

PRISM-Consult: A Panel-of-Experts Architecture for Clinician-Aligned Diagnosis

Lionel Levine, MS
University of California, Los Angeles
Los Angeles, CA 90095, USA
ORCID: 0000-0002-6926-7438

John Santerre, PhD
UC Berkeley School of Information
Berkeley, CA 94720, USA
(ORCID not available)

Alexander S. Young, MD MSHS
UCLA David Geffen School of Medicine
Los Angeles, CA 90095, USA
ORCID: 0000-0002-9367-9213

T. Barry Levine, MD
ABLE Medical Consulting
Savannah, GA 31411, USA
(ORCID not available)

Francis Campion, MD
MITRE Corporation
Bedford, MA 01730, USA
ORCID: 0000-0002-0757-9305

Majid Sarrafzadeh, PhD
University of California, Los Angeles
Los Angeles, CA 90095, USA
(ORCID not available)

Abstract—We present PRISM-Consult, a clinician-aligned panel-of-experts architecture that extends the compact PRISM sequence model into a routed family of domain specialists. Episodes are tokenized as structured clinical events; a lightweight router reads the first few tokens and dispatches to specialist models (Cardiac-Vascular, Pulmonary, Gastro-Oesophageal, Musculoskeletal, Psychogenic). Each specialist inherits PRISM’s small transformer backbone and token template, enabling parameter efficiency and interpretability. On real-world Emergency Department cohorts, specialists exhibit smooth convergence with low development perplexities across domains, while the router achieves high routing quality and large compute savings versus consult-all under a safety-first policy. We detail the data methodology (initial vs. conclusive ICD-9 families), routing thresholds and calibration, and report per-domain results to avoid dominance by common events. The framework provides a practical path to safe, auditable, and low-latency consult at scale, and we outline validation steps—external/temporal replication, asymmetric life-threat thresholds, and multi-label arbitration—to meet prospective clinical deployment standards.

Index Terms—Clinical decision support, Emergency department, Triage, Healthcare AI, Large language models, Transformers, Mixture-of-experts, Routing, Probability calibration, Natural language processing, Tokenization, ICD-9-CM, Multi-label classification, Real-time systems, Safety-critical systems, Model interpretability, Electronic health records

I. BACKGROUND

Clinical care unfolds as time-ordered sequences of events that include presenting complaints, assessing vitals, diagnostic orders, laboratory observations, and evolving diagnoses. The core challenge is to turn the earliest, sparsest portion of this sequence into reliable predictions that (1) anticipate likely next events and differentials and (2) support time-critical decision-making. Our goal is to achieve this with models that are fast, auditable, and robust to noisy early signals. In the Emergency Department (ED), where minutes matter and presentations are often ambiguous, we represent encounters as tokenized event streams and learn to predict the next event and, by extension, the patient’s diagnostic trajectory.

We previously introduced (*PRISM*) a compact, domain-focused approach to sequence modeling [1]. We focused

on the ED setting as a proving ground for sequence-based clinical decision support, emphasizing early-prefix prediction that can surface high-value differentials and guide downstream diagnostic actions. PRISM represented Emergency Department (ED) episodes as tokenized event streams and trained a small decoder-only transformer to predict the next event and, by extension, the evolving diagnostic trajectory [1]. PRISM emphasized *low-latency, interpretable modeling* by (i) restricting the event schema to diagnostics, labs, and diagnoses (excluding procedures and medications), (ii) employing a hand-designed token template (e.g., [DIAG]_ICD9_410.xx, [OBS]_LAB_TROP:HIGH, [ACTION]_ORD_ECG), (iii) capping sequence length (e.g., 512 tokens) and vocabulary size (compact custom lexicon), and (iv) utilizing a moderate-capacity backbone (e.g., 6 layers, $d_{model}=256$) suitable for clinical deployment constraints.

Initially applied to chestpain presentations in the ED, PRISM demonstrated strong calibration and efficient inference, validating the premise that carefully scoped tokenization and model size can yield clinically useful predictions without the cost or opacity of very large general-purpose models. However, the same design choices that improve efficiency, namely a streamlined vocabulary and a single-target clinical domain, also *narrow* the applicability of the model: real-world emergency department presentations often traverse multiple organ systems, and differential diagnoses for a single symptom (e.g., chest pain) span cardiac, pulmonary, gastro-oesophageal, musculoskeletal, and psychogenic causes. This motivates an extension from a single-domain backbone to a *routed panel of domain specialists*.

II. INTRODUCTION

In this paper we present **PRISM-Consult**, a routed *panel-of-experts* architecture that extends the PRISM methodology from a single-domain model to a family of specialist models orchestrated by a lightweight router. The router interprets the earliest events (symptoms and first diagnostic cues) and dispatches the episode to one or more specialist models aligned

with clinical organ systems. Each specialist inherits the PRISM tokenization template, shares embeddings for consistency, and is fine-tuned on domain-filtered cohorts labeled by definitive ICD-9 families. This preserves PRISM’s efficiency and interpretability while expanding clinical coverage.

Our contributions are threefold. First, we formalize a *clinical routing surface* based on initial ICD-9 symptom codes and map it to specialist targets defined by conclusive diagnostic families. Second, we implement a minimalist router and parameter-efficient specialist adapters that retain PRISM’s compact footprint while providing cross-domain accuracy. Third, we report partial results across cardiac, pulmonary, gastro-oesophageal, musculoskeletal, and psychogenic domains, showing smooth convergence and low development perplexities (near ~ 2.0 for all five completed specialists), thereby validating that the PRISM backbone generalizes when routed to domain-consistent corpora.

A. Why a Panel of Experts? Tokenization Trade-offs and Clinical Coverage

Its tokenization strategy sits at the core of PRISM’s efficiency. A *more specific* token inventory (e.g., separate tokens for finely binned labs, highly granular diagnoses) captures nuanced patterns and can improve in-domain accuracy and calibration. Yet specificity increases vocabulary size, fragmenting data across rare tokens and slowing inference. Conversely, a *more generic* inventory (broader bins, fewer code variants) compacts the vocabulary and speeds inference but risks losing discriminative power across heterogeneous diseases.

This creates a fundamental tension in single-model designs: optimizing the token schema for one domain (e.g., cardiology) may degrade performance in another (e.g., pulmonary embolism vs. pneumothorax), where different signals and thresholds matter. **PRISM-Consult** resolves this by *holding the token template constant*—preserving interpretability and shared embeddings—while *routing* episodes to *domain-specialized fine-tunings*. Each specialist selectively expands or emphasizes a *local* subset of tokens relevant to its domain (e.g., pleural processes, gas exchange markers for pulmonary; ischemic markers and ECG patterns for cardiac), without forcing the global vocabulary to balloon. The router thus acts as a clinical analogue of referral: “right expert, right time,” mitigating the specificity–coverage trade-off inherent to a single global model.

III. RELATED WORK

A. Clinical foundation models and domain adaptation

Foundation models trained on biomedical and clinical corpora underpin many recent advances in clinical NLP. Early domain-adapted encoders such as BioBERT [2] and ClinicalBERT [3] demonstrated that continued pretraining on PubMed/PMC and EHR notes yields substantial gains for NER, RE, QA, and risk prediction. PubMedBERT [4] further showed that training *from scratch* on biomedical text can outperform continual-pretraining approaches on multiple benchmarks. Generative clinical LMs have since emerged:

BioGPT [5] brought GPT-style pretraining to biomedical literature, while GatorTron and GatorTronGPT scaled to billions of parameters using mixed corpora that include large EHR collections, improving a wide array of downstream tasks [6], [7]. Complementing these efforts, the Med-PaLM line explored instruction tuning and medical QA evaluation at scale, reporting expert-competitive answers on MultiMedQA and related evaluations [8], [9].

Despite strong aggregate performance, deploying a single generalist model for all problems can be suboptimal in clinical operations. Differences in vocabulary, labeling practices, and decision latencies across services (e.g., cardiology vs. psychiatry) motivate domain-aligned specializations with small, auditable backbones, provided we can direct cases reliably to the right specialist. This motivates the *routed panel-of-experts* framing adopted in PRISM-Consult.

B. Mixture-of-experts and routing among specialists

Mixture-of-Experts (MoE) and conditional computation architectures enable input-dependent parameter activation and have been central to efficient scaling. GShard [10] and Switch Transformer [11] demonstrated that sparse expert routing can train trillion-parameter class models with near-constant per-token cost, provided stability and load-balancing constraints are met. Subsequent work proposed alternative routing rules (e.g., Expert Choice) to improve load balancing and efficiency [12]. In parallel, “routing among models” at inference—choosing which pre-trained system to run for a given query—has become an active area, with routers trained from preference data or heuristics to trade off cost, quality, and latency [13], [14].

Beyond general LLM systems, ensemble/agent-style methods explicitly coordinate multiple specialists. Recent “mixture-of-agents” approaches show that structured collaboration among models can improve reasoning and robustness [15]. In healthcare, most prior work focuses on single-model evaluation (e.g., LLMs for triage or diagnosis) [16], [17] or traditional clinical decision support (CDS) [18]. There remains a gap for lightweight, auditable routers that triage *early* clinical prefixes to compact, domain-specific experts, optimizing safety and compute while preserving traceability—precisely the niche that PRISM-Consult targets.

C. Positioning PRISM-Consult

PRISM-Consult operationalizes a routed specialist architecture tailored to ED presentations: (i) compact specialists (same PRISM backbone and schema) trained on domain-consistent corpora; (ii) a calibrated router that reads the first K tokens and dispatches to one or more experts under safety-first thresholds; and (iii) per-domain evaluation and macro-averaging to avoid dominance by common presentations. Compared with monolithic clinical LMs, this design aligns with operational realities (service-specific vocabulary and SLAs), supports explainability (specialist audit trails and calibrated probabilities), and can scale horizontally by adding experts or refining routing

policies, drawing on MoE and LLM-routing principles while remaining deployable in resource-limited clinical settings.

IV. METHODS

A. Overview.

PRISM-Consult extends the original PRISM framework—a compact, domain-focused sequence model over structured clinical events—into a routed, multi-specialist system. All data handling, event schema design, tokenization, and baseline modeling choices follow PRISM; below we restate those elements in full and then describe the extensions for the router and specialist models. Throughout, we reference the prior work for provenance [1] while ensuring this section is self-contained.

B. Data and Preprocessing

1) *Data Source:* Data for this study was extracted from data from the MIMIC-IV database, an extensive electronic health records repository collected from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019. The database encompasses detailed clinical information, including demographics, vital signs, laboratory test results, diagnostic procedures, and discharge diagnoses captured during hospital stays [19].

2) *Data structure and cohort window:* Each record for this model is comprised of a comprehensive longitudinal record of a patient’s diagnostic journey, across clinical episodes. Episodes begin at admission and end at patient discharge.

Timelines consist of timestamped *structured* events drawn from:

- 1) **Patient Information** (E.g., age, gender)
- 2) **Admission and Discharge Events**
- 3) **Diagnostics and orders** (e.g., ECG ordered, CTA chest ordered).
- 4) **Laboratory observations** (e.g., troponin value, D-dimer positive/negative).
- 5) **Diagnoses** encoded in ICD-9-CM.

Procedures and medication administrations are *intentionally excluded*, mirroring PRISM’s emphasis on a concise and clinically interpretable vocabulary focused on diagnostic reasoning. All events are normalized to a patient-relative clock (Initial admission = 0) and sorted chronologically and then in a structured manner detailed below to ensure consistency in batched event tokenization.

3) *Inclusion/exclusion Criteria:* Our study included all adult ED encounters with at least one qualifying initial (pre-diagnosis) symptom code recorded during the index window, along with at least one of the identified terminal diagnoses, in order to have a comprehensive set of patients traversing the diagnostic progression across clinical disciplines.

4) *Initial (Pre-Diagnosis) ICD-9-CM Code Set — Router Inputs:* Figure 1 details the set of preliminary diagnostic codes a patient presents with that are notionally indicative of a diagnosis pathway. Inclusion of at least one such code is a requisite for inclusion in the study.

Category	Symptom / Provisional De- ICD-9-CM scriptor	ICD-9-CM
Core chest-pain & cardiorespiratory	Chest pain, unspecified	786.50
	Precordial pain	786.51
	Other chest pain (incl. pleuritic)	786.52
	Shortness of breath / dyspnea	786.05
	Tachypnea	786.06
	Palpitations	785.1
Pulmonary-leaning	Syncopal and collapse	780.2
	Hemoptysis	786.3
GI-leaning	Cough	786.2
	Heartburn / pyrosis	787.1
Musculoskeletal / trauma	Nausea and vomiting	787.0
	Epigastric abdominal pain	789.06
	Abdominal pain, site unspecified	789.00
	Costochondritis / Tietze’s syn-	733.6
Psychogenic / functional	drome (provisional)	
	Contusion of chest wall	924.01
	Closed fracture of rib	807.02
	Panic disorder (episodic paroxysmal anxiety)	300.01
	Anxiety state, unspecified	300.00
	Hyperventilation syndrome	306.4

Fig. 1. Initial set of diagnoses, notionally tied to associated diagnostic grouping

Specialist	Condition Family	ICD-9-CM (incl. 4th/5th digits)
Cardiac-Vascular	Acute myocardial infarction (AMI)	410.xx
	Unstable angina / intermediate coronary syndrome	411.1
	Pericardial-myocardial inflammation (incl. pericarditis/myocarditis) ⁶	423.9
	Aortic dissection (thoracic/abdominal)	441.0x
	Hypertrophic cardiomyopathy (HCM) flare/decompensation	425.1
Pulmonary	Pulmonary embolism / infarction	415.1x
	Pneumothorax (spontaneous/tension/other)	512.8x
	Pleurisy / pleural processes	511.x
	Severe pneumonia (unspecified organism)	486
Gastro-Oesophageal	Gastro-oesophageal reflux disease (GERD)	530.81
	Diffuse oesophageal spasm	530.5
	Spontaneous oesophageal rupture (Boerhaave)	530.4
Musculoskeletal	Peptic-ulcer perforation (gastric / duodenal)	533.1x; 533.4x
	Costochondritis / Tietze’s syndrome	733.6
	Rib fracture (closed)	807.0x
Psychogenic / Functional	Chest-wall contusion	924.01
	Panic disorder	300.01
	Anxiety state, unspecified	300.00
	Hyperventilation syndrome	306.4

Fig. 2. Set of final ‘Gold Label’ diagnostic codes, representing an ultimate diagnostic determination

5) *Final (Conclusive) ICD-9-CM Code Set — Specialist Gold Labels:* Figure 2 details the set of final, or target, Gold-Label diagnostic codes representing an ultimate diagnostic determination by a clinical expert. Patients must ultimately be diagnosed with one of these conditions for inclusion in the study cohort.

6) *Sequence construction:* Once eligible patients were determined, their medical histories were extracted and converted

⁶Per prior protocol, include locally used acute pericarditis/myocarditis codes in this family and harmonize under the same training label.

to token sequences using a fixed event-precedence policy, based foremost on chronological timing and then a predetermined hierarchy of token types: `DIAG > LAB > ORDER`, with explicit time-gap markers inserted between events when inter-event intervals exceed pre-specified thresholds (e.g., 1 hour, 6 hours) to preserve chronology. A final alphabetical ordering within each token-type is then done to ensure that batched sequences are ordered consistently across episodes (this ensures consistency in prediction accuracy given the singular ‘next token’ generation schema to the PRISM model). Each sequence is truncated or windowed to a maximum length of 512 tokens for an entire patient chronology, potentially spanning multiple admission episodes. The final (gold-label) diagnosis token is *not* included on the input side for training tasks that predict it, preventing label leakage.

7) *Tokenization template and vocabulary.*: We use a compact, hand-designed token schema with type prefixes to preserve semantics:¹

- `[DIAG]_ICD9_{code}` for diagnoses (e.g., `[DIAG]_ICD9_786.50`).
- `[OBS]_LAB_{test}:{bin/value}` for labs (e.g., `[OBS]_LAB_TROP:HIGH`).²
- `[ACTION]_ORD_{order}` for diagnostic orders (e.g., `[ACTION]_ORD_ECG`).
- `[GAP]_H{k}` for time-gap markers (e.g., `[GAP]_H1` for >1 hour).
- Special sentinels: `[BOS]`, `[EOS]`, `[PAD]`, `[UNK]`.

Vocabulary growth is controlled by (i) whitelisting code families relevant to target presentations and (ii) merging infrequent variants into `[UNK]` or higher-level bins (e.g., rare lab subtypes). This yields a compact vocabulary that supports low-latency inference while maintaining clinical interpretability.

C. Model backbone (PRISM) Configuration and Training

The baseline PRISM model is a decoder-only transformer with:

- $L=6$ transformer blocks; model dimension $d_{model}=256$; MLP expansion $4d$; $n_{heads}=4$.
- Learned absolute positional embeddings; tied input/output token embeddings for parameter efficiency.
- Dropout 0.1 in attention and MLP layers; layer normalization pre-attention.

The objective is next-token prediction over the event sequence (autoregressive cross-entropy). For temporal awareness, an auxiliary time-to-next-event head (Huber loss) can be added during pre-training; this head is discarded at inference.

1) *Optimization.*: We train with AdamW (weight decay 0.01), linear warmup over the first 5% of steps followed by cosine decay to 10% of the peak learning rate. Typical settings: batch size 64–128 sequences, peak LR in $1e-3$ to $2e-4$ depending on corpus size, gradient clipping at 1.0, early

¹These templates are inherited from PRISM and reused verbatim. New domains add tokens but do not alter templates.

²Continuous labs are discretized into clinically meaningful bins (e.g., LOW/NORMAL/HIGH/CRITICAL) defined a priori clinically-validated thresholds; raw numeric values are not emitted as free-text.

stopping on dev NDCG@3. Mixed-precision training is used when supported.

2) *Evaluation (PRISM baseline)*: Primary metrics for model performance were Top-1/Top-3 next-event recall, NDCG@k, and calibration (Brier score, reliability curves). For diagnosis-prediction ablations, AUROC/PR for specific gold-label families were measured. All metrics are computed per domain and macro-averaged to avoid dominance by common events.

D. PRISM-Consult (panel of experts) Configuration and Training

PRISM-Consult’s framework extends PRISM by the expansion of the following elements:

- 1) A **router** trained on the earliest events of each episode to emit one or more domain flags (Cardiac–Vascular, Pulmonary, Gastro–Oesophageal, Musculoskeletal, Psychogenic). Inputs are the first 2–5 coded events represented as bag-of-concepts TF–IDF features projected to a 256-dim space, optionally concatenated with simple temporal features (e.g., time since triage). The router is a 2-layer transformer classifier with sigmoid outputs and focal loss to up-weight life-threatening domains.
- 2) A set of **specialist models**—one per domain—initialized from the PRISM backbone and fine-tuned on domain-filtered corpora defined by the final gold-label codes specified above. To preserve parameter sharing and inference speed, each specialist uses low-rank adaptation (LoRA) on attention and feed-forward layers while keeping shared embeddings frozen; adapter ranks and learning rates are selected on a held-out dev set per domain.
- 3) A **dispatch policy** at inference: if any life-threatening domain (e.g., AMI, PE, dissection) exceeds a calibrated threshold, a single high-priority specialist is invoked; otherwise the top-2 domains are consulted in parallel and their suggestions are merged by a deterministic arbitration layer (Cardiac > Pulmonary > Gastro > Musculoskeletal > Psychogenic).

1) *Labeling surface (initial & final codes)*: For each episode, only codes occurring *before* the first definitive diagnosis are considered “initial” to prevent leakage; when multiple definitive diagnoses are present (e.g., AMI + PE), multi-label targets are assigned.

2) *Calibration, safety, and auditability*: All classifier outputs (router and specialists) were temperature-calibrated on a development fold using cross-entropy minimization. We logged router logits, specialist logits, and the final arbitration decision per episode to enable post hoc review. If the router emits uniformly low confidence (all domain probabilities below a safety threshold) or vitals cross pre-defined danger thresholds, the system fails open to parallel consultation of all specialists.

3) *Compute and reproducibility.*: Experiments for the router were run on modern GPUs (e.g., A100-class) using deterministic seeds. Data preprocessing and training pipelines are

version-controlled; model checkpoints, tokenizer vocabularies, and ICD-9 codebooks (both initial and final sets) are archived with SHA checksums.

4) *Overview of Light-Weight Router Design:* We model early Emergency Department (ED) episodes as a prefix time-series classification task. After the first K coded events (default $K=5$), the router produces calibrated probabilities over five specialist domains: *Cardiac-Vascular*, *Pulmonary*, *Gastro-Oesophageal*, *Musculoskeletal*, and *Psychogenic*. Episodes may be multi-label in principle, though the present corpus yielded disjoint (single-label) episodes. The router dispatches to one or more PRISM-Consult specialists under a safety-first policy.

5) *Data ingestion and harmonization:* Tokenized episodes reside in five domain directories (Cardiac, Pulmonary, Gastro-Oesophageal, Musculoskeletal, Psychogenic).

Each record is mapped to a canonical schema:

episode = (episode_id, $\mathbf{e} = [e_1, \dots, e_L]$, time_feats $\in \mathbb{R}^{\leq 2}$),

where optional time_feats contain simple scalars (minutes-to-first-order; max inter-event gap). Episodes with the same episode_id across directories are merged (longest token list retained; non-empty time_feats preferred), and a 5-dim multi-hot label $\mathbf{y} \in \{0, 1\}^5$ indicates domain membership. In this study, directories were disjoint by construction, so $\|\mathbf{y}\|_0=1$.

6) *Proportional sampling:* Let $\mathcal{E} = \{(\mathbf{e}_i, \mathbf{y}_i)\}_{i=1}^N$ denote the merged episode table. To obtain a target size T while preserving domain mixture, we compute per-domain prevalence p_d on $\mathcal{E}$ and allocate quotas $q_d \approx p_d T$. We first fill quotas with domain-exclusive episodes, then top up from remaining sets (including potential multi-label cases), avoiding duplicates; if fewer than T remain after deduplication, we top up uniformly at random. Setting $T=0$ disables sampling.

7) *Prefix expansion (anytime supervision):* To enable predictions after each early event, we expand episodes into prefixes of lengths $\ell \in \{1, \dots, \min(K, L_i)\}$:

$$\mathcal{P} = \{(i, \ell, \mathbf{e}_{i,1:\ell}, \mathbf{y}_i, w_\ell = \ell/K)\}.$$

Each prefix inherits the final label $\mathbf{y}_i$ but only uses the first ℓ tokens as input, preventing label leakage. We optionally use w_ℓ as a sample weight to mildly emphasize later (more informative) prefixes.

8) *Featurization: TF-IDF $\rightarrow$ SVD:* We join prefix tokens with spaces to form a short “document” and compute 1–2-gram TF-IDF with min_df=2. A 256-dim truncated SVD yields a compact vector $\mathbf{z}_{i,\ell} \in \mathbb{R}^{256}$; optional time_feats (≤ 2 scalars) are concatenated to form $\tilde{\mathbf{z}}_{i,\ell}$.

9) *Router model and calibration:* We train one-vs-rest logistic regression heads (saga, L2, $C=2.0$, max_iter=3000) for the five domains on the prefix table $\mathcal{P}$, using stratified patient-level splits (70/10/20 train/dev/test on the indicator of any positive label). Probabilities are calibrated per head on the development split with Platt scaling (3-fold). Let $\hat{p}_d(\tilde{\mathbf{z}})$ be the calibrated probability for domain d .

10) *Routing policy and threshold tuning:* At inference, with $\hat{\mathbf{p}} = \{\hat{p}_d\}_{d=1}^5$:

- 1) If a life-threatening domain (Cardiac, Pulmonary) satisfies $\hat{p}_d \geq \tau_{hi}$, route *top-1* (the argmax).
- 2) Else if $\max_d \hat{p}_d \geq \tau_{lo}$, route *top-2*.
- 3) Else, *fail-open* to all five experts.

We grid-search (τ_{hi}, τ_{lo}) on the development split to minimize expected experts per episode subject to a safety constraint on life-threatening recall (Cardiac/Pulmonary). Unless otherwise noted, we target ≥ 0.98 on development; if no grid point satisfies the constraint, we select the point with maximal life-threat recall (tie-break: lower compute).

11) *Safety behavior:* If calibrated probabilities are uniformly low (i.e., $\max_d \hat{p}_d < 0.25$) or if triage vitals cross pre-defined danger thresholds, the router fails open to all experts and emits an audit record (raw logits, thresholds, selected route). Temperature scaling is re-checked on each model refresh to maintain probability fidelity.

12) *Evaluation: Discrimination.* ROC-AUC and PR-AUC are reported per domain on the held-out test set using calibrated probabilities.

a) *Routing quality:* With routed set R_i and truth Y_i for episode i :

$$\text{Recall}_{\text{any}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}[R_i \cap Y_i \neq \emptyset],$$

$$\text{Recall}_{\text{all}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}[Y_i \subseteq R_i].$$

Life-threat recall is computed on:

$$\{i : Y_i \cap \{\text{Cardiac, Pulmonary}\} \neq \emptyset\}$$

and requires R_i to include at least one of these domains.

b) *Compute proxy and latency:* We report $\mathbb{E}[|R_i|] = \frac{1}{N} \sum_i |R_i|$ and an estimated latency $L_i = L_{\text{router}} + \sum_{d \in R_i} L_d$ using fixed per-expert times.

c) *Anytime curves:* Metrics are stratified by prefix length $\ell = 1, \dots, K$ to quantify earliness.

d) *Anytime performance:* To quantify earliness, we repeat the above at each prefix length $\ell \in \{1, \dots, K\}$, plotting discrimination and routing recalls as functions of ℓ .

13) *Ablations and robustness checks:* We assess: (i) $K \in \{2, 3, 5\}$; (ii) text-only vs. text+time features; (iii) uncalibrated vs. calibrated probabilities; and (iv) alternative linear heads (linear SVM with Platt). As sanity baselines we compare against *consult-all*, *fixed Cardiac+Pulmonary*, and a single generalist PRISM variant (no routing).

V. RESULTS

A. Specialist Model Training & Performance

1) *Cohorts and training setup:* A total of $N=20,436$ Emergency Department (ED) episodes met all inclusion criteria for PRISM-Consult. Domain-eligible cohorts are not disjoint (multi-label episodes may appear in multiple domain pools). Where noted, some domains were *scoped* to a capped training subset to ensure compute parity and class balance. All specialists inherit the PRISM backbone and tokenization;

Domain	Eligible N	Train N	Best Epoch	Val Loss	PPL Δ	Val Loss (%)
Cardiac-Vascular	1,128	1,128	5	0.7917	2.21	-56.97
Pulmonary	3,150	3,150	5	0.7041	2.02	-52.27
Gastro-Oesophageal	9,350 (scoped 4,500)	4,500	5	0.7004	2.01	-38.03
Musculoskeletal	523	523	5	1.2692	3.56	-52.87
Psychogenic	14,924 (scoped 4,500)	4,500	5	0.6289	1.88	-29.42

Fig. 3. Development-set performance by specialist model. Perplexity is $\exp(\text{Val Loss})$. Δ Val Loss is the percentage reduction from epoch 1 to the best epoch. Scoped domains used capped training sets for balance.

optimization and early-stopping follow the protocol in the Methods section. Unless otherwise stated, training used the default causal language modeling loss of autoregressive cross-entropy) with mixed precision.

2) *Domain-specific outcomes (Specialist Model Set Performance)*: Each specialist exhibited stable convergence with steadily improving validation loss across epochs, consistent with the baseline PRISM behavior on cardiac timelines. Table V-A2 reports the best development loss, derived perplexity ($\text{PPL} = \exp(\text{loss})$), and percentage reduction from epoch 1 to the best epoch.

a) *Cardiac-Vascular (AMI/UA/pericardial/aortic/HCM)*.: Validation loss decreased from 1.8397 at epoch 1 to 0.7917 at epoch 5 (-56.97%; $\text{PPL}=2.21$), closely mirroring the original PRISM cardiac behavior and supporting the transferability of the backbone.

b) *Pulmonary (PE/pneumothorax/pleura/pneumonia)*.: Validation loss improved from 1.4753 to 0.7041 by epoch 5 (-52.27%; $\text{PPL}=2.02$), indicating a well-captured structure in early pulmonary presentations.

c) *Gastro-Oesophageal (GERD/spasm/Boerhaave/ulcer perforation)*.: On a scoped training pool of 4,500 episodes, validation loss fell from 1.1303 to 0.7004 (-38.03%; $\text{PPL}=2.01$) by epoch 5, with steady epoch-on-epoch gains and no signs of overfitting.

d) *Musculoskeletal (costochondritis/rib injury/chest-wall contusion)*.: Despite the smallest cohort ($N=523$), validation loss dropped from 2.6927 to 1.2692 (-52.87%; $\text{PPL}=3.56$). The higher perplexity is consistent with data scarcity and etiologic heterogeneity; nevertheless, the relative improvement and stable trajectory indicate learnability under limited data.

e) *Psychogenic (panic/anxiety/hyperventilation)*.: Training on a scoped subset of 4,500 episodes converged from 0.8910 to 0.6289 (-29.42%; $\text{PPL}=1.88$), the lowest perplexity among the specialists, reflecting a relatively compact symptom-to-diagnosis mapping for this domain.

3) *Cross-domain interpretation*: Figure 4 visualizes the training of all specialist models across epochs. All five specialists converge to low development perplexities (three near ~ 2.0 and one below 2.0), with smooth validation curves and substantial loss reductions from epoch 1 (Table V-A2). This *cross-disciplinary consistency*—achieved without altering tokenization, model size, or optimization hyperparameters—validates our central design choice: the compact PRISM backbone, when routed to domain-consistent corpora, yields strong and stable learning dynamics beyond its original cardiac scope. The musculoskeletal model’s higher perplexity likely

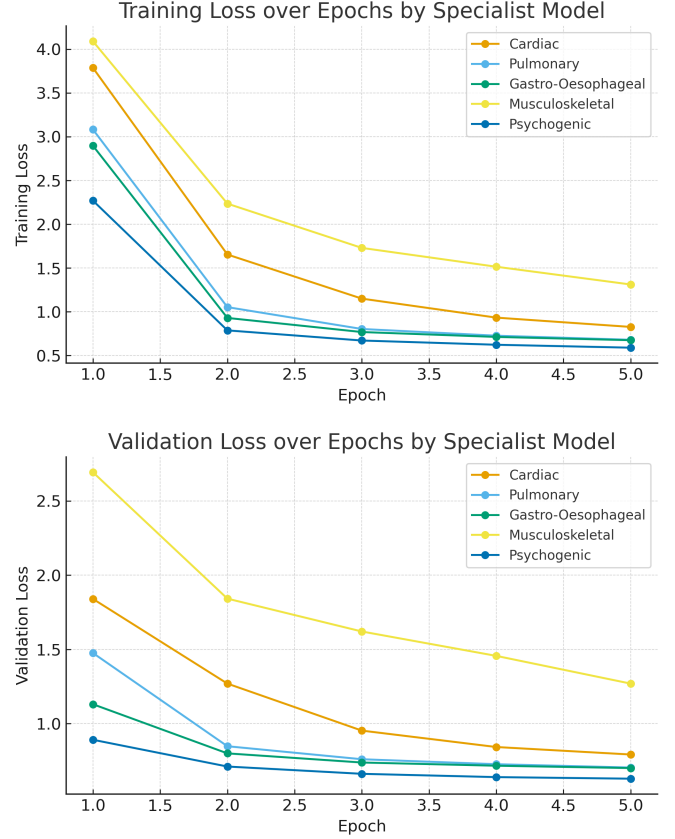


Fig. 4. Cross-specialist Training Loss over epochs for both training and validation cohorts

reflects limited sample size and broader label variance; we anticipate improvements with targeted data augmentation and modest adapter-rank increases.

4) *Training curves (dev) and calibration*: For each specialist, validation loss decreased monotonically through epoch 5 with no divergence from training loss, suggesting adequate regularization (dropout 0.1; early stopping at epoch 5). Final checkpoints correspond to the minima reported above. Temperature scaling for probabilistic outputs is fit on the development fold, with reliability plots to be included in the appendix.

B. Router Training Results

1) *Cohort and training rows*: Across domain directories we ingested $N=13,801$ unique episodes: Cardiac 1,128 (8.2%), Pulmonary 3,150 (22.8%), Gastro-Oesophageal 4,500 (32.6%), Musculoskeletal 523 (3.8%), Psychogenic 4,500 (32.6%). Episodes were single-label by design (no cross-directory duplicates). Prefix expansion with $K=5$ yielded 69,005 training/evaluation rows.

C. Cohort and Routing Policy Summary

1) *Selected thresholds and routing mix*: Fig 5. details specific thresholds and overall results mix. The development grid selected $(\tau_{hi}, \tau_{lo}) = (0.70, 0.30)$ under our objective.

	Card	Pulm	GI	MSK	Psych
Episodes (unique)	1,128	3,150	4,500	523	4,500
Share (%)	8.2	22.8	32.6	3.8	32.6
Prefixes ($K=5$)	69,005 total rows				
Selected thresholds	$\tau_{hi} = 0.70$, $\tau_{lo} = 0.30$ (dev grid)				
Dev life-threat recall	0.973 (Card $\vee$ Pulm)				
Dev recall _{any} / recall _{all}	1.000 / 0.945				
Dev expected experts	1.549 (vs. 5.0 consult-all; $\sim 69\%$ reduction)				
Dev latency (ms)	88.4 (router + routed specialists)				

Fig. 5. Light-Weight Router Training Cohort and Overall Results Summary

Domain	Dev		Test	
	AUROC	AP	AUROC	AP
Cardiac	0.984	0.985	0.964	0.958
Pulmonary	0.965	0.960	0.968	0.964
Gastro–Oesophageal	0.970	0.972	0.974	0.978
Musculoskeletal	0.939	0.910	0.951	0.936
Psychogenic	0.972	0.969	0.961	0.958
Macro-avg	0.966	0.959	0.964	0.959

Fig. 6. Domain-specific Discriminative results

With these thresholds, the router predominantly returned *top-1* or *top-2* routes; fail-open was rare. The expected consulted experts per episode on test was 1.565, a $\sim 69\%$ reduction versus consult-all (5 experts).

D. Discrimination (per-domain; macro-averaged)

Fig. 6 details domain specific discriminative results, along with macro-averaged results across domains for both the development and testing cohorts.

1) *Safety and routing quality*: On the development split, life-threatening recall (Cardiac/Pulmonary) was 0.973; on test it was 0.965. Both $Recall_{any}$ metrics were 1.000 (the routed set always included a correct domain), and $Recall_{all}$ was 0.945 (dev) and 0.942 (test). Estimated latencies, using the fixed per-expert constants, were 88.4 ms (dev) and 89.5 ms (test).

VI. DISCUSSION

A. Cross-disciplinary performance of PRISM

PRISM-Consult extends the original PRISM design from a cardiology-focused study to a routed, multi-specialist system while preserving the compact backbone and token template. Empirically, all five specialists trained with the same optimization recipe and achieved smooth, monotonic declines in development loss with low final perplexities (near ~ 2.0 for Cardiac, Pulmonary, Gastro–Oesophageal, and Psychogenic; higher but improving for Musculoskeletal). This cross-disciplinary consistency, obtained without increasing model width/depth or altering tokenization, indicates that the PRISM backbone transfers effectively when exposed to domain-consistent corpora

and calibrated with light routing. That is, a single representational space and schema can support heterogeneous organ-system reasoning as long as data partitioning and supervision align with clinical provenance.

Additionally the router’s near-perfect $Recall_{any}$ and high $Recall_{all}$ further suggest that early tokens carry enough discriminative signal to select the appropriate specialist(s) reliably, even under strict latency constraints.

B. Next steps: validation and extension

We outline three directions to harden and extend this framework:

- 1) **External and temporal validation.** Replicate training/evaluation on (i) a second site with distinct coding habits and lab panels; (ii) a temporally held-out slice to assess drift.
- 2) **Safety-first routing refinements.** Enforce a hard development constraint of life-threatening recall ≥ 0.98 via (a) asymmetric thresholds (lower τ_{hi} for Cardiac/Pulmonary than for other domains), (b) a guardrail that includes both Cardiac and Pulmonary whenever $\max(\hat{p}_{card}, \hat{p}_{pulm}) \geq \tau_{lo}^{life}$, and (c) per-head isotonic calibration when under-confidence is detected near decision boundaries. Re-report the compute–safety Pareto frontier (expected experts vs. life-threat recall).
- 3) **Broader coverage and richer inputs.** Add specialists (e.g., aortic catastrophes vs. general vascular, neurologic mimics) as new gold-label families mature; incorporate small, domain-agnostic time features (minutes-to-first-order; max inter-event gap) that are already supported by the router. Where multi-label episodes exist (e.g., AMI+PE), explicitly train/evaluate multi-label routing quality ($Recall_{all}$ at top- k) and measure arbitration behavior.

Operationally, we recommend prefix-stratified audits (metrics by $\ell = 1..K$), route-mix telemetry (top-1/top-2/fail-open), and episode-level audit logs (router/specialist logits, thresholds, decision) to support prospective QA and IRB-facing safety reviews.

C. Limits

Three limitations temper interpretation. First, the present corpus is effectively single-label across directories, so the multi-label routing regime was not strongly exercised; future cohorts should include confirmed multi-diagnosis cases to probe arbitration. Second, class imbalance (e.g., the relatively small Musculoskeletal pool) likely contributes to higher perplexity and calibration variance for that head; targeted sampling, modest adapter rank increases, and data augmentation (token normalization for near-synonymous events) are straightforward mitigations. Third, learned absolute positional embeddings fix the maximum context length and may not extrapolate; if longer horizons are needed, rotary or ALiBi encodings can be substituted with minimal disruption to the rest of the stack. Finally, threshold selection in the current run favored a near-constraint point (life-threat recall < 0.98

on test); future reports should hard-enforce prespecified safety constraints and present confidence intervals via bootstrap over episodes.

VII. CONCLUSION

PRISM-Consult operationalizes a clinician-aligned *panel-of-experts* by pairing a calibrated, light-weight router with parameter-efficient PRISM specialists trained on domain-consistent corpora. Using only the earliest tokens, the router attains high coverage and substantial compute savings relative to consult-all, while specialists exhibit smooth, low-perplexity convergence across disparate organ systems—evidence that the PRISM backbone generalizes beyond its original cardiac scope. With minor policy and calibration refinements to meet strict life-threat recall targets, the framework provides an auditable, low-latency pathway to deployable clinical decision support that routes each episode to the right expert at the right time.

REFERENCES

- [1] L. Levine, J. Santerre, *et al.*, "PRISM: A Transformer-based Language Model of Structured Clinical Event Data," *arxiv.org/abs/2506.11082*, 2025.
- [2] J. Lee, W. Yoon, S. Kim, *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *arXiv:1901.08746*, 2019.
- [3] K. Huang, A. Altosaar, D. R. R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," *arXiv:1904.05342*, 2019.
- [4] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-Specific Language Model Pretraining for Biomedical NLP," in *ACL*, 2021.
- [5] R. Luo, L. Sun, Y. Xia, *et al.*, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *arXiv:2210.10341*, 2022.
- [6] X. Yang, A. W. S. Zhang, Y. Bian, *et al.*, "A large language model for electronic health records," *npj Digital Medicine* 5, 194 (2022).
- [7] C. Peng, X. Yang, Y. Bian, *et al.*, "A study of generative large language model for medical AI," *npj Digital Medicine* 6, 205 (2023).
- [8] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge," *Nature* 620, 172–180 (2023).
- [9] K. Singhal, S. Azizi, E. K. Lai, *et al.*, "Med-PaLM 2: Towards expert-level medical question answering with LLMs," *arXiv:2305.09617*, 2023.
- [10] D. Lepikhin, H. Lee, Y. Xu, *et al.*, "GShard: Scaling giant models with conditional computation and automatic sharding," *arXiv:2006.16668*, 2020.
- [11] W. Fedus, B. Zoph, N. Shazeer, "Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv:2101.03961*, 2021.
- [12] Z. Yanqi, L. Tao, *et al.*, "Mixture-of-Experts with Expert Choice Routing," Google Research Blog (Nov. 2022), NeurIPS 2022 spotlight.
- [13] I. Ong, A. Huang, P. Lu, *et al.*, "RouteLLM: Learning to Route LLMs with Preference Data," *OpenReview*, 2024.
- [14] Z. Zhao, F. Huang, Y. Xu, *et al.*, "Eagle: Efficient Training-Free Router for Multi-LLM Inference," NeurIPS ML For Systems Workshop, 2024.
- [15] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, J. Zou, "Mixture-of-Agents Enhances Large Language Model Capabilities," *arXiv:2406.04692*, 2024.
- [16] L. Masannek, M. Günther, S. Körber, *et al.*, "Triage Performance Across Large Language Models and Healthcare Professionals," *J Med Internet Res* 26:e53297 (2024).
- [17] S. R. Ranji, A. F. Hernandez-Boussard, "Large Language Models—Misdiagnosing Diagnostic Reasoning?," *JAMA Network Open* 7(10):e2433381 (2024).
- [18] A. T. M. Wasylewicz, M. W. M. Jaspers, "Clinical Decision Support Systems," in *Health Informatics: eHealth* (NIH/NCBI Bookshelf), 2018.
- [19] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data* 10(1), 1–18 (2023). doi:10.1038/s41597-022-01899-x.