MOSAIC : MIXTURE-OF-SPECIALTY ADAPTATION FOR INDIVIDUALIZED CLINICAL-PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health record (EHR) foundation models share one set of parameters across all patients: a patient's history enters only through the input context, and adapting these models to a clinical setting requires updating millions of parameters. Yet clinical personalization has natural structure: specialties reason differently about the same findings, and patients within a specialty follow different trajectories. We present MOSAIC (**M**ixture-**O**f-**S**pecialty **A**daptation for **I**ndividualized **C**linical-prediction), which captures both levels through composition over a frozen LLM rather than per-patient parameter updates. At the cohort level, evolutionary prompt optimization discovers specialty-specific clinical reasoning rules in natural language. At the patient level, micro-adapters are composed through trajectory-conditioned routing, yielding adapter weights that reflect each patient's disease history. On a frozen Qwen2.5-7B backbone with only 78K trainable parameters, MOSAIC exceeds 18 baselines on MIMIC-IV next-visit diagnosis prediction at four levels of diagnostic granularity and outperforms standard LoRA at a matched parameter budget. Cohort reasoning contributes mainly at coarse granularity and patient calibration mainly at fine granularity. The approach transfers to eICU, a multi-center critical-care database. MOSAIC also predicts across a broad diagnostic vocabulary and improves monotonically with patient history length, where deep EHR baselines plateau.

1 INTRODUCTION

Personalized clinical prediction, the task of anticipating what will happen to *this* patient rather than the average patient, remains an open problem. One setting where this matters is next-visit diagnosis prediction, a standard benchmark for longitudinal EHR models that requires reasoning over a patient's full history to predict from a vocabulary of thousands of possible diagnoses (Mukherjee et al., 2023; Li et al., 2020); we treat it as a testbed for adaptation mechanisms rather than as a deployable clinical instrument. Electronic health record (EHR) foundation models (Steinberg et al., 2021; Li et al., 2020; Rasmy et al., 2021; Renc et al., 2024; Yang et al., 2023; Pang et al., 2024; Kraljevic et al., 2024; Shmatko et al., 2025) learn powerful sequential representations from millions of patient records, but the same parameters generate predictions for every patient regardless of specialty context, disease acuity, or individual trajectory. Adapting these models to a clinical context requires updating millions of parameters. More fundamentally, they provide no mechanism for explicitly or persistently modeling an individual: a patient's history enters only through the input context, and the mapping from that context to a prediction is shared across the entire population.

We observe that clinical personalization has natural structure. A cardiologist and a nephrologist reason differently as they attend to different risk factors, anticipate different complications, and apply different clinical heuristics (Gold et al., 2022). But two cardiac patients also differ from each other: one with stable chronic disease and another with rapidly evolving multimorbidity require fundamentally different predictions even within the same specialty. Effective personalization must therefore operate at multiple levels: adapting clinical reasoning to the cohort while adapting predictions to the individual. We implement this insight as MOSAIC , a two-level adaptation hierarchy on a frozen Qwen 2.5-7B backbone (Qwen et al., 2025). At the *cohort level*, GEPA, an evolutionary prompt optimization framework, discovers specialty-specific clinical reasoning rules in natural language: which comorbidity patterns signal progression, when to predict recurrence versus novel diagnosis, and which risk factors dominate in each clinical context (Agrawal et al., 2026). At the

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

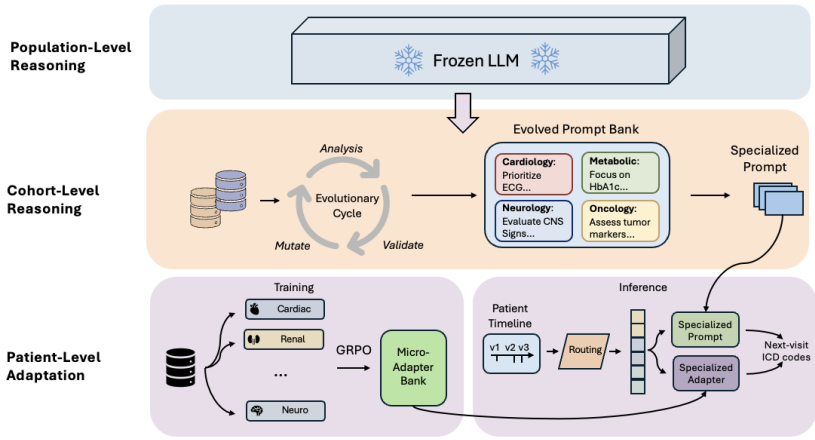

Figure 1: Overview of MOSAIC . A frozen LLM supplies population-level medical knowledge, evolved specialty prompts encode cohort-level reasoning, and dynamically routed micro-adapters provide patient-level calibration.

patient level, we train micro-adapters of 9.8K parameters per specialty using TinyLoRA (Morris et al., 2026), which extends LoRA (Hu et al., 2022) and LoRA-XS (Bałazy et al., 2025). These adapters are composed through trajectory-conditioned dynamic routing, producing patient-conditioned weights that reflect each individual's evolving disease profile at prediction time. These two mechanisms operate at different levels: GEPA encodes the clinical logic of *how* to reason about a specialty, while dynamically routed adapters calibrate *what* to predict for this particular patient. Each patient's diagnosis history determines a unique routing distribution over specialty adapters and a unique input context, so the same frozen backbone produces patient-conditioned predictions even though no parameters are updated for any individual.

We evaluate MOSAIC on MIMIC-IV (Johnson et al., 2023) and eICU (Pollard et al., 2018) on next-visit diagnosis prediction against 18 deep EHR and LLM-based baselines. MOSAIC exceeds all baselines on MIMIC-IV at every granularity and matches or exceeds them on eICU except Chapter F1. It predicts across a broader vocabulary than all tested deep models, covering conditions other models miss entirely. The two adaptation levels contribute at different diagnostic granularities: evolved cohort reasoning contributes primarily at coarse granularity while patient-level calibration contributes primarily at fine granularity. The contributions of this paper are as follows:

1. MOSAIC , **a two-level personalization decomposition** into cohort reasoning and patient calibration that exceeds all baselines on MIMIC-IV at every diagnostic granularity with 78K trainable parameters.
2. A **granularity-specific contribution finding**: cohort-level reasoning and patient-level calibration contribute at different diagnostic granularities, with each addressing a level of the prediction problem the other does not.
3. Evidence that this decomposition **recovers the diagnostic long tail**: MOSAIC covers all clinical body systems and improves monotonically with patient history length.

2 METHOD

Given a patient's longitudinal EHR history $H = (e_1, e_2, \ldots, e_t)$, a chronologically ordered sequence of timestamped diagnoses, medications, and procedures observed up to cutoff time t, we predict the set $D \subseteq$ ICD-10 of diagnosis codes recorded at the patient's next clinical visit. Each event $e = (\text{age}, \text{type}, \text{code})$ carries the patient's age at occurrence, the event type, and a standardized code (ICD-10 for diagnoses, drug name for medications).

Our system decomposes this prediction into three sources of knowledge (Figure 1): *population-level* medical knowledge from a frozen pretrained LLM, *cohort-level* clinical reasoning from evolutionary prompt optimization, and *patient-level* distributional calibration from dynamically routed micro-

adapters. The population level requires no adaptation: rather than training a domain-specific model on EHR sequences, we use a frozen Qwen2.5-7B-Instruct backbone (Qwen et al., 2025), quantized to 4-bit NF4 via BitsAndBytes (Dettmers et al., 2023). We describe the two adapted levels in turn, then their composition at inference time.

2.1 COHORT-LEVEL REASONING

A pretrained LLM possesses broad medical knowledge, including specialty-relevant clinical reasoning. However, this knowledge is general-purpose as it reflects textbook medicine rather than the specific diagnostic patterns, comorbidity trajectories, and coding conventions of a particular patient population. A cardiologist and a gastroenterologist interpret the same finding differently; these specialty-specific heuristics must be discovered from clinical data. Encoding them through weight updates would sacrifice the parameter efficiency of the frozen backbone. Instead, we discover them through evolutionary prompt optimization.

Cohort assignment. We partition patients into eight cohorts by a deterministic rule over the body-system categories of the Clinical Classifications Software Refined (CCSR) (Healthcare Cost and Utilization Project (HCUP), 2024): a patient is assigned to the cohort whose CCSR bucket accounts for at least 30% of the pre-cutoff diagnoses in their record, and to a *generalist* cohort otherwise. The resulting cohorts are cardiac, metabolic, renal/genitourinary, digestive, neuro/mental, oncology/hematology, musculoskeletal/trauma, and generalist. Cohort prototypes used for adapter composition (Section 2.2) are the mean ICD-10 chapter distributions of each cohort's training patients, computed from pre-cutoff diagnoses.

Evolutionary prompt optimization. For each of the eight cohorts we evolve a natural-language *clinical guidance* block, prepended to a fixed task instruction ("predict the next n ICD-10 codes for this patient"), using GEPA (Agrawal et al., 2026), a reflective prompt-evolution framework; only the guidance block evolves. We initialize a population of $P{=}8$ candidates from four trajectory-specific seed templates that encode meta-strategy (patient specificity, disease progression, ICD-version matching) but no disease-specific knowledge, then run $T{=}15$ iterations per cohort. Each iteration samples a parent with weight proportional to the number of validation patients on which it achieves population-maximum reward, following GEPA's per-instance Pareto frontier, which preserves prompts that excel on specific patient subgroups. A reflection LLM (Qwen2.5-7B-Instruct) inspects the parent's lowest-scoring training cases, diagnoses the reasoning failure, and emits a mutated guidance string. Fitness on every set is per-patient set-F1 over Agency for Healthcare Research and Quality (AHRQ) Clinical Groups. Evolution terminates on a no-improvement patience of 8 iterations or budget exhaustion; the output for each cohort is the highest-scoring Pareto-frontier candidate.

2.2 PATIENT-LEVEL CALIBRATION

Evolved clinical reasoning applies the same logic to every patient within a specialty. A cardiac patient with stable hypertension and a cardiac patient with acute decompensating heart failure receive identical clinical guidance. Closing this gap requires patient-level adaptation, but full fine-tuning per patient is computationally infeasible. We address this through micro-adapters composed via trajectory-conditioned dynamic routing.

Dynamic routing. At inference, each patient's adapter parameters are composed on-the-fly as a continuous mixture of the eight specialty micro-adapters, with mixture weights derived from the patient's own diagnosis history. We summarize the pre-cutoff ICD-10 trajectory as a chapter-frequency distribution $\mathbf{h}$ and measure its cosine similarity to each cohort prototype $\bar{\mathbf{h}}_s$—the average chapter distribution of cohort s's training patients. Similarities pass through a temperature-scaled softmax ($\tau{=}0.5$) to produce mixture weights w_s, and the patient's adapter parameters are composed as $\mathbf{v}_{\text{patient}} = \sum_{s=1}^{8} w_s \mathbf{v}_s$.

Because w_s varies continuously with $\mathbf{h}$, the eight cohort adapters define a mixture simplex over which each patient lands at a trajectory-determined point: a patient with predominantly cardiac history and emerging renal findings blends the cardiac and renal–genitourinary adapters in proportion to the chapter signal that patient actually presents, rather than being hard-assigned to a single cohort. The

78,400 trainable parameters are amortized across the entire population via the shared TinyLoRA basis.

Micro-adapter architecture. Each specialty adapter applies a TinyLoRA (Morris et al., 2026) update to the frozen backbone. TinyLoRA builds on LoRA-XS (Bałazy et al., 2025), which constrains the LoRA update to the principal subspace of each weight matrix: for $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with truncated rank-r SVD $W \approx U\Sigma V^\top$ ($U \in \mathbb{R}^{d_{\text{out}} \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$ diagonal, $V \in \mathbb{R}^{d_{\text{in}} \times r}$), LoRA-XS learns a free mixing matrix $R \in \mathbb{R}^{r \times r}$ giving $\Delta W = U\Sigma R V^\top$, with (U, Σ, V) frozen. TinyLoRA further parameterizes R through a fixed random projection tensor $P \in \mathbb{R}^{u \times r \times r}$ and a trainable vector $\mathbf{v} \in \mathbb{R}^u$:

$$R(\mathbf{v}) \;=\; \sum_{i=1}^{u} v_i \, P_i, \qquad \Delta W \;=\; U \, \Sigma \, R(\mathbf{v}) \, V^\top. \tag{1}$$

Only $\mathbf{v}$ is learned per module; the SVD factors and P are frozen. We sample P once with a fixed seed and *share it across all specialties* (though not across modules within a specialty), so that the trainable vectors $\mathbf{v}_s$ live in a common parameter basis. This shared basis is what makes the dynamic routing well-defined: the mixture $\mathbf{v}_{\text{patient}} = \sum_s w_s \mathbf{v}_s$ corresponds to a genuine convex combination of specialty updates rather than an arbitrary sum of vectors from incompatible projections.

Training. The adapter is trained with the GEPA-evolved specialty prompt active in the input context, so the adapter's contribution is measured as the marginal gain on top of GEPA rather than as an independent component. Each specialty adapter is trained independently via group relative policy optimization (Shao et al., 2024) for 500 steps with batch size 8 and $K{=}4$ rollouts per patient. The reward uses a graded matching scheme: each predicted code receives a score of 1.0 for an exact ICD-10 match, 0.5 for a same-chapter match (correct first letter), and 0.1 for a same-type (diagnosis) match. The advantage is standardized across the full batch rather than within each prompt's rollout group, preserving cross-patient difficulty signal. Generation temperature anneals from 1.0 to 0.7 over training steps. For training and inference, we use one NVIDIA B200 GPU.

2.3 Inference

At prediction time, the patient's history yields routing weights over the 8 specialty adapters; these compose a patient-specific adapter, and the dominant specialty's evolved prompt frames the input. The frozen LLM with the composed adapter generates ICD-10 codes at temperature 0.7, requiring approximately 6 GB of VRAM.

3 Related Work

EHR foundation models. Next-encounter diagnosis prediction is a core task in clinical informatics (Mukherjee et al., 2023). Attention-based RNNs (Choi et al., 2017; Ma et al., 2017) gave way to transformer and generative architectures (Li et al., 2020; Yang et al., 2023; Kraljevic et al., 2024; Shmatko et al., 2025) and to EHR foundation models for general patient representation (Steinberg et al., 2021; 2024; Rasmy et al., 2021; Pang et al., 2021; 2024; Renc et al., 2024). All are trained end-to-end on clinical sequences, entangling general medical knowledge, cohort-level reasoning, and patient context in a single parameter set shared across the population. `MOSAIC` instead assigns each function to a separate mechanism at its own granularity.

LLMs for clinical prediction. Clinical LLMs such as GatorTron (Yang et al., 2022) and Med-PaLM (Singhal et al., 2023) target unstructured narratives; whether medical-domain pretraining transfers to structured longitudinal prediction is open, and Section 5 evaluates two such models at matched scale. On structured records, CPLLM (Shoham & Rappoport, 2024) and EHR-R1 (Liao et al., 2025) reformulate the task as per-disease binary classification, and EviCare (Zhang et al., 2026) conditions an LLM on retrieved deep-model evidence. `MOSAIC` treats it as multi-label set prediction, needs no auxiliary model, and keeps the backbone frozen.

Parameter-efficient adaptation and personalization. Parameter-efficient fine-tuning (Houlsby et al., 2019; Hu et al., 2022; Bałazy et al., 2025; Dettmers et al., 2023) reduces trainable parameters by orders of magnitude but adapts uniformly. Composition extends this: AdapterFusion combines task

Table 1: Results on MIMIC-IV next-visit diagnosis prediction. F1 at four diagnostic granularities with 95% bootstrap confidence intervals. Blocks group models by training exposure: trained end-to-end, fine-tuned LLMs, and frozen LLMs evaluated zero-shot. Bold marks the best point estimate in each column.

Model	Chapter F1	CCSR F1	Clin. Group F1	Code F1
MOTOR	0.112 [0.108, 0.116]	0.142 [0.136, 0.147]	0.182 [0.175, 0.189]	0.011 [0.010, 0.012]
Delphi	0.121 [0.115, 0.128]	0.152 [0.144, 0.160]	0.191 [0.182, 0.200]	0.018 [0.015, 0.021]
Foresight	0.147 [0.141, 0.152]	0.188 [0.181, 0.194]	0.256 [0.247, 0.265]	0.021 [0.019, 0.022]
CEHR-BERT	0.149 [0.143, 0.155]	0.188 [0.181, 0.195]	0.259 [0.250, 0.268]	0.021 [0.020, 0.023]
CEHR-GPT	0.151 [0.145, 0.156]	0.191 [0.184, 0.198]	0.266 [0.256, 0.275]	0.022 [0.020, 0.023]
ETHOS	0.154 [0.147, 0.159]	0.194 [0.187, 0.201]	0.271 [0.263, 0.280]	0.024 [0.023, 0.026]
CLMBR	0.156 [0.149, 0.161]	0.197 [0.190, 0.204]	0.271 [0.262, 0.281]	0.024 [0.022, 0.025]
BEHRT	0.156 [0.149, 0.162]	0.196 [0.189, 0.203]	0.273 [0.263, 0.282]	0.023 [0.022, 0.024]
Med-BERT	0.154 [0.148, 0.159]	0.195 [0.188, 0.202]	0.273 [0.263, 0.282]	0.023 [0.021, 0.024]
TransformEHR	0.210 [0.196, 0.223]	0.248 [0.238, 0.257]	0.340 [0.327, 0.352]	0.063 [0.057, 0.067]
CPLLM	0.202 [0.194, 0.209]	0.250 [0.242, 0.259]	0.334 [0.323, 0.345]	0.072 [0.065, 0.075]
EHR-R1-8B	0.198 [0.188, 0.207]	0.239 [0.229, 0.250]	0.289 [0.279, 0.300]	0.095 [0.084, 0.098]
Qwen2.5-7B-Instruct	0.232 [0.220, 0.244]	0.286 [0.273, 0.298]	0.349 [0.338, 0.361]	0.099 [0.088, 0.104]
Qwen2.5-14B-Instruct	0.236 [0.223, 0.250]	0.283 [0.266, 0.299]	0.331 [0.319, 0.342]	0.078 [0.068, 0.083]
Qwen2.5-32B-Instruct	0.239 [0.229, 0.249]	0.288 [0.276, 0.300]	0.346 [0.335, 0.357]	0.085 [0.075, 0.090]
Gemma4-31B	0.225 [0.209, 0.241]	0.265 [0.253, 0.276]	0.307 [0.297, 0.318]	0.076 [0.066, 0.081]
Meditron-7B	0.163 [0.153, 0.173]	0.202 [0.190, 0.216]	0.240 [0.229, 0.252]	0.060 [0.051, 0.065]
BioMistral-7B	0.111 [0.103, 0.119]	0.132 [0.122, 0.141]	0.161 [0.151, 0.171]	0.033 [0.025, 0.037]
MOSAIC	**0.258 [0.246, 0.270]**	**0.308 [0.297, 0.319]**	**0.377 [0.364, 0.387]**	**0.124 [0.112, 0.129]**

adapters non-destructively (Pfeiffer et al., 2021) and LoRAMoE (Dou et al., 2024) routes over LoRA experts in the lineage of sparsely-gated MoE (Shazeer et al., 2017). Separately, automatic prompt optimization (Zhou et al., 2023; Yang et al., 2024; Pryzant et al., 2023; Fernando et al., 2023; Agrawal et al., 2026) adapts behavior with zero parameter updates. Prior clinical work treats personalization as emergent from end-to-end training (Choi et al., 2017; Ma et al., 2017; Shoham & Rappoport, 2024; Rasmy et al., 2021) or learns unsupervised patient representations (Miotto et al., 2016), with no explicit inference-time mechanism. MOSAIC combines both lines: routing weights derive from the patient's diagnostic trajectory rather than learned per-token gates, and evolved prompts encode cohort-level reasoning rather than task instructions.

4 Experimental Setup

4.1 Task

Given a patient's clinical history, which we define as a chronological sequence of diagnoses, medications, and procedures, the task is to predict the set of ICD-10 diagnosis codes that will appear at the patient's next clinical encounter.

4.2 Datasets and Evaluation Protocol

We evaluate primarily on MIMIC-IV v3.1 (Johnson et al., 2023), a longitudinal EHR dataset from Beth Israel Deaconess Medical Center (2008–2022) containing 178,821 patients after preprocessing. Event types include diagnoses (ICD-9 mapped to ICD-10 via AHRQ 3-character prefix mapping, 97.8% patient-level coverage), medications, and procedures. As an additional dataset, we use the eICU Collaborative Research Database (Pollard et al., 2018), a critical-care cohort with substantially different acuity and diagnostic distribution. All baselines are evaluated on the identical patient set using the same scoring pipeline. We split patients 70/15/15 into training, validation, and test sets, with no patient in more than one split. Cohort assignment, prototypes, and routing inputs are computed from pre-cutoff history only, and the test set is used solely for final evaluation.

Table 2: Ablation on MIMIC-IV. GEPA and the adapter contribute at different diagnostic granularities.

System	Params	Chapter F1	CCSR F1	Clin. Group F1	Code F1
Qwen2.5-7B-Instruct	0	0.232 [0.219, 0.244]	0.286 [0.273, 0.298]	0.349 [0.338, 0.361]	0.099 [0.088, 0.104]
+ GEPA	0	0.252 [0.238, 0.264]	0.297 [0.285, 0.308]	0.357 [0.346, 0.368]	0.100 [0.089, 0.104]
LoRA matched-78K	78K	0.245 [0.233, 0.256]	0.290 [0.277, 0.302]	0.331 [0.319, 0.342]	0.103 [0.091, 0.108]
LoRA r=4	10M	0.261 [0.246, 0.276]	0.302 [0.290, 0.313]	0.365 [0.353, 0.375]	0.115 [0.103, 0.120]
LoRA r=8	20M	0.252 [0.238, 0.264]	0.295 [0.283, 0.305]	0.354 [0.343, 0.365]	0.116 [0.104, 0.121]
MOSAIC	78K	**0.258 [0.246, 0.270]**	**0.308 [0.297, 0.319]**	**0.377 [0.364, 0.387]**	**0.124 [0.112, 0.129]**

4.3 Baselines

We compare against 18 baselines spanning three paradigms, grouped in Table 1 by training exposure.

Deep EHR models are trained end-to-end on MIMIC-IV and include autoregressive (CLMBR (Steinberg et al., 2021), MOTOR (Steinberg et al., 2024)), masked language modeling (BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), CEHR-BERT (Pang et al., 2021)), generative (CEHR-GPT (Pang et al., 2024), TransformEHR (Yang et al., 2023), Foresight (Kraljevic et al., 2024), Delphi (Shmatko et al., 2025)), and hybrid (ETHOS (Renc et al., 2024)) architectures, with trainable parameter counts ranging from 3.92M (Foresight) to 10.87M (Delphi).

Fine-tuned LLMs include CPLLM (Shoham & Rappoport, 2024), which fine-tunes Llama-2 for per-disease binary classification, and EHR-R1 (Liao et al., 2025), a reasoning-enhanced 8B-parameter model for diagnosis classification. Both reformulate the task as independent binary decisions rather than multi-label set prediction.

Frozen LLMs receive no clinical training and are evaluated zero-shot. This group serves two purposes. It isolates what the backbone contributes before any adaptation, and it separates the two explanations that could otherwise account for MOSAIC 's performance: model scale and medical-domain pretraining. We therefore evaluate Qwen2.5-Instruct at 7B, 14B and 32B (Qwen et al., 2025) and Gemma4-31B (Team, 2026) as general-purpose models across a range of scales, and Meditron-7B (Chen et al., 2023) and BioMistral-7B (Labrak et al., 2024) as medical-domain models matched in scale to our backbone. All six are prompted identically to MOSAIC but without evolved guidance or adapters.

4.4 Evaluation Metrics

We report F1 at four levels of diagnostic granularity, from coarsest to finest: **Chapter F1** (single-letter ICD chapter, A–Z), **CCSR Body System F1** (AHRQ 22-body-system classification (Healthcare Cost and Utilization Project (HCUP), 2024)), **Clinical Group F1** (12 clinically actionable disease groups, e.g., Acute Cardiac, Diabetes/Metabolic, Renal), and **Code F1** (exact 3-character ICD-10 match). All confidence intervals are 95% bootstrap intervals computed over 1,000 resamples of the test set.

5 Results

Main results. Table 1 reports F1 at four diagnostic granularities on MIMIC-IV. MOSAIC exceeds every baseline at every granularity; paired bootstrap tests confirm these differences are statistically significant across baselines and granularities (Appendix B). The relative improvement over the strongest baseline is largest at the finest level: MOSAIC achieves a 25% relative gain at Code F1 (0.124 vs. 0.099) compared to roughly 7–8% at the three coarser granularities. The two fine-tuned LLM baselines (CPLLM at 0.072, EHR-R1-8B at 0.095) fall between TransformEHR (0.063) and frozen Qwen (0.099) at Code F1 despite being fine-tuned LLMs of similar or larger size, indicating that scaling and fine-tuning a larger LLM on this task does not reliably improve over the frozen 7B backbone.

Table 3: Results on eICU next-visit diagnosis prediction. F1 at four diagnostic granularities with 95% bootstrap confidence intervals. Blocks group models by training exposure: trained end-to-end, fine-tuned LLMs, and frozen LLMs evaluated zero-shot. Bold marks the best point estimate in each column.

Model	Chapter F1	CCSR F1	Clin. Group F1	Code F1
MOTOR	0.218 [0.210, 0.226]	0.252 [0.241, 0.262]	0.329 [0.316, 0.340]	0.087 [0.082, 0.091]
Delphi	0.163 [0.146, 0.179]	0.189 [0.168, 0.212]	0.234 [0.221, 0.245]	0.075 [0.066, 0.081]
Foresight	0.236 [0.227, 0.244]	0.282 [0.270, 0.292]	0.356 [0.343, 0.367]	0.127 [0.120, 0.133]
CEHR-BERT	0.265 [0.252, 0.276]	0.311 [0.297, 0.325]	0.378 [0.365, 0.390]	0.156 [0.148, 0.162]
CEHR-GPT	0.236 [0.224, 0.248]	0.279 [0.265, 0.294]	0.350 [0.337, 0.361]	0.116 [0.109, 0.122]
ETHOS	0.248 [0.237, 0.259]	0.288 [0.274, 0.301]	0.358 [0.345, 0.368]	0.129 [0.122, 0.135]
CLMBR	0.245 [0.234, 0.256]	0.288 [0.275, 0.300]	0.359 [0.345, 0.369]	0.138 [0.130, 0.144]
BEHRT	0.266 [0.254, 0.278]	0.312 [0.298, 0.325]	0.380 [0.367, 0.392]	0.157 [0.148, 0.163]
Med-BERT	0.271 [0.250, 0.293]	0.303 [0.290, 0.316]	0.375 [0.361, 0.386]	0.157 [0.149, 0.163]
TransformEHR	0.330 [0.301, 0.361]	0.346 [0.331, 0.362]	0.391 [0.377, 0.404]	0.277 [0.257, 0.287]
CPLLM	0.323 [0.301, 0.344]	0.342 [0.327, 0.356]	0.385 [0.372, 0.397]	0.283 [0.264, 0.294]
EHR-R1-8B	0.279 [0.259, 0.297]	0.315 [0.287, 0.338]	0.352 [0.338, 0.364]	0.273 [0.251, 0.283]
Qwen2.5-7B-Instruct	0.254 [0.233, 0.275]	0.277 [0.253, 0.299]	0.324 [0.310, 0.336]	0.256 [0.232, 0.268]
Qwen2.5-14B-Instruct	**0.343** [0.318, 0.363]	0.341 [0.315, 0.367]	0.391 [0.378, 0.405]	0.308 [0.286, 0.318]
Qwen2.5-32B-Instruct	0.340 [0.316, 0.361]	0.356 [0.328, 0.382]	0.398 [0.385, 0.411]	0.312 [0.290, 0.324]
Gemma4-31B	0.330 [0.307, 0.351]	0.352 [0.323, 0.381]	0.380 [0.366, 0.394]	0.295 [0.271, 0.307]
Meditron-7B	0.295 [0.270, 0.321]	0.293 [0.275, 0.309]	0.347 [0.331, 0.363]	0.252 [0.227, 0.265]
BioMistral-7B	0.146 [0.131, 0.159]	0.143 [0.129, 0.156]	0.168 [0.156, 0.180]	0.072 [0.059, 0.080]
MOSAIC	0.336 [0.314, 0.357]	**0.363** [0.335, 0.391]	**0.399** [0.387, 0.411]	**0.339** [0.312, 0.352]

Ablation and granularity-specific contribution. Table 2 decomposes our system's gains. Starting from the Qwen backbone, GEPA-evolved reasoning improves Chapter F1 by +0.020 and Clinical Group F1 by +0.008 but produces minimal change at Code level (+0.001). Adding the adapter on top of evolved prompts yields a further +0.006 at Chapter but +0.024 at Code. The two components contribute at different granularities: cohort-level reasoning at coarse granularity, where specialty-specific comorbidity patterns matter most, and patient-level calibration at fine granularity, where distributional correction to individual trajectories is most consequential. The adapter is trained conditional on the evolved prompt, so we report its contribution as the marginal gain on top of GEPA. Standard LoRA provides a calibration point. At a matched 78K budget it trails MOSAIC at every granularity. Scaling it to rank 4 (10M params, $128\times$ our budget) and rank 8 (20M, $256\times$) closes the gap at Chapter and CCSR but not at Clinical Group or Code, where MOSAIC remains ahead. Additional low-rank capacity therefore recovers coarse performance but not the fine-grained calibration the decomposition provides.

Efficiency and coverage. Figure 2(a) plots Clinical Group F1 against trainable parameter count. The nearest baseline in performance (TransformEHR, 0.340) requires $\sim$3.9M trainable parameters; MOSAIC reaches 0.377 with 78K, and the largest-budget deep EHR models (Delphi 10.87M, MOTOR 8.5M) score lowest. Since MOSAIC runs on a frozen 7B backbone, this measures adaptation cost rather than model size; the parameter-controlled comparison is Table 2, where every row shares that backbone. That budget does not come at the cost of vocabulary coverage. Figure 2(b) reports unique ICD-10 codes predicted across the test set: most deep EHR baselines emit 28 to 53, collapsing onto high-frequency diagnoses, and TransformEHR reaches 184. Delphi predicts 581, comparable to our 585, but at 0.018 Code F1, so breadth without calibration yields negligible precision. Matching that coverage at $7\times$ the Code F1 indicates accurate prediction across the diagnostic long tail, not merely frequent conditions.

Where the gains fall. Three stratifications indicate the gains are broad rather than concentrated. Across 25 individual conditions (Figure 3), MOSAIC produces a near-uniform contour, indicating stable discriminative power across the disease spectrum, while baselines are competitive on a handful of high-prevalence conditions and collapse on others. Across demographics (Table 4), the adapter

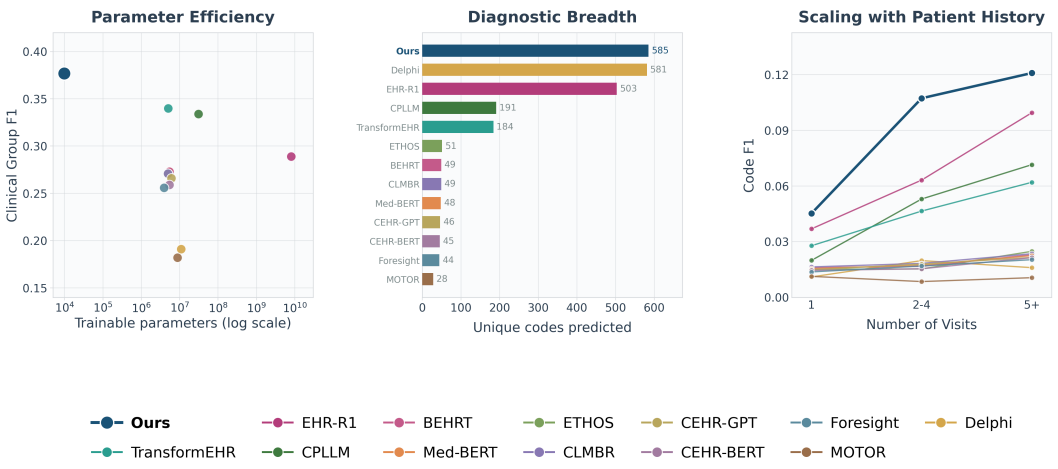

Figure 2: Analysis beyond F1. **(a)** MOSAIC achieves the highest Clinical Group F1 among trained systems with 78K trainable parameters, fewer than any trained baseline. **(b)** MOSAIC predicts 585 unique ICD-10 codes, matching Delphi's breadth while achieving 7× higher Code F1. **(c)** Code F1 improves monotonically with patient history length, while deep EHR baselines plateau after 2–4 visits.

delta under Code F1 is significantly positive on 9 of 11 strata by sex, age band and self-reported race, with no significant negative effect; the two non-significant strata are the smallest by sample size. Across history length (Figure 2(c)), Code F1 rises monotonically from approximately 0.05 with a single prior visit to 0.12 with five or more, while deep EHR baselines plateau beyond 2 to 4 visits, most remaining below 0.03 even at 5+ visits.

Replication on eICU. Table 3 evaluates the approach on eICU, a critical-care dataset with substantially different patient acuity, institutional practices and coding conventions. MOSAIC attains the highest point estimate at CCSR, Clinical Group and Code level, while Qwen2.5-14B and 32B are slightly ahead at Chapter level. The largest gain is at Code level (0.339 vs. 0.312 for the strongest baseline), indicating that the decomposition transfers to a different, multi-center patient population, most clearly at fine granularity.

6 Discussion

Three knowledge sources, two adaptation mechanisms. End-to-end EHR models entangle general medical knowledge, specialty-specific reasoning, and patient-level calibration in a single parameter set. These operate at different granularities, and our results suggest entangling them in one parameter set is wasteful. MOSAIC separates them: general medical knowledge is sourced from a frozen pretrained LLM with zero adaptation, while the two pieces that do require adaptation are handled at their natural granularity: specialty reasoning by zero-parameter evolved prompts, and patient calibration by 78K trainable parameters. The LoRA comparisons confirm this is not merely an efficiency result. At matched 78K-parameter budget, standard LoRA achieves 0.103 Code F1 versus 0.124 for MOSAIC ; even at 128–256× our budget, it still trails at Clinical Group and Code. The bottleneck in clinical adaptation is the structure of adaptation, not model capacity.

Personalization through composition. In MOSAIC , patient-specific behavior arises from composition rather than per-patient parameter updates: diagnosis history sets the routing weights, the event history sets the input context, and the dominant specialty's evolved prompt sets the reasoning frame. The results are consistent with this view. Performance scales with patient history length (Figure 2c), and the adapter improves Code F1 in nine of eleven demographic strata (Table 4). Personalization here is a property of the components' composition rather than of any single one.

Limitations and future work. Our evaluation spans two US datasets; validating on non-US healthcare systems and outpatient populations would further test generality. Because the evolved

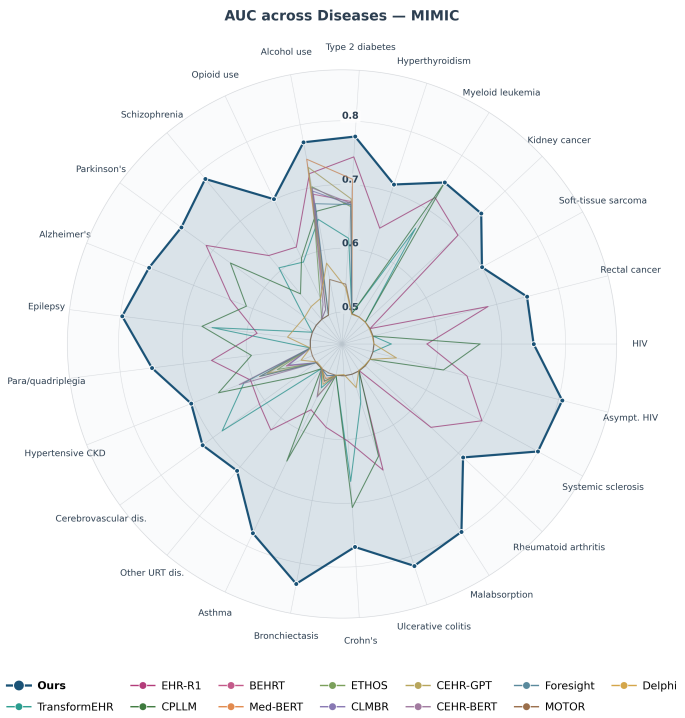

Figure 3: Per-disease AUROC on MIMIC-IV across 25 conditions. Score per (patient, disease) cell: $(K - r + 1)/K$ if the disease appears at rank r in the system's top-K predictions.

guidance is expressed in natural language, it can be directly inspected by clinicians; incorporating expert review into the evolution loop is a promising direction.

7 Conclusion

Personalized clinical prediction has been pursued primarily through scale: larger EHR foundation models, more parameters, longer training. MOSAIC argues for the opposite. Once general medical knowledge is sourced from a frozen pretrained LLM, the residual adaptation problem factors into two structurally different pieces, specialty-level reasoning and patient-level calibration, that admit qualitatively different mechanisms and together require only 78K trainable parameters to exceed end-to-end EHR models. The implication is that the bottleneck in clinical adaptation is structural, not capacitive: the parameters in current EHR foundation models are predominantly recovering knowledge a pretrained LLM already possesses, and the small fraction that remains genuinely patient-specific does not need a foundation model's worth of capacity to capture. We see this as evidence that future work on clinical personalization should invest in decomposition over scale, identifying which components of clinical reasoning are already latent in general-purpose models and adapting only what is not.

Table 4: Adapter contribution by demographic subgroup.

Axis	Level	Adapter Δ
Sex	M	**+0.018** [+0.008, +0.024]
Sex	F	**+0.032** [+0.022, +0.039]
Age	<50	**+0.019** [+0.008, +0.027]
Age	50–64	**+0.026** [+0.013, +0.031]
Age	65–79	**+0.015** [+0.008, +0.022]
Age	80+	**+0.010** [+0.004, +0.014]
Race	White	**+0.031** [+0.021, +0.038]
Race	Black	**+0.015** [+0.005, +0.021]
Race	Hispanic	+0.008 [-0.001, +0.015]
Race	Asian	+0.001 [-0.008, +0.009]
Race	Other	**+0.008** [+0.000, +0.015]

AI USE STATEMENT

Large language models are a component of the method itself, and their role is described in Sections 2.1 and 2.2. A frozen Qwen2.5-7B-Instruct backbone supplies population-level medical knowledge and generates all predictions, and a separate reflection model produces the candidate guidance blocks in the prompt-evolution loop. No guidance block is accepted on the model's assertion: candidates are selected by predictive fitness against held-out clinical data, and all reported predictions are scored against recorded next-visit codes.

We also used a general-purpose AI assistant to improve the clarity and flow of the writing. It was not used to generate research ideas, design experiments, produce or interpret results, or write the analysis code. All AI-assisted text was reviewed and edited by the authors, and every experimental result reported here was produced and verified by the authors.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning, 2026. URL https://arxiv.org/abs/2507.19457.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. Lora-xs: Low-rank adaptation with extremely small number of parameters, 2025. URL https://arxiv.org/abs/2405.17604.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, 2017. URL https://arxiv.org/abs/1608.05745.

Tim Dettmers et al. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*, 2023.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1932–1945, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.106. URL https://aclanthology.org/2024.acl-long.106/.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Jonathan G Gold, Christopher L Knight, Jennifer G Christner, Christopher E Mooney, David E Manthey, and Valerie J Lang. Clinical reasoning education in the clerkship years: a cross-disciplinary national needs assessment. *PLoS One*, 17(8):e0273250, 2022.

Healthcare Cost and Utilization Project (HCUP). Clinical classifications software refined (CCSR) for ICD-10-CM diagnoses. https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp, 2024. URL https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp. Agency for Healthcare Research and Quality (AHRQ), Rockville, MD. Version 2024.1.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2790–2799, 2019.

Edward J. Hu et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Alistair E.W. Johnson et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 2023.

Zeljko Kraljevic et al. Foresight: a generative pretrained transformer for modeling patient timelines. *Lancet Digital Health*, 2024.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

Yikuan Li et al. Behrt: Transformer for electronic health records. *Scientific Reports*, 2020.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, Jinjie Gu, Ya Zhang, Yanfeng Wang, Yu Wang, and Weidi Xie. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis, 2025. URL https://arxiv.org/abs/2510.25628.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1903–1911. ACM, August 2017. doi: 10.1145/3097983.3098088. URL http://dx.doi.org/10.1145/3097983.3098088.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1): 26094, 2016.

John X. Morris, Niloofar Mireshghallah, Mark Ibrahim, and Saeed Mahloujifar. Learning to reason in 13 parameters, 2026. URL https://arxiv.org/abs/2602.04118.

Pritam Mukherjee, Marie Humbert-Droz, Jonathan H Chen, and Olivier Gevaert. Scope: predicting future diagnoses in office visits using electronic health records. *Scientific Reports*, 13(1):11005, 2023.

Chao Pang et al. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. *ML4H*, 2021.

Chao Pang et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint*, 2024.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39/.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

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

Laila Rasmy et al. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 2021.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7 (1):256, 2024.

Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models (introduces grpo). *arXiv preprint arXiv:2402.03300*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

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

Ofir Ben Shoham and Nadav Rappoport. Cpllm: Clinical prediction with large language models, 2024. URL https://arxiv.org/abs/2309.11295.

Karan Singhal et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Ethan Steinberg, Ken Jung, et al. Language models are an effective representation learning technique for electronic health record data. In *JAMIA*, 2021.

Ethan Steinberg et al. Motor: A time-to-event foundation model for structured medical records. In *ICLR*, 2024.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers, 2024. URL https://arxiv.org/abs/2309.03409.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022. URL https://arxiv.org/abs/2203.03540.

Zhichao Yang et al. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 2023.

Hengyu Zhang, Xuyun Zhang, Pengxiang Zhan, Linhao Luo, Hang Lv, Yanchao Tan, Shirui Pan, and Carl Yang. Evicare: Enhancing diagnosis prediction with deep model-guided evidence for in-context reasoning, 2026. URL https://arxiv.org/abs/2604.10455.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers, 2023. URL https://arxiv.org/abs/2211.01910.

Table 5: Paired bootstrap differences in F1 between MOSAIC and each baseline on MIMIC-IV. For each of 1,000 bootstrap resamples of the test patient set, F1 was computed for both systems on the same resampled patients and the difference recorded. Reported intervals are 2.5th to 97.5th percentiles of paired differences. All comparisons are significant at the 95% level.

Comparator	Δ Chapter F1	Δ CCSR F1	Δ Clin. Group F1	Δ Code F1
MOTOR	+0.146 [+0.134, +0.158]	+0.166 [+0.155, +0.177]	+0.194 [+0.182, +0.206]	+0.113 [+0.101, +0.119]
Delphi	+0.137 [+0.123, +0.149]	+0.155 [+0.143, +0.168]	+0.185 [+0.172, +0.198]	+0.105 [+0.093, +0.112]
Foresight	+0.112 [+0.100, +0.124]	+0.120 [+0.108, +0.133]	+0.120 [+0.108, +0.133]	+0.103 [+0.091, +0.109]
CEHR-BERT	+0.109 [+0.097, +0.121]	+0.119 [+0.107, +0.131]	+0.117 [+0.105, +0.131]	+0.102 [+0.090, +0.108]
CEHR-GPT	+0.107 [+0.095, +0.120]	+0.117 [+0.104, +0.128]	+0.111 [+0.098, +0.123]	+0.102 [+0.090, +0.108]
ETHOS	+0.104 [+0.092, +0.117]	+0.114 [+0.101, +0.126]	+0.105 [+0.093, +0.117]	+0.099 [+0.087, +0.105]
CLMBR	+0.103 [+0.090, +0.115]	+0.111 [+0.099, +0.122]	+0.105 [+0.093, +0.117]	+0.100 [+0.088, +0.106]
BEHRT	+0.103 [+0.091, +0.115]	+0.112 [+0.099, +0.124]	+0.104 [+0.091, +0.116]	+0.101 [+0.088, +0.107]
Med-BERT	+0.105 [+0.092, +0.116]	+0.112 [+0.101, +0.124]	+0.103 [+0.092, +0.116]	+0.101 [+0.089, +0.107]
TransformEHR	+0.049 [+0.033, +0.065]	+0.060 [+0.047, +0.073]	+0.037 [+0.024, +0.051]	+0.061 [+0.050, +0.067]
CPLLM	+0.057 [+0.044, +0.070]	+0.057 [+0.045, +0.070]	+0.043 [+0.030, +0.055]	+0.052 [+0.041, +0.059]
EHR-R1-8B	+0.061 [+0.049, +0.073]	+0.068 [+0.056, +0.081]	+0.087 [+0.074, +0.099]	+0.029 [+0.020, +0.037]
Qwen2.5-7B-Instruct	+0.026 [+0.017, +0.036]	+0.022 [+0.013, +0.031]	+0.027 [+0.018, +0.036]	+0.025 [+0.018, +0.031]
Qwen2.5-14B-Instruct	+0.022 [+0.010, +0.033]	+0.025 [+0.008, +0.041]	+0.045 [+0.034, +0.056]	+0.046 [+0.035, +0.054]
Qwen2.5-32B-Instruct	+0.019 [+0.009, +0.030]	+0.020 [+0.009, +0.030]	+0.030 [+0.019, +0.040]	+0.038 [+0.029, +0.046]
Gemma4-31B	+0.033 [+0.021, +0.047]	+0.043 [+0.032, +0.054]	+0.069 [+0.058, +0.081]	+0.047 [+0.037, +0.056]
Meditron-7B	+0.095 [+0.082, +0.108]	+0.105 [+0.089, +0.119]	+0.136 [+0.123, +0.149]	+0.063 [+0.053, +0.072]
BioMistral-7B	+0.147 [+0.134, +0.159]	+0.176 [+0.162, +0.188]	+0.215 [+0.202, +0.227]	+0.091 [+0.079, +0.098]

A GEPA-EVOLVED PROMPTS

Below we excerpt representative clinical rules for 2 specialties:

Digestive specialty (emb_3):

> *1. When a patient has active GI diagnoses, predict RELATED but DISTINCT GI conditions, not the same code. A patient with K21 is already showing GI involvement. The next step is often a different GI pathology (K52) rather than repetition.*
>
> *2. Drug signatures should drive diagnosis prediction more aggressively. When a patient is receiving Metronidazole (an antibiotic targeting anaerobes/infections) along with supportive drugs like Ondansetron and Magnesium, this signals an INFECTIOUS process, not just abdominal pain.*

Oncology/hematology specialty (emb_5):

> *1. Detect oncologic trajectories and predict metastatic spread, not cardiovascular defaults. When a patient has ANY prior neoplasm code (C18, C25, C76, or any C-chapter code), the model must prioritize oncologic progression over cardiovascular predictions. Do NOT default to I10, I71, or I69 when the patient has active or recent cancer.*
>
> *2. For patients with active advanced cancer, predict symptom management and supportive care codes, not acute cardiopulmonary events. In advanced cancer patients, predict R33 (retention of urine) and Z79 (long-term drug use) rather than acute cardiopulmonary events.*

B PAIRED STATISTICAL ANALYSIS

Because all systems are evaluated on the same patients, we report paired bootstrap confidence intervals: for each of 1,000 resamples of the test patients, we compute F1 for both systems on the same resampled set and record the difference. Intervals excluding zero indicate paired significance at the 95% level. On MIMIC-IV (Table 5), MOSAIC 's gains are significant against all 18 baselines at all four granularities, with paired differences ranging from $+0.019$ (versus Qwen2.5-32B-Instruct at Chapter) to $+0.215$ (versus BioMistral-7B at Clinical Group).