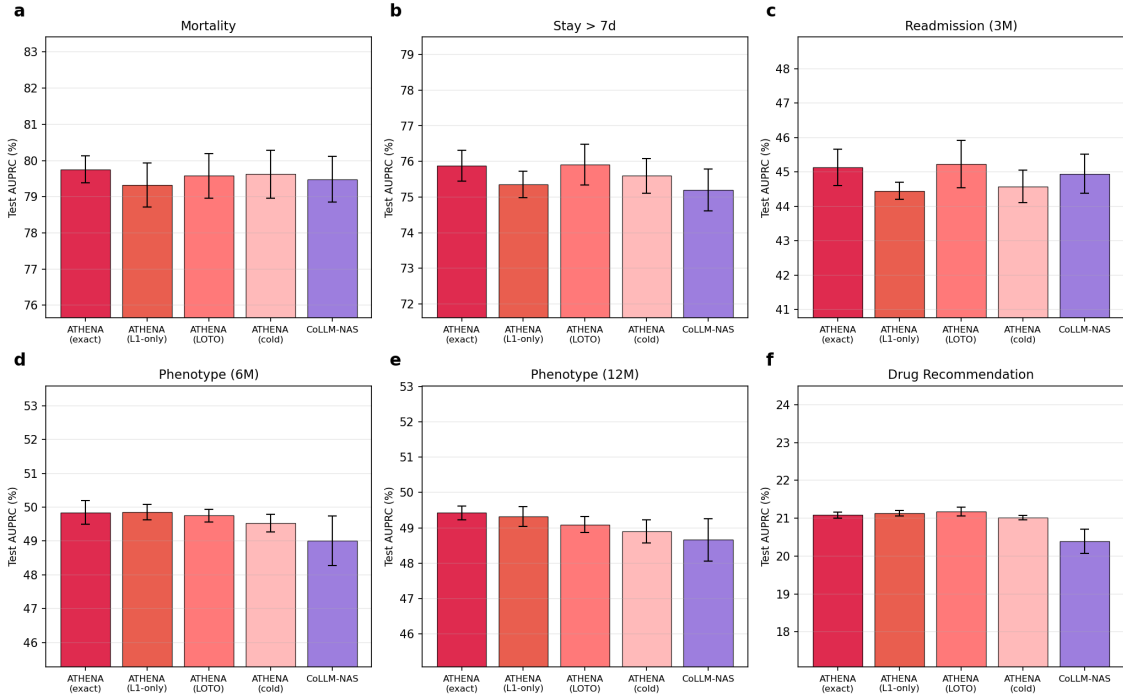


Figure 6: Cross-hospital prior ablation on the external MIMIC-IV target. Bars and error bars show the mean and SD, respectively, of test AUPRC across five random seeds for exact task-matched retrieval, *L1-only* retrieval without the *Layer 2* architecture-effect prior, *LOTO* Retrieval excluding exact task matches, Cold Start without either prior layer, and CoLLM-NAS as the representative baseline.

for multilabel outcomes. Notably, LOTO Retrieval remained competitive even when exact task matches were excluded from the source knowledge base. This result suggests that the transferred information is not limited to reusing architectures previously optimized for the same endpoint; architectural patterns learned from related prediction problems may also provide useful guidance for a new task. At the same time, the observed task-specific architecture effects argue against a single universally optimal Transformer configuration for longitudinal EHR data. Clinical prediction tasks differ in outcome structure, prediction horizon, and the information that must be integrated across a patient's record, and these differences may favor different model capacities and configurations. Cross-hospital transfer of architecture knowledge should therefore complement, rather than replace, adaptation to the target task.

ATHENA extends existing LLM-guided NAS approaches that primarily derive search guidance from information generated during the current search process [16, 17]. Rather than beginning each search without prior architectural evidence, ATHENA incorporates structured knowledge accumulated from previously evaluated hospital-task settings and updates this guidance using validation performance at the target site. This distinction is particularly relevant to multi-institutional EHR research, where models are repeatedly developed or adapted across institutions and clinical endpoints. Differences in patient populations, coding practices, clinical care patterns, data density, and outcome prevalence can affect both model performance and the architecture best suited to a prediction problem. Nevertheless, architecture selection is often repeated independently for each new dataset or endpoint. Performing an extensive architecture search for every hospital-task combination can become computationally burdensome, particularly for pretrained Transformer models. Reusing evidence from previous architecture evaluations provides a practical middle ground between applying the same architecture across all settings and repeating a computationally intensive search from scratch for every new study.

The external evaluation on MIMIC-IV further illustrates the potential value of this approach. Architecture knowledge derived from the source hospitals remained informative when transferred to a target health system with different cohort characteristics and clinical data distributions. Importantly, ATHENA did not directly transfer a source architecture as the final model. Instead, source knowledge was used to guide candidate generation, while architecture selection continued to depend on validation performance at the target site. This separation between transferred guidance and local evaluation is important for multi-institutional clinical modeling because an architecture that performs well in one health system may not remain optimal in another because of differences in patient populations, clinical care patterns, and outcome distributions. Architecture transfer may therefore be most useful when prior evidence narrows and informs the search while target-site data retain the final role in model selection.

Several limitations should be considered. First, the evaluation included two target health systems and six downstream

prediction tasks. Additional institutions, clinical specialties, coding systems, patient populations, and prediction horizons will be needed to determine how broadly the observed transfer patterns generalize. Although MIMIC-IV provided an external target health system, it represents a single academic medical center and does not capture the heterogeneity of potential real-world deployment environments. Second, the search space was limited to Transformer embedding dimension, depth, number of attention heads, and MLP expansion ratio. Other design choices, including temporal representations, attention mechanisms, prediction heads, tokenization strategies, and training hyperparameters, were held fixed and may interact with the architectural components examined here. Third, although supernet-based evaluation generally preserved architecture rankings, ranking fidelity was lower for some binary outcomes, particularly Readmission (3M), indicating that the accuracy of weight-inherited subnetwork evaluation may vary by prediction task. Fourth, the reported target-search cost does not include the upfront computation required to construct the source-site architecture knowledge base. The efficiency advantage of ATHENA therefore depends on reuse: its potential benefit increases when accumulated architecture evidence can support multiple subsequent modeling studies rather than a single search.

In conclusion, ATHENA demonstrates a framework for reusing architecture-search knowledge across hospitals and clinical prediction tasks while preserving local model selection. By combining retrieved architecture examples, architecture-effect guidance, target-site validation, and weight-sharing evaluation, ATHENA identified strong Transformer configurations under limited search budgets while retaining task- and site-specific adaptation. More broadly, this approach reframes architecture search as knowledge that can accumulate across EHR modeling studies rather than a computational process that must be repeated independently for each new prediction problem. Future work should evaluate this strategy across more diverse health systems and model families, expand the range of transferable architectural and training information, and examine whether more efficient and reproducible architecture development can support robust external validation and prospective evaluation of EHR prediction models.

CRedit authorship contribution statement

Deyi Li: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Qi Xu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Lingyao Li:** Validation, Writing – original draft. **Tiansheng Wang:** Validation, Writing – original draft. **Muxuan Liang:** Validation, Writing – original draft. **Mei Liu:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The patient-level OneFlorida+ data used in this study are not publicly available because of privacy, institutional, and data-use restrictions. Researchers may request access through the OneFlorida+ Clinical Research Network Front Door or Coordinating Center (<https://onefl.net/front-door/>); access is contingent on eligibility, network and institutional review, applicable data-use agreements, and ethical or institutional review board approvals. MIMIC-IV is a de-identified, credentialed-access resource available through PhysioNet (<https://physionet.org/>). Access requires PhysioNet credentialing, completion of the required research training, and acceptance of the applicable data-use agreement. The study-specific analytic datasets derived from OneFlorida+ and MIMIC-IV cannot be redistributed by the authors; eligible researchers must obtain the source data independently under the respective data-governance requirements.

References

- [1] J. Wang, J. Luo, M. Ye, X. Wang, Y. Zhong, A. Chang, G. Huang, Z. Yin, C. Xiao, J. Sun, et al., Recent advances in predictive modeling with electronic health records, in: IJCAI: proceedings of the conference, Vol. 2024, 2024, p. 8272.
- [2] W. Ren, J. Zhu, Z. Liu, T. Zhao, V. Honavar, A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models, arXiv preprint arXiv:2507.12774 (2025).
- [3] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ digital medicine 4 (1) (2021) 86.
- [4] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, H. Yu, Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records, Nature communications 14 (1) (2023) 7857.
- [5] R. Poulain, R. Beheshti, Graph transformers on ehrs: Better representation improves downstream performance, in: The Twelfth International Conference on Learning Representations, 2024.
- [6] D. Li, Z. Yao, Q. Xu, M. Liang, L. Li, Z. Xu, M. Liu, Dt-behrt: Disease trajectory-aware transformer for interpretable patient representation learning, arXiv preprint arXiv:2603.10180 (2026).
- [7] M. Kirchler, M. Ferro, V. Lorenzini, R. P. van de Water, F. G. A. 3, C. Lippert, A. Ganna, Large language models improve transferability of electronic health record-based predictions across countries and coding systems, npj Digital Medicine 9 (1) (2026) 177.
- [8] S. Salmani Pour Avval, N. D. Eskue, R. M. Groves, V. Yaghoubi, Systematic review on neural architecture search: Ssp avval et al., Artificial Intelligence Review 58 (3) (2025) 73.
- [9] T. Elsken, J. H. Metzen, F. Hutter, Neural architecture search: A survey, Journal of Machine Learning Research 20 (55) (2019) 1–21.
- [10] B. Zoph, Q. V. Le, Neural architecture search with reinforcement learning, arXiv preprint arXiv:1611.01578 (2016).
- [11] E. Real, A. Aggarwal, Y. Huang, Q. V. Le, Regularized evolution for image classifier architecture search, in: Proceedings of the aaai conference on artificial intelligence, Vol. 33, 2019, pp. 4780–4789.
- [12] H. Liu, K. Simonyan, Y. Yang, Darts: Differentiable architecture search, arXiv preprint arXiv:1806.09055 (2018).
- [13] K. Kandasamy, W. Neiswanger, J. Schneider, B. Póczos, E. P. Xing, Neural architecture search with bayesian optimisation and optimal transport, Advances in neural information processing systems 31 (2018).
- [14] Z. Guo, X. Zhang, H. Mu, W. Heng, Z. Liu, Y. Wei, J. Sun, Single path one-shot neural architecture search with uniform sampling, in: European conference on computer vision, Springer, 2020, pp. 544–560.
- [15] Z. Ji, G. Zhu, C. Yuan, Y. Huang, Rz-nas: Enhancing llm-guided neural architecture search via reflective zero-cost strategy., in: ICML, 2025.
- [16] M. Zheng, X. Su, S. You, F. Wang, C. Qian, C. Xu, S. Albanie, Can gpt-4 perform neural architecture search?, arXiv preprint arXiv:2304.10970 (2023).
- [17] Z. Li, Z. Lin, Y. Wang, Collm-nas: Collaborative large language models for efficient knowledge-guided neural architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 3273–3282.
- [18] X. Su, Q. Mao, Z. Wu, X. Lin, S. You, Y. Liao, C. Xu, Large language models driven neural architecture search for universal and lightweight disease diagnosis on histopathology slide images, npj Digital Medicine 8 (1) (2025) 682.
- [19] M. Chen, H. Peng, J. Fu, H. Ling, Autoformer: Searching transformers for visual recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 12270–12280.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [21] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, Behrt: transformer for electronic health records, Scientific reports 10 (1) (2020) 7155.
- [22] C. Pang, X. Jiang, K. S. Kalluri, M. Spotnitz, R. Chen, A. Perotte, K. Natarajan, Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks, in: Machine Learning for Health, PMLR, 2021, pp. 239–260.
- [23] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, K. Rahimi, Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records, IEEE journal of biomedical and health informatics 27 (2) (2022) 1106–1117.
- [24] Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J. A. Yeung, A. Deng, A. Balston, J. Ross, E. Idowu, et al., Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study, The Lancet Digital Health 6 (4) (2024) e281–e290.
- [25] M. Wornow, R. Thapa, E. Steinberg, J. Fries, N. Shah, Ehrshot: An ehr benchmark for few-shot evaluation of foundation models (2023), URL <https://arxiv.org/abs/2307.02028> 2307.
- [26] C. White, M. Safari, R. Sukthankar, B. Ru, T. Elsken, A. Zela, D. Dey, F. Hutter, Neural architecture search: Insights from 1000 papers, arXiv preprint arXiv:2301.08727 (2023).
- [27] D. So, Q. Le, C. Liang, The evolved transformer, in: International conference on machine learning, PMLR, 2019, pp. 5877–5886.
- [28] H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, S. Han, Hat: Hardware-aware transformers for efficient natural language processing, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7675–7688.
- [29] J. Xu, X. Tan, R. Luo, K. Song, J. Li, T. Qin, T.-Y. Liu, Nas-bert: Task-agnostic and adaptive-size bert compression with neural architecture search, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1933–1943.
- [30] J. Gao, H. Xu, H. Shi, X. Ren, P. L. Yu, X. Liang, X. Jiang, Z. Li, Autobert-zero: Evolving bert backbone from scratch, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 10663–10671.
- [31] D. R. So, W. Manke, H. Liu, Z. Dai, N. Shazeer, Q. V. Le, Primer: Searching for efficient transformers for language modeling, 2022, URL <https://arxiv.org/abs/2109.08668> (2022).
- [32] M. Bal-Ghaoui, M. Tiouti, Metallmix: An xai aided llm-meta-learning based approach for hyper-parameters optimization, arXiv preprint arXiv:2509.09387 (2025).

- [33] Z. Xu, D. R. So, A. M. Dai, Mufasa: Multimodal fusion architecture search for electronic health records, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10532–10540.
- [34] S. Cui, J. Wang, Y. Zhong, H. Liu, T. Wang, F. Ma, Automated fusion of multimodal electronic health records for better medical predictions, in: Proceedings of the... SIAM International Conference on Data Mining. SIAM International Conference on Data Mining, Vol. 2024, 2024, p. 361.
- [35] A. Chen, D. Dohan, D. So, Evoprompting: Language models for code-level neural architecture search, *Advances in neural information processing systems* 36 (2023) 7787–7817.
- [36] C. Yu, X. Liu, Y. Wang, Y. Liu, W. Feng, D. Xiong, C. Tang, J. Lv, Gpt-nas: Evolutionary neural architecture search with the generative pre-trained model, *arXiv preprint arXiv:2305.05351* (2023).
- [37] M. U. Nasir, S. Earle, J. Togelius, S. James, C. Cleghorn, Llmatic: neural architecture search via large language models and quality diversity optimization, in: proceedings of the Genetic and Evolutionary Computation Conference, 2024, pp. 1110–1118.
- [38] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical concepts, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1495–1504.
- [39] T. Zhang, M. Chen, A. A. T. Bui, Diagnostic prediction with sequence-of-sets representation learning for clinical events, in: Proc. Int. Conf. Artificial Intelligence in Medicine (AIME), Vol. 12299 of LNAI, Springer, 2020, pp. 348–358.
- [40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
- [41] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
- [42] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, *Nature Machine Intelligence* 2 (1) (2020) 2522–5839.
- [43] R. L. Fleurence, L. H. Curtis, R. M. Califf, R. Platt, J. V. Selby, J. S. Brown, Launching pcornet, a national patient-centered clinical research network, *Journal of the American Medical Informatics Association* 21 (4) (2014) 578–582. doi:10.1136/amiajnl-2014-002747. URL <https://doi.org/10.1136/amiajnl-2014-002747>
- [44] W. R. Hogan, et al., The oneflorida data trust: a centralized, translational research data infrastructure of statewide scope, *Journal of the American Medical Informatics Association* 29 (4) (2022) 686–693. doi:10.1093/jamia/ocab221.
- [45] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al., MIMIC-IV, a freely accessible electronic health record dataset, *Scientific data* 10 (1) (2023) 1.
- [46] R. L. Fleurence, L. H. Curtis, R. M. Califf, R. Platt, J. V. Selby, J. S. Brown, Launching pcornet, a national patient-centered clinical research network, *Journal of the American Medical Informatics Association* 21 (4) (2014) 578–582.
- [47] R. Xu, M. K. Ali, J. C. Ho, C. Yang, Hypergraph transformers for ehr-based clinical predictions, *AMIA Summits on Translational Science Proceedings* 2023 (2023) 582.
- [48] W. Yao, K. Yin, W. K. Cheung, J. Liu, J. Qin, Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 38, 2024, pp. 16416–16424.

Supplementary materials for:

ATHENA: Knowledge-guided agentic neural architecture search for AutoFormer-based electronic health record modeling

Deyi Li^{1,†,*}, Qi Xu^{1,†}, Lingyao Li², Tiansheng Wang^{3,4}, Muxuan Liang⁵ and Mei Liu^{1,**}

¹Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA

²College of Information Science, University of Arizona, Tucson, AZ, USA

³Department of Pharmaceutical Health Outcomes and Policy, College of Pharmacy, University of Houston, Houston, TX, USA

⁴Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁵Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

[†]These authors contributed equally to this work.

*Corresponding author: lideyi@ufl.edu

**Corresponding author: mei.liu@ufl.edu

Method S1. Cohort construction and inclusion/exclusion criteria

The following criteria were applied to all cohorts.

- **Multimodal completeness.** An admission was included if it had at least one record in each of the diagnosis, medication, procedure, and laboratory tables. Admissions missing any of these modalities were excluded.
- **Age.** Only adult admissions (≥ 18 years) were included. Ages of 89 years and older were coded as 90 years and discretized into 20 equal-width bins.
- **Number of clinical events.** No minimum number of clinical events was required beyond the multimodal completeness criterion.
- **Number of visits.** No minimum number of admissions per patient was required. To limit computational burden, patients with more than eight distinct admissions were excluded.

Method S2. Implementation details

- **Architecture constraints and output heads.** The target-specific parameter limit was $P_{\max} = 4,000,000$. The componentwise maximal supernet configuration was $(d, L, \gamma, h) = (256, 8, 8, 8)$, and the internal query/key/value width was fixed at 256 for every subnet. All search methods operated on the legal space $\mathcal{A}_{H,t}$ defined in the main-text Methods and used the same hospital-specific pretrained supernet.
- **Cross-hospital prior.** The source prior used four de-identified OneFlorida+ sites and 100 distinct architectures per source-hospital-task pair; both target cohorts were excluded from prior construction. Source metrics were averaged over the three checkpoints with the highest source-validation AUPRC. ATHENA retrieved $K = 5$ source architectures from the task-matched or fallback source set. In the task descriptor, $\tilde{C}_t = \log(C_t)/\log(20)$ and $\tilde{\Delta}_t = \Delta_t/365$, with $\Delta_t = 0$ for same-admission outcomes. Architecture-effect labels and interaction rules used normal-approximation 95% CIs, and interaction cells required at least three records.
- **Agentic search.** ATHENA permitted at most $R = 3$ critique passes per search round and terminated after $F_{\max} = 3$ consecutive rounds without an accepted legal candidate. The maximum evaluation budget was $B = 30$. Results were summarized after 5, 20, and 30 evaluations.
- **Masked language modeling (MLM) and supernet pretraining.** For MLM, 15% of non-special tokens were selected; 80% of selected tokens were replaced by [MASK], 10% by a random token of the same type, and 10% were unchanged. Supernet pretraining ran for at most 100 epochs with validation-loss patience 5. Optimization used AdamW, batch size 64, learning rate 2×10^{-4} , weight decay 10^{-2} , and gradient clipping at norm 1.0. The learning-rate schedule used linear warm-up over the first 10% of epochs followed by cosine decay to 1% of the peak rate.
- **Candidate fine-tuning and checkpoint selection.** Candidate fine-tuning used the same optimizer settings for at most 30 epochs with validation-AUPRC patience 5. Search-time metrics were averaged over the three epochs with the highest validation AUPRC, whereas the weights from the single best validation-AUPRC epoch were retained for final test evaluation. Attention dropout, residual dropout, and drop-path rate were each 0.1.
- **Large language model (LLM) configuration.** All LLM-based methods used `anthropic/claude-3.5-haiku`. The Proposal, Critic, and Strategy modules returned structured JSON; invalid responses were retried and otherwise contributed to the consecutive-failure criterion.

Table S1. Definitions of the six clinical prediction tasks. Each task is defined relative to the *target admission*, i.e., the hospital admission being modeled.

Mortality (binary): Whether the patient dies during the target admission (1 if a recorded in-hospital death time is present, 0 otherwise). The outcome is assessed only during the target admission, from admission to discharge or death, with no post-discharge follow-up.

Stay > 7d (binary): Whether the length of stay of the target admission exceeds seven days (1 if > 7 days, 0 otherwise), measured from admission to discharge. The outcome is defined for the single target admission rather than as a cumulative measure across admissions, and all admissions are included regardless of discharge disposition.

Readmission (3M) (binary): Whether the patient has a subsequent hospital admission within 3 months (90 days) of the target admission (1 if a subsequent admission occurs within 90 days, 0 otherwise). The interval is measured from the discharge date of the target admission to the admission date of the subsequent admission.

Phenotype (6M) (multilabel, 18 classes): The 18 predefined phenotype categories present among the diagnoses recorded at the patient’s next hospital admission, provided that the next admission occurs within 6 months (180 days) of the target admission. Only target admissions followed by a qualifying next admission are included.

The 18 categories are: (1) acute and unspecified renal failure; (2) acute cerebrovascular disease; (3) acute myocardial infarction; (4) cardiac dysrhythmias; (5) chronic kidney disease; (6) chronic obstructive pulmonary disease; (7) conduction disorders; (8) congestive heart failure (nonhypertensive); (9) coronary atherosclerosis and related; (10) disorders of lipid metabolism; (11) essential hypertension; (12) fluid and electrolyte disorders; (13) gastrointestinal hemorrhage; (14) hypertension with complications; (15) other liver diseases; (16) other lower respiratory disease; (17) pneumonia; and (18) septicemia (except in labor).

Phenotype (12M) (multilabel, 18 classes): The same 18 phenotype categories are used as in the 6-month task, but the next hospital admission must occur within 12 months (365 days) of the target admission. The label is an 18-dimensional binary vector, and only target admissions followed by a qualifying next admission are included.

Drug Recommendation (multilabel, 55 classes): The medication classes prescribed during the target admission, based on the admission diagnoses and the patient’s prior visit history. Ground truth consists of physician-prescribed drugs mapped to Anatomical Therapeutic Chemical (ATC) level-4 classes. Of the 140 ATC level-4 classes, the 55 classes present in $\geq 1\%$ of admissions (prevalence, 1.0–21.8%) are retained. The resulting label is a 55-dimensional binary vector. For model input, diagnoses from the target admission are retained, whereas medication, laboratory, and procedure tokens from that admission are masked; all tokens from prior visits remain available.

The 55 ATC level-4 classes are: (1) antithrombotic agents (B01A); (2) other analgesics and antipyretics (N02B); (3) stomatological preparations (A01A); (4) laxatives (A06A); (5) drugs for peptic ulcer and gastro-oesophageal reflux (A02B); (6) other mineral supplements (A12C); (7) lipid-modifying agents (C10A); (8) irrigating solutions (B05C); (9) calcium (A12A); (10) beta-blocking agents (C07A); (11) opioids (N02A); (12) antidepressants (N06A); (13) antiemetics and antinauseants (A04A); (14) ACE inhibitors, plain (C09A); (15) antacids (A02A); (16) thyroid preparations (H03A); (17) pancreatic hormones (H04A); (18) insulins and analogues (A10A); (19) selective calcium-channel blockers (C08C); (20) other beta-lactam antibacterials (J01D); (21) high-ceiling diuretics (C03C); (22) vitamin B12 and folic acid (B03B); (23) intestinal anti-infectives (A07A); (24) anxiolytics (N05B); (25) adrenergics, inhalants (R03A); (26) antiepileptics (N03A); (27) other cardiac preparations (C01E); (28) thiazide diuretics (C03A); (29) cardiac vasodilators (C01D); (30) topical antihemorrhoidals (C05A); (31) class I and III antiarrhythmics (C01B); (32) angiotensin-II receptor blockers (C09C); (33) quinolone antibacterials (J01M); (34) antipsychotics (N05A); (35) antigout preparations (M04A); (36) potassium (A12B); (37) intestinal anti-inflammatory agents (A07E); (38) penicillins (J01C); (39) general anesthetics (N01A); (40) arteriolar smooth-muscle agents (C02D); (41) anti-inflammatory and antirheumatic NSAIDs (M01A); (42) dermatological corticosteroids (D07A); (43) propulsives (A03F); (44) antipruritics (D04A); (45) hypnotics and sedatives (N05C); (46) other dermatological preparations (D11A); (47) drugs used in addictive disorders (N07B); (48) cardiac stimulants, excluding glycosides (C01C); (49) topical nasal decongestants (R01A); (50) potassium-sparing agents (C03D); (51) direct-acting antivirals (J05A); (52) sulfonamides and trimethoprim (J01E); (53) cough suppressants (R05D); (54) bacterial vaccines (J07A); and (55) blood glucose-lowering drugs, excluding insulins (A10B).

Table S2. Baseline characteristics of the six study cohorts. Sites A–D are OneFlorida+ source sites used to construct the cross-hospital architecture prior, Site E is the held-out OneFlorida+ internal target, and MIMIC-IV is the external target. Continuous variables are reported as mean (SD) or median [IQR] and categorical variables as n (%). Sex percentages are calculated among patients with recorded sex, with missing sex reported separately. Comorbidities are reported as the number and percentage of admissions with the corresponding diagnosis.

Characteristic	OneFlorida+ Source Sites				Internal Target	External Target
	Site A	Site B	Site C	Site D	Site E	MIMIC-IV
<i>Demographics</i>						
Patients, n	52,831	55,735	88,527	58,282	84,479	60,706
Admissions, n	144,313	119,579	200,911	97,303	135,955	84,492
Female, n (%)	5,547 (48.8)	25,439 (49.0)	18,511 (45.7)	26,223 (46.3)	44,078 (54.6)	30,198 (49.7)
Sex not recorded, n (%)	41,470 (78.5)	3,863 (6.9)	48,059 (54.3)	1,653 (2.8)	3,816 (4.5)	0 (0.0)
<i>Utilization</i>						
Admissions per patient, mean (SD)	2.7 (1.7)	2.1 (1.5)	2.3 (1.4)	1.7 (1.1)	1.6 (1.1)	1.4 (0.9)
Length of stay, days, median [IQR]	2 [0–4]	4 [1–7]	4 [2–7]	5 [3–16]	4 [2–7]	4 [2–8]
Diagnoses per admission, mean (SD)	10.8 (7.7)	14.2 (10.6)	19.1 (15.6)	12.5 (8.3)	5.4 (4.2)	11.9 (7.2)
Medications per admission, mean (SD)	11.3 (7.1)	3.5 (2.7)	13.1 (9.9)	3.2 (2.3)	3.9 (2.8)	3.3 (2.5)
<i>Outcomes</i>						
In-hospital mortality, n (%)	644 (0.4)	1,263 (1.1)	4,955 (2.5)	2,226 (2.3)	5,317 (3.9)	3,577 (4.2)
Stay > 7d, n (%)	13,790 (9.6)	26,063 (21.8)	30,841 (15.4)	33,049 (34.0)	28,317 (20.8)	23,031 (27.3)
Readmission (3M), n (%)	30,264 (21.0)	24,455 (20.5)	47,990 (23.9)	19,495 (20.0)	23,242 (17.1)	10,727 (12.7)
<i>Comorbidities (n, % of admissions)</i>						
Essential hypertension	50,545 (35.0)	52,639 (44.0)	85,260 (42.4)	33,248 (34.2)	26,591 (19.6)	35,443 (41.9)
Disorders of lipid metabolism	50,156 (34.8)	33,518 (28.0)	55,822 (27.8)	23,269 (23.9)	5,368 (3.9)	30,498 (36.1)
Cardiac dysrhythmias	26,555 (18.4)	20,763 (17.4)	37,695 (18.8)	12,110 (12.4)	7,644 (5.6)	19,681 (23.3)
Congestive heart failure	20,014 (13.9)	19,278 (16.1)	28,832 (14.4)	10,740 (11.0)	6,979 (5.1)	13,789 (16.3)
Coronary atherosclerosis	8,808 (6.1)	4,809 (4.0)	9,244 (4.6)	3,055 (3.1)	1,543 (1.1)	19,675 (23.3)
Acute/unspecified renal failure	18,337 (12.7)	18,995 (15.9)	29,141 (14.5)	12,409 (12.8)	12,003 (8.8)	13,861 (16.4)
Chronic kidney disease	10,894 (7.5)	9,064 (7.6)	15,732 (7.8)	6,697 (6.9)	3,791 (2.8)	13,489 (16.0)
COPD	11,814 (8.2)	11,984 (10.0)	20,920 (10.4)	4,581 (4.7)	3,957 (2.9)	7,634 (9.0)
Pneumonia	6,468 (4.5)	7,449 (6.2)	15,576 (7.8)	4,235 (4.4)	6,086 (4.5)	5,971 (7.1)
Fluid and electrolyte disorders	29,744 (20.6)	36,240 (30.3)	49,495 (24.6)	29,854 (30.7)	13,787 (10.1)	19,550 (23.1)

Table S3. *Layer 2* pairwise architecture-effect rules supplied to the search prompt (M_{meta}). For each task, the interaction model uses the two most influential architecture features. The table reports feature-level combinations with estimated effects whose 95% confidence intervals (CIs) exclude zero.

Task	Feature pair	Rule supplied to the prompt
Mortality	<code>embed_dim</code> $\times$ <code>heads</code>	Prefer <code>embed_dim</code> $\in \{128, 256\}$, with the preference for 256 applying when <code>heads</code> ≥ 2 ; discourage <code>embed_dim</code> $\in \{32, 64\}$.
Stay > 7d	<code>depth</code> $\times$ <code>embed_dim</code>	Prefer <code>depth</code> ≥ 2 ; discourage <code>depth</code> = 1 across all <code>embed_dim</code> levels.
Readmission (3M)	<code>heads</code> $\times$ <code>embed_dim</code>	Prefer <code>heads</code> $\in \{4, 8\}$; discourage <code>heads</code> $\in \{1, 2\}$ across all <code>embed_dim</code> levels.
Phenotype (6M)	<code>embed_dim</code> $\times$ <code>depth</code>	Prefer <code>embed_dim</code> $\in \{128, 256\}$; discourage <code>embed_dim</code> $\in \{32, 64\}$ across all <code>depth</code> levels.
Phenotype (12M)	<code>embed_dim</code> $\times$ <code>depth</code>	Prefer <code>embed_dim</code> $\in \{128, 256\}$; discourage <code>embed_dim</code> $\in \{32, 64\}$ across all <code>depth</code> levels.
Drug Recommendation	<code>embed_dim</code> $\times$ <code>depth</code>	Prefer <code>embed_dim</code> $\in \{128, 256\}$; discourage <code>embed_dim</code> $\in \{32, 64\}$ across all <code>depth</code> levels.

Table S4. Test AUROC at a search budget of 30 architecture evaluations (mean $\pm$ standard deviation over five random seeds). Within each hospital–task comparison, the best mean is shown in **bold** and the second-best mean is underlined; ties share the same mark. *Avg. Rank* denotes the mean rank across the 12 hospital–task comparisons (1 = best; ties receive the average rank), with the number of best-in-row comparisons shown in parentheses.

Target	Task	Random Search	EA	GENIUS	CoLLM-NAS	ATHENA
OneFlorida+	Mortality	89.15 $\pm$ 0.72	89.71 $\pm$ 0.51	89.07 $\pm$ 1.15	88.94 $\pm$ 1.49	<u>89.30 $\pm$ 0.95</u>
OneFlorida+	Stay > 7d	90.32 $\pm$ 0.31	90.31 $\pm$ 0.13	90.46 $\pm$ 0.13	<u>90.39 $\pm$ 0.30</u>	90.30 $\pm$ 0.25
OneFlorida+	Readmission (3M)	70.22 $\pm$ 0.28	70.52 $\pm$ 0.48	70.38 $\pm$ 0.19	70.29 $\pm$ 0.37	<u>70.51 $\pm$ 0.35</u>
OneFlorida+	Phenotype (6M)	<u>78.74 $\pm$ 0.55</u>	78.82 $\pm$ 0.31	78.39 $\pm$ 0.32	78.18 $\pm$ 0.71	78.82 $\pm$ 0.25
OneFlorida+	Phenotype (12M)	<u>79.86 $\pm$ 0.40</u>	79.84 $\pm$ 0.44	79.64 $\pm$ 0.26	78.85 $\pm$ 0.25	80.05 $\pm$ 0.10
OneFlorida+	Drug Recommendation	71.21 $\pm$ 0.08	71.26 $\pm$ 0.08	71.31 $\pm$ 0.07	71.20 $\pm$ 0.18	<u>71.28 $\pm$ 0.09</u>
MIMIC-IV	Mortality	96.09 $\pm$ 0.14	<u>96.18 $\pm$ 0.06</u>	96.13 $\pm$ 0.17	<u>96.18 $\pm$ 0.16</u>	96.19 $\pm$ 0.16
MIMIC-IV	Stay > 7d	85.14 $\pm$ 0.26	<u>85.43 $\pm$ 0.20</u>	85.33 $\pm$ 0.25	85.06 $\pm$ 0.36	85.56 $\pm$ 0.39
MIMIC-IV	Readmission (3M)	67.89 $\pm$ 0.19	<u>67.90 $\pm$ 0.78</u>	67.77 $\pm$ 0.45	67.84 $\pm$ 0.23	68.08 $\pm$ 0.40
MIMIC-IV	Phenotype (6M)	80.04 $\pm$ 0.47	79.89 $\pm$ 0.48	79.83 $\pm$ 0.52	79.47 $\pm$ 0.58	80.07 $\pm$ 0.25
MIMIC-IV	Phenotype (12M)	<u>79.38 $\pm$ 0.24</u>	<u>79.38 $\pm$ 0.27</u>	79.15 $\pm$ 0.44	78.89 $\pm$ 0.61	79.53 $\pm$ 0.19
MIMIC-IV	Drug Recommendation	78.25 $\pm$ 0.17	78.37 $\pm$ 0.07	77.90 $\pm$ 0.46	77.41 $\pm$ 0.46	<u>78.32 $\pm$ 0.17</u>
<i>Avg. Rank (best-in-row)</i>		3.29 (0)	<u>2.21 (4)</u>	3.42 (2)	4.38 (0)	1.71 (7)

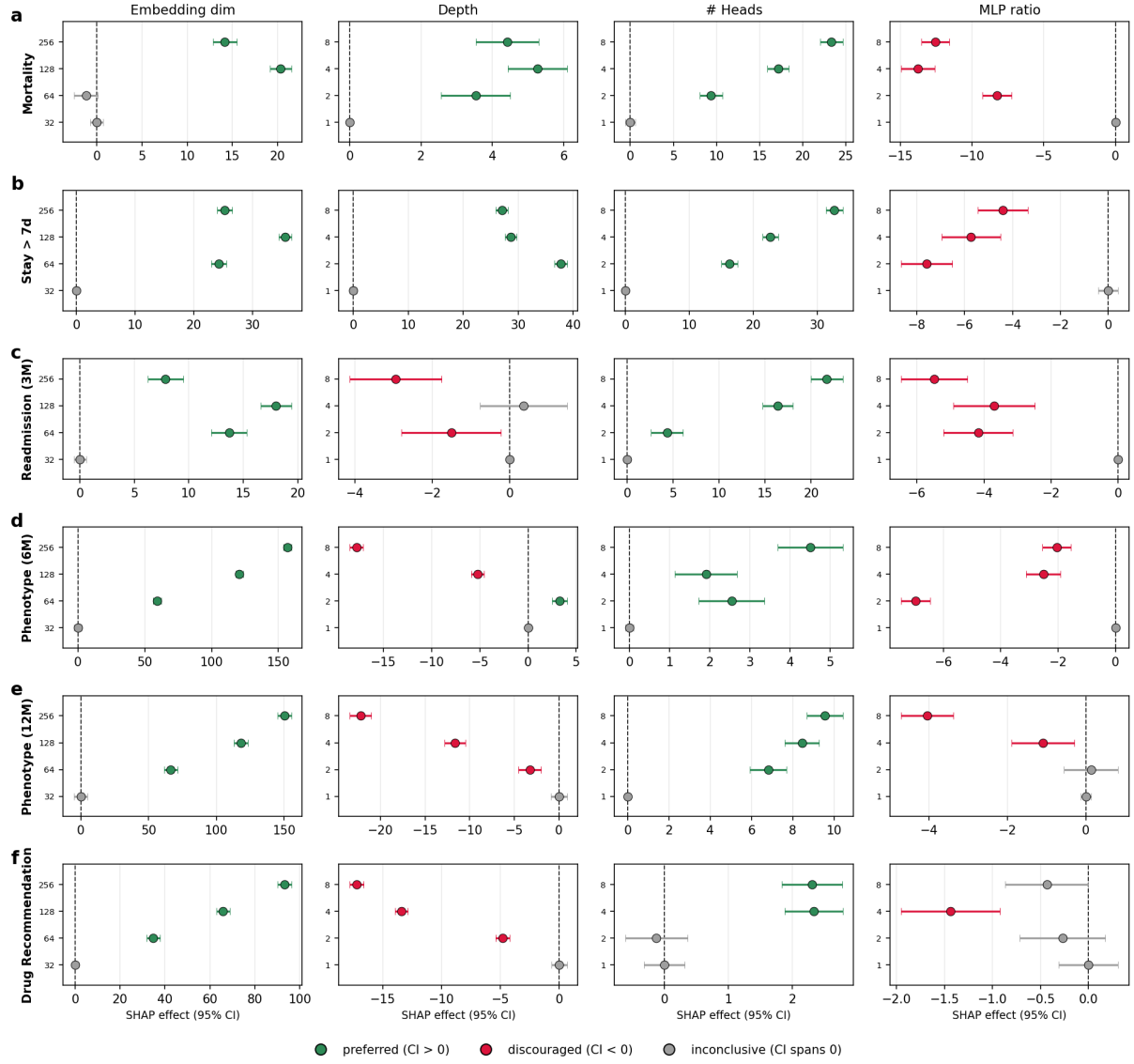


Figure S1. Cross-hospital architecture prior learned by the *Layer 2* meta-regression across four One-Florida+ source sites. Rows denote the six prediction tasks: (a) Mortality, (b) Stay > 7d, (c) Readmission (3M), (d) Phenotype (6M), (e) Phenotype (12M), and (f) Drug Recommendation; columns denote architecture dimensions. Points show mixed-effects SHapley Additive exPlanations (SHAP) estimates with 95% CIs for each hyperparameter level. Green indicates preferred levels (CI > 0), red indicates discouraged levels (CI < 0), and gray indicates inconclusive levels whose CIs span 0. Gray points at zero without CIs denote reference levels.

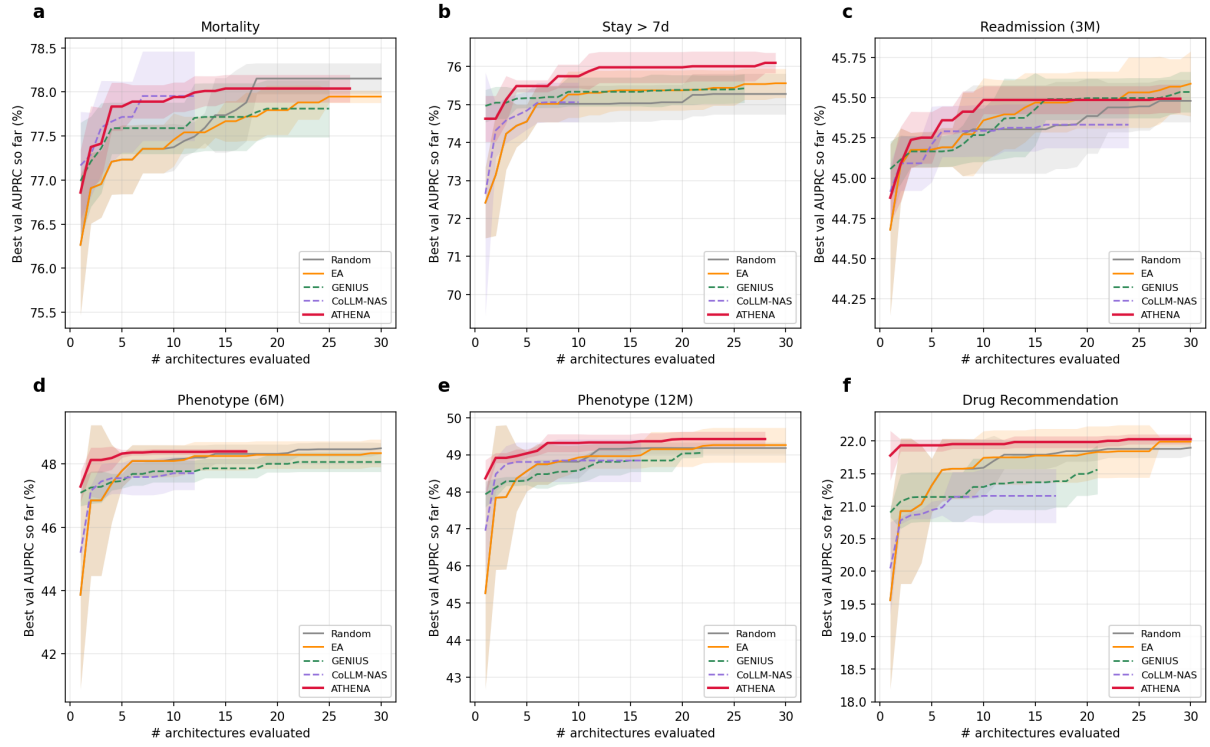


Figure S2. Search trajectories on the held-out MIMIC-IV target for (a) Mortality, (b) Stay > 7d, (c) Readmission (3M), (d) Phenotype (6M), (e) Phenotype (12M), and (f) Drug Recommendation. Lines and shaded bands show the mean and SD, respectively, of the best validation AUPRC attained across five random seeds as a function of the number of evaluated architectures. Curves may end before the full budget when proposal saturation yields no new legal architecture.

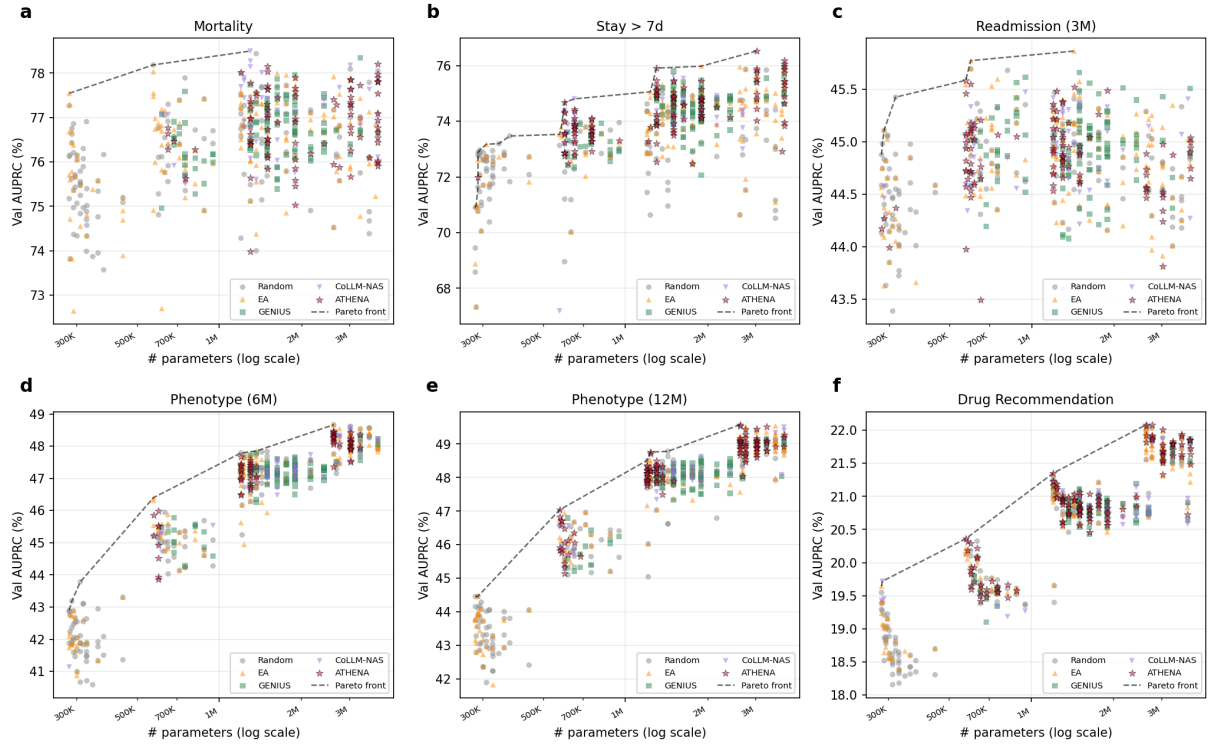


Figure S3. Validation AUPRC versus parameter count on the held-out MIMIC-IV target for (a) Mortality, (b) Stay > 7d, (c) Readmission (3M), (d) Phenotype (6M), (e) Phenotype (12M), and (f) Drug Recommendation. Markers denote search methods, and the dashed line indicates the empirical Pareto frontier across evaluated architectures. Validation performance guided architecture search and selection; the selected architecture was evaluated only once on the held-out test set.

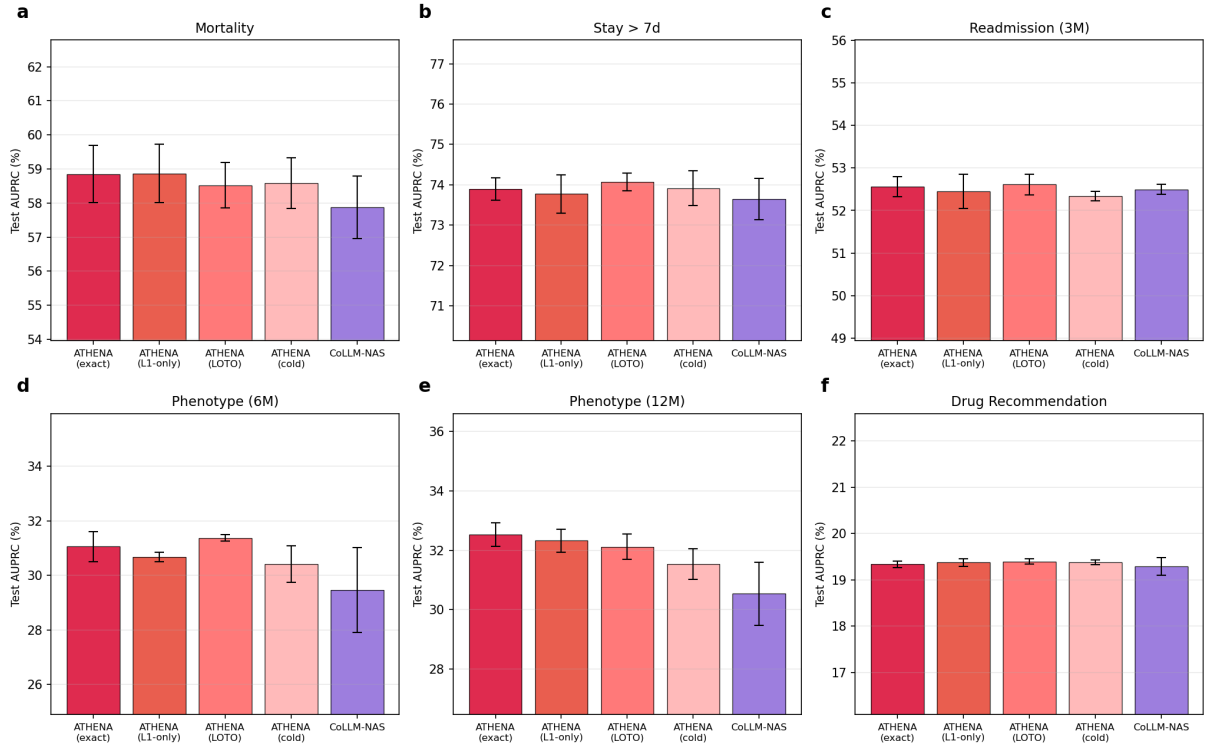


Figure S4. Cross-hospital prior ablation on the held-out OneFlorida+ target. Bars and error bars show the mean and SD, respectively, of test AUPRC across five random seeds for exact task-matched retrieval, *L1-only* retrieval without the *Layer 2* architecture-effect prior, Leave-One-Task-Out (LOTO) Retrieval excluding exact task matches, Cold Start without either prior layer, and CoLLM-NAS as the representative baseline.