THERAWORLD: A MEDICAL WORLD MODEL IN HYPERBOLIC SPACE FOR TREATMENT-DRIVEN DISEASE TRAJECTORY SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modeling treatment-driven disease trajectories is crucial for clinical decision support. However, existing large language models typically rely on Euclidean representation spaces that poorly capture the hierarchical structure of medical concepts and the rapidly expanding state space of severe patients. We introduce TheraWorld, a medical world model that predicts future lab values and patient-state trends from historical records and observed treatment sequences in hyperbolic space. TheraWorld adapts pretrained LLMs with LoRA in hyperbolic space, and further introduces hyperbolic consistency and severity-aware radial regularization to preserve clinical structure and align representation geometry with patient severity. We construct a treatment-round-based EHR dataset and evaluate TheraWorld across multiple model scales. TheraWorld achieves the best average score of 72.10, outperforming medical LLMs and large models such as MedGemma-27B (67.30) and GLM-5.1 (69.06). Further analysis shows that the learned hyperbolic radius reflects patient severity, demonstrating the interpretability of hyperbolic geometry for LLM-based medical world modeling.

1 INTRODUCTION

Patient outcome prediction is fundamental to patient care, as it supports early risk identification and informs timely clinical intervention (Yang et al., 2025c; Ding et al., 2025). Physicians must continuously assess patient conditions and predict how they will evolve, based on a wide range of clinical evidence such as patient history, lab results, and medications. This challenge becomes especially difficult in intensive care units, where patient conditions can deteriorate rapidly and treatments are administered frequently. In such settings, predicting how a patient's condition is likely to evolve under an observed treatment sequence is essential for timely clinical decisions (Rajkomar et al., 2018; Harutyunyan et al., 2019).

Recent advances in foundation models have opened new opportunities for modeling complex clinical trajectories (Li et al., 2026; Feng et al., 2025; Li et al., 2025). However, existing EHR-based world and trajectory models (Mu et al., 2026; Adam et al., 2026) are built upon standard Transformer representations, which typically operate in Euclidean space (Yang et al., 2025b). This flat geometry may overlook two intrinsic structures of clinical data, as illustrated in Figure 1. At the token level, medical concepts follow a long-tailed frequency distribution, where a few general terms appear frequently while specialized clinical terms remain rare (Patil et al., 2025). At the patient-state level, illness severity induces an exponentially expanding state space, since critically ill patients can exhibit diverse combinations of organ dysfunctions (Nickel & Kiela, 2017; Micic & Chu, 2018). Our pilot experiments in Section 3.2 further support both properties. Consequently, Euclidean representations may compress rare medical concepts and diverse high-acuity states into limited representational space, reducing their separability.

Hyperbolic geometry provides a natural solution to both problems. Unlike Euclidean space, hyperbolic space expands exponentially with radius, which improves token-level separation by offering more capacity for rare clinical terms and alleviates patient-state crowding by allowing increasingly severe states to spread out farther from the center (Chen et al., 2024b; Yang et al., 2025b).

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.

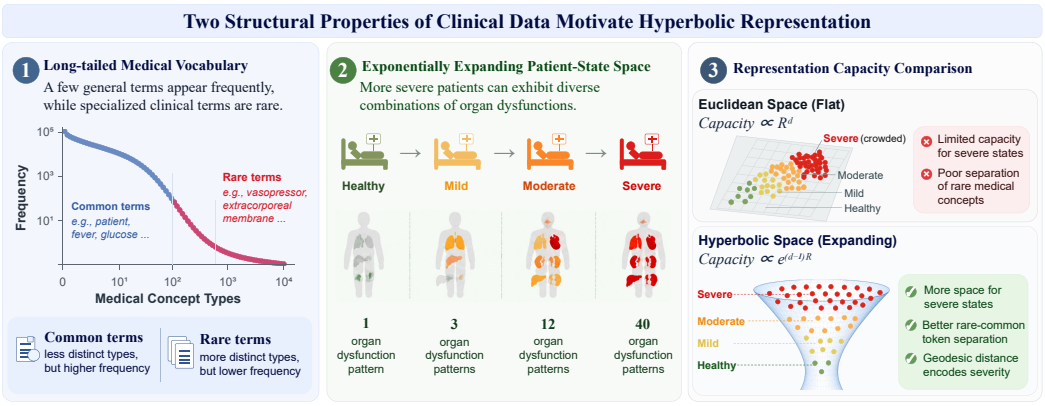


Figure 1: Hyperbolic space better matches the exponentially expanding and long-tailed structure of EHR data, alleviating crowding of severe states and improving separation between rare and common medical concepts.

Building on this insight, we propose **TheraWorld**, a hyperbolic foundation model adaptation framework built around three components. First, we apply HypLoRA (Yang et al., 2025b) to the clinical domain, injecting hyperbolic geometry into the model's attention and feed-forward adaptations. Second, we introduce hyperbolic state consistency regularization to align predicted and ground-truth next states in hyperbolic space. Third, severity-aware radial regularization links hyperbolic radius to clinical severity through radial ranking, calibration, and boundary control. TheraWorld predicts individual-level outcomes conditioned on observed treatment sequences, as it does not perform causal identification or estimate intervention effects. To evaluate this framework, we construct a treatment-driven prediction dataset from MIMIC-IV, where patient trajectories are segmented into treatment rounds based on medications and procedures. Experiments across multiple model scales show that TheraWorld achieves the strongest overall performance and learns a hyperbolic radius that reflects patient severity.

Our contributions are as follows:

- We construct a treatment-round-based EHR dataset for *medical world modeling*, where each sample uses prior history and interventions to predict next-round laboratory outcomes and severity-related patient states.
- We propose **TheraWorld**, a hyperbolic foundation model adaptation framework that combines HypLoRA, hyperbolic state consistency regularization, and severity-aware radial regularization to predict individual disease trajectories conditioned on observed treatment sequences in hyperbolic space.
- We conduct extensive experiments, showing that TheraWorld achieves the strongest overall performance, reaching the best average score of 72.10 and outperforming both medical LLMs and substantially larger models.

2 Related Work

2.1 Clinical Trajectory Models

Early clinical trajectory models treat prediction as a numeric value problem over EHR features, forecasting discrete outcomes from coded diagnoses and time-series vitals (Moen et al., 2025; Rajkomar et al., 2018; Harutyunyan et al., 2019). These approaches produce point predictions rather than full state, limiting their ability to capture the compositional nature of clinical deterioration. Subsequent work leverages LLMs for full clinical-state simulation, fine-tuning or prompting biomedical models to generate future clinical summaries from textual case histories (Makarov et al., 2025; Pellegrini et al., 2026; Adam et al., 2026). More recent efforts move simulation into latent space, using diffusion models or structured dynamics to decouple generation from surface tokens (Yang et al., 2025c; Ding et al., 2025), with explicit counterfactual intervention support (Mu et al., 2026).

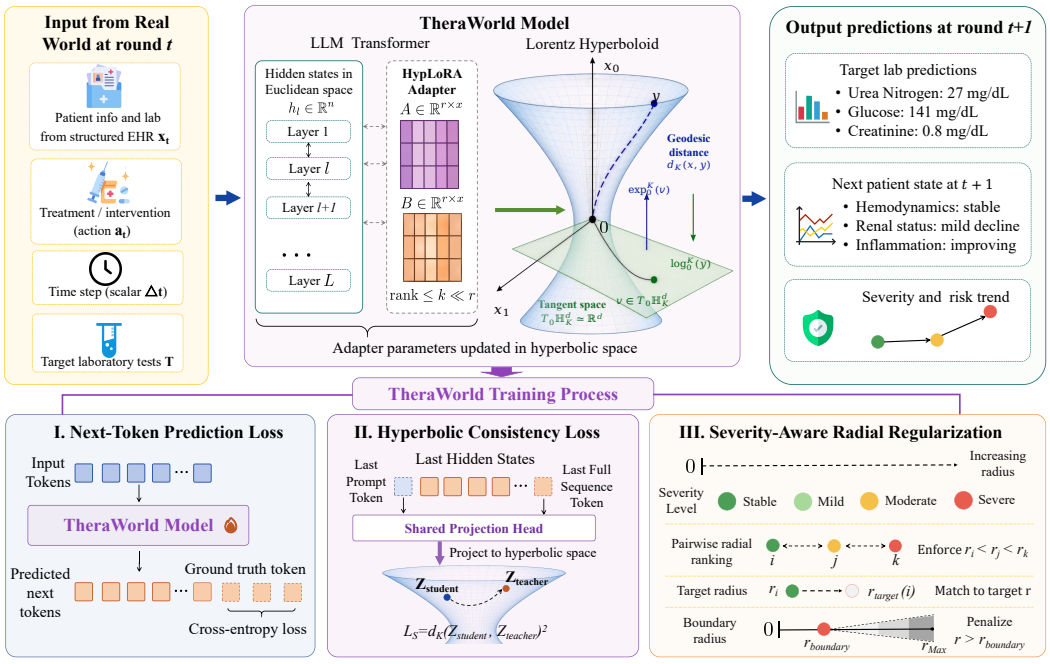


Figure 2: **Overview of TheraWorld Model**. Given structured EHR context and interventions, TheraWorld generates structured natural-language predictions of future laboratory values. HypLoRA adapts the frozen LLM by introducing hyperbolic transformations in the adapter pathway, while training jointly optimizes next-token prediction, hyperbolic state consistency, and severity-aware radial regularization to support clinical trajectory simulation.

Across these works, the latent space is implicitly Euclidean, with little attention to geometric structure or explicit representation of clinical acuity beyond final scalar severity prediction.

2.2 Geometric Representation Learning

A parallel line argues that world models should reason in latent space rather than over raw observations, and that the *geometry* of the space matters (Balestriero & LeCun, 2025; Huang et al., 2025; Chen et al., 2026). Work in this area has shown that imposing geometric structure can improve planning and state representation in general domains (Zhang et al., 2026; Chung et al., 2026), though neither approach has been adapted to clinical data. Hyperbolic geometry naturally supports hierarchical representation learning by embedding tree-like structures with low distortion (Nickel & Kiela, 2017; Micic & Chu, 2018). Prior work has shown that hyperbolic distance provides an interpretable generic-to-specific axis (Srivastava & Wu, 2025), enables hyperbolic LoRA-based LLM adaptation (Yang et al., 2025b), and helps preserve hierarchical structure while reducing representation collapse (Chen et al., 2024b).

However, these methods are all for static language tasks rather than action-conditioned dynamics, and none has been applied to clinical data.

3 Preliminary

3.1 Data Construction

Treatment-Round Partitioning. To support treatment-driven prediction, we construct a treatment-round-based EHR dataset from MIMIC-IV (Johnson et al., 2023) for *medical world modeling*. Existing EHR datasets are commonly organized by fixed time windows or sliding intervals, which are convenient for temporal modeling but do not directly correspond to clinical intervention cycles. In contrast, our goal is to predict how a patient's condition evolves after specific treatments. We therefore define round boundaries according to therapeutic intervention events, including medication

and procedures. By centering the dataset around treatment rounds rather than clock time alone, the prediction task better reflects treatment-driven disease trajectory simulation.

Table 1: Patient state diversity across severity groups. **Unique**: distinct organ-dysfunction configurations; e^H: effective number of equally common patterns; **Density**: unique patterns per 100 patients.

Severity	***N***	**Unique**	e^H	**Density**
Healthy	7,448	4	2.6	0.05
Mild	2,456	16	12.8	0.65
Moderate	1,172	33	18.5	2.82
Severe	223	40	20.7	17.94

SOFA-Based Severity Scoring. In addition to laboratory value prediction, we introduce a severity signal to support analysis of patient-state evolution. Since our work studies whether patient states can be geometrically organized by clinical severity, the dataset requires an interpretable estimate of illness severity for each round. We adopt a subset of the SOFA score (Vincent et al., 1996), using platelet count, total bilirubin, and creatinine to assess coagulation, liver, and kidney dysfunction. This SOFA-based score provides a lightweight but clinically grounded approximation of patient severity.

The final dataset is split into training, validation, and test sets, containing 10,001, 302, and 996 samples, respectively. Detailed scoring rules for the SOFA-based severity levels, together with full demographic and event statistics, are provided in Appendix A.

3.2 Pilot Experiment

We conduct pilot experiments on the above dataset to verify the structural properties of clinical data at both the token and patient-state levels. At the patient-state level, Table 1 shows that organ-dysfunction configurations become more diverse with increasing illness severity. Despite fewer severe cases, severe patients exhibit substantially more unique and effective state patterns, suggesting a rapidly expanding state space. At the token level, Figure 3 reveals a clear long-tailed distribution of medical concepts. As word frequency decreases, the number of distinct word types increases sharply from very common words to rare words. This indicates that rare clinical concepts are both numerous and sparse, requiring greater representational capacity for effective separation. Together, these findings empirically support the two structural properties motivating our method: rare medical concepts require better token-level separation, and severe patient states require more representational capacity to avoid crowding. More experiment details are provided in Appendix B.

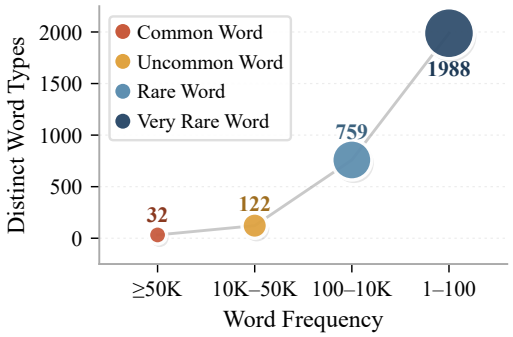


Figure 3: Distribution of distinct word types across frequency groups. The bubble size indicates the relative number of word types in each group.

3.3 Task Definition

We formulate medical world modeling as a treatment-driven patient-state prediction task, as illustrated in Figure 2. At each time step t, the model receives four inputs from the real clinical world, including the patient's current structured EHR state $\mathbf{x}_t$, the treatment or intervention action a_t, the elapsed time scalar Δt, and the set of target laboratory tests T to be predicted. TheraWorld models the transition function

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t, a_t, \Delta t, T; \theta), \tag{1}$$

where $\mathbf{x}_{t+1}$ denotes the predicted next-round patient state.

The output includes predicted laboratory values, SOFA-based next-state severity estimation, and the corresponding risk trend after treatment.

4 THERAWORLD MODEL

4.1 OVERVIEW

TheraWorld adapts a frozen pretrained LLM for treatment-driven disease trajectory modeling through hyperbolic parameter-efficient adaptation. As shown in Figure 2, the framework combines HypLoRA for injecting hyperbolic structure into the adapter pathway, hyperbolic state consistency loss for aligning predicted and ground-truth next states, and severity-aware radial regularization for organizing patient representations by clinical severity. In contrast to prior work on this task (Mu et al., 2026; Adam et al., 2026), which primarily treats disease trajectory modeling as structured numerical prediction, TheraWorld retains the free-form text generation capability of the pretrained LLM, while only the adapters, projection head, and learnable geometric scalars are trained.

4.2 HYPLORA: HYPERBOLIC LOW-RANK ADAPTATION

Standard LoRA (Hu et al., 2021) applies low-rank updates in Euclidean space, which is suboptimal for hierarchical medical concepts and severe patient states. We therefore apply HypLoRA (Yang et al., 2025b) to inject hyperbolic geometry into model adaptation. HypLoRA maps activations onto the Lorentz hyperboloid, applies low-rank updates in hyperbolic space, and maps them back to Euclidean space. We apply HypLoRA to all attention projection and feed-forward layers, with learnable curvature and tangent-norm parameters controlling the geometry at each layer. The full derivation is provided in Appendix D.

Formally, let $\mathbf{x} = (x_1, \ldots, x_T)$ denote the input prompt tokens and $\mathbf{y} = (y_1, \ldots, y_M)$ denote the target answer tokens. Since TheraWorld is trained as a text generation model, the primary task objective is the language modeling loss:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{|\mathcal{T}_{\text{ans}}|} \sum_{i \in \mathcal{T}_{\text{ans}}} \log P(y_i \mid y_{<i},\, \mathbf{x}_{\text{prompt}}) \tag{2}$$

4.3 HYPERBOLIC STATE CONSISTENCY LOSS

We introduce a hyperbolic state consistency loss as a self-supervised representation distillation objective which is demonstrated in Figure 2. During training, the teacher representation is extracted from the last token of the full output sequence, which encodes both the input context and the ground-truth next-state observations, giving it privileged access to the future patient state. The student representation is extracted from the last prompt token before generation and must anticipate this future without having observed it, which means minimizing the geodesic distance between the two representations forces the student to internalize information about the expected next patient state, effectively using the future as a self-supervised training signal. At inference time, only the student representation is used, as the ground-truth observations are unavailable. Both representations are mapped into the Lorentz hyperboloid through a shared projection head and the consistency loss penalizes their squared geodesic distance with B denoting the batch size:

$$\mathcal{L}_S = \frac{1}{B} \sum_{b=1}^{B} d_K\big(\mathbf{z}_b^S,\, \mathbf{z}_b^T\big)^2, \tag{3}$$

where $d_K(\cdot, \cdot)$ denotes the geodesic distance on the Lorentz hyperboloid with curvature K, and $\mathbf{z}_b^S$ and $\mathbf{z}_b^T$ are the projected student-state and teacher-reference embeddings, respectively. This loss encourages predicted patient-state transitions to match real clinical outcomes in hyperbolic space.

4.4 SEVERITY-AWARE RADIAL REGULARIZATION

We use the hyperbolic radius to encode clinical severity in patient-state representations. Each patient state is projected onto the Lorentz hyperboloid, where larger geodesic distances from the origin indicate higher severity. We enforce this severity-aware radial structure through a set of regularization losses based on SOFA labels.

Pairwise radial ranking loss. For each sample i, let s_i denote the previous-round severity label and r_i denote the hyperbolic radius of its representation. We construct an ordered pair set $\mathcal{P}$, where each pair (i, j) indicates that sample i is more severe than sample j. The loss is defined as:

$$\mathcal{L}_P = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max(0,\, m - (r_i - r_j)), \tag{4}$$

where m is the ranking margin. This loss encourages more severe samples to have larger radius, and penalizes pairs that violate this ordering.

Target-based radius loss. To provide a soft absolute scale, each sample is assigned a severity-dependent target radius $\hat{r}_i = r_{\min} + r_{\text{scale}} \cdot \min(s_i, s_{\max})/s_{\max}$, where $r_{\min}$ and r_{scale} define the target radius range, and $s_{\max}$ caps the severity value. The loss is:

$$\mathcal{L}_T = \frac{1}{|B_{\text{valid}}|} \sum_{i\in B_{\text{valid}}} \text{SmoothL1}(r_i - \hat{r}_i,\, \beta), \tag{5}$$

where B_{valid} is the set of samples with valid severity labels, and β is the SmoothL1 threshold. This loss encourages the learned radius r_i to stay close to its severity-dependent target $\hat{r}_i$.

Soft boundary penalty. To prevent radius from saturating near the maximum range, we introduce a soft boundary penalty:

$$\mathcal{L}_B = \frac{1}{B} \sum_{i=1}^{B} \max(0,\, r_i - r_{\text{soft}})^2, \tag{6}$$

where B is the batch size and r_{soft} is the soft radius threshold. This loss penalizes only representations whose radius exceeds r_{soft}.

4.5 Training Objective

The full training objective combines the language modeling loss with state consistency and radial regularization:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_S \cdot \mathcal{L}_S + \lambda_P \cdot \mathcal{L}_P + \lambda_T \cdot \mathcal{L}_T + \lambda_B \cdot \mathcal{L}_B \tag{7}$$

Overall, the training objective jointly optimizes predictive accuracy and geometric structure. The language modeling loss trains TheraWorld to generate next-round clinical outcomes, the hyperbolic state consistency loss aligns predicted transitions with ground-truth next-state references, and the radial regularization terms organize patient representations according to clinical severity. Through this joint optimization, TheraWorld learns a hyperbolic patient-state space that is both effective for treatment-driven prediction and interpretable in terms of disease severity.

5 Experiments

5.1 Baselines

We compare TheraWorld against a representative set of strong baselines spanning three categories. Closed-source proprietary models include GPT-5-Mini (OpenAI, 2025) and Claude-Haiku-4.5 (Anthropic, 2025); open-source general-purpose models include GLM-5.1 (GLM et al., 2024) and DeepSeek-V4-Flash (DeepSeek-AI, 2026); and open-source medical models include MedGemma-27B-IT (Sellergren et al., 2026) and Baichuan-M2-32B (Team et al., 2025). All baseline models are evaluated under their default inference configurations. A complete list of evaluated models and implementation details are provided in Appendix F.1. We further include standard LoRA (Hu et al., 2021) and SFT (Ouyang et al., 2022) applied to the same backbone models, to isolate the contribution of hyperbolic geometry from fine-tuning gains. We also compare against non-neural baselines including gradient boosted regression trees (GBRT) and random forests (RF), reported in Appendix G.6. We exclude prior disease trajectory models (Adam et al., 2026; Makarov et al., 2025; Moen et al., 2025; Mu et al., 2026) because they either assume predefined structured numerical outputs or are not open source, and thus are not readily applicable to our natural-language generation setting.

Table 2: Overall performance on the laboratory prediction task. S@k is the proportion of numerical predictions with relative error within $k\%$. Avg is calculated as the mean of S@10, S@15, S@25, and Stat-F1. ↑ indicates higher is better, and ↓ indicates lower is better. The best results are bolded and the second-best results are underlined.

Model	Model Size	S@10↑	S@15↑	S@25↑	Stat-F1↑	sMAPE↓	Avg↑
Closed-Source Models							
GPT-5-Mini	–	55.28	64.59	73.62	**87.16**	32.89	70.16
Claude-Haiku-4.5	–	56.00	65.27	75.47	86.38	31.17	70.78
Open-Source General-Purpose Models							
Qwen3-4B	4B	48.33	55.72	64.57	83.68	51.50	63.07
Qwen3-8B	8B	49.98	57.64	66.42	84.14	47.69	64.55
Qwen3-14B	14B	50.49	58.19	67.81	84.87	41.21	65.34
Qwen3.5-122B-A10B	122B	56.54	65.65	75.36	86.53	30.70	71.02
Qwen3-235B-A22B	235B	56.23	65.37	75.72	86.30	**28.85**	70.90
MiniMax-M2.7	229B	54.77	63.68	73.38	86.09	35.62	69.48
DeepSeek-V4-Flash	284B	54.32	62.78	72.54	85.60	35.75	68.81
Llama-4-Maverick-17B	400B	54.32	63.13	73.16	86.00	34.69	69.15
GLM-5.1	744B	55.08	63.37	72.45	85.34	38.25	69.06
Open-Source Medical Models							
MedGemma-4B-IT	4B	46.52	53.39	60.94	82.96	62.93	60.95
HuatuoGPT-o1-8B	8B	48.78	56.72	65.33	83.41	52.99	63.56
FineMedLM-o1	8B	52.28	62.64	72.13	81.79	38.65	67.21
Baichuan-M2-32B	32B	44.44	52.07	60.20	82.98	66.26	59.92
MedGemma-27B-IT	27B	53.15	61.07	70.00	84.97	44.75	67.30
Ours (TheraWorld)							
Qwen3-4B	4B	55.47	63.99	73.53	85.78	36.01	69.69
Qwen3-8B	8B	<u>57.40</u>	<u>66.13</u>	<u>76.28</u>	86.11	30.64	<u>71.48</u>
Qwen3-14B	14B	**58.01**	**66.73**	**76.81**	<u>86.86</u>	<u>30.14</u>	**72.10**

5.2 Evaluation

We evaluate prediction quality along two complementary dimensions. For numerical precision, we report sMAPE to measure scale-invariant regression error, and threshold-based accuracy metrics S@10, S@15, and S@25, which reflect the fraction of predictions within 10%, 15%, and 25% relative error of the ground truth. Beyond raw numerical accuracy, we report clinical status F1 (Stat-F1) for classifying predicted lab values as normal or abnormal against established clinical reference ranges. Formal definitions of all metrics are provided in Appendix E.

5.3 Implementation Details

We train three TheraWorld variants, TheraWorld-4B, TheraWorld-8B, and TheraWorld-14B, by fine-tuning Qwen3-4B, Qwen3-8B, and Qwen3-14B (Yang et al., 2025a) foundation models on our treatment-round-based EHR dataset. For API-based models, all experiments are repeated three times and the averaged results are reported. For locally deployed models, we adopt greedy decoding, so the generation results are deterministic. More details can be found in Appendix F.

Table 3: Results of different training methods on the laboratory value prediction task across Qwen3 backbones. All values are percentages. LoRA and TheraWorld use matched rank and scaling factor at each backbone scale (16/32 for Qwen3-8B and 64/128 for Qwen3-14B).

	Qwen3-8B				Qwen3-14B			
Method	**S@10↑**	**S@15↑**	**S@25↑**	**sMAPE↓**	**S@10↑**	**S@15↑**	**S@25↑**	**sMAPE↓**
Base	49.98	57.64	66.42	47.69	50.49	58.19	67.81	41.21
LoRA	56.44	65.04	74.72	34.27	55.73	64.12	73.85	38.36
SFT	<u>56.81</u>	<u>65.62</u>	<u>75.57</u>	<u>32.11</u>	<u>56.52</u>	<u>65.14</u>	<u>74.99</u>	<u>34.51</u>
TheraWorld	**57.40**	**66.13**	**76.28**	**30.64**	**58.01**	**66.73**	**76.81**	**30.14**

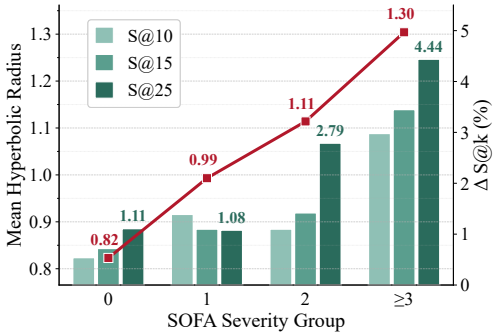


Figure 4: Hyperbolic radius and prediction gains across SOFA severity groups. The left axis corresponds to the mean hyperbolic radius of TheraWorld, shown by the red line. The right axis corresponds to ΔS@k between TheraWorld and LoRA at the 14B scale, shown by the green bars.

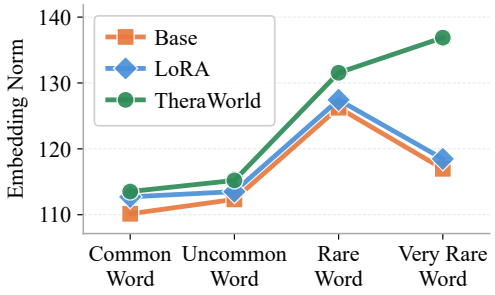


Figure 5: Last-layer hidden-state norm distributions across token frequency groups on Qwen3-8B. The y-axis denotes the ℓ_2 norm of final-layer hidden states, while the x-axis groups tokens by frequency.

5.4 Main Results

Table 2 presents the comparison between TheraWorld and the evaluated set of baselines. TheraWorld-14B achieves the best average score of 72.10, outperforming all evaluated baselines including much larger models. Among closed-source models, TheraWorld-14B surpasses both GPT-5-Mini and Claude-Haiku-4.5 despite being a fine-tuned open-source model, demonstrating that targeted hyperbolic adaptation on clinical data is more effective than general-purpose pretraining at scale. Among open-source general-purpose models, the largest evaluated systems, including Qwen3-235B-A22B and GLM-5.1, fall short of TheraWorld-14B by 1.20 and 3.04 average points respectively, confirming that TheraWorld exceeds models over 16$\times$ its parameter count. Open-source medical models, despite domain-specific training, consistently underperform general-purpose models of comparable scale: MedGemma-27B-IT achieves an average of only 67.30, and Baichuan-M2-32B scores 59.92, suggesting that standard medical fine-tuning without geometric inductive bias provides limited benefit for treatment-driven trajectory prediction.

As shown in Table 3, TheraWorld consistently and substantially outperforms both LoRA and full SFT across all model scales. Since LoRA and SFT are both Euclidean adaptation methods applied to the same backbone models with the same training data, this performance gap cannot be attributed to differences in training signal or model capacity. Rather, it directly implicates the geometric inductive bias: standard Euclidean fine-tuning, whether parameter-efficient or full, operates in a flat representation space that is structurally misaligned with the hierarchical and severity-dependent organization of clinical data. These consistent gains across scales confirm that HypLoRA and its auxiliary regularization objectives drive the improvements beyond what standard adaptation methods can achieve, and that the geometry of the latent space is itself a critical factor in treatment-driven trajectory prediction.

Table 4: Ablation study of TheraWorld on Qwen3-8B. State denotes $\mathcal{L}_S$, Ranking denotes $\mathcal{L}_P$, and Radius denotes $\mathcal{L}_T$ and $\mathcal{L}_B$. The row without HypLoRA but with all three losses active denotes the method that applies the identical supervision objectives to standard (non-hyperbolic) LoRA. Best results are bolded and second-best results are underlined.

HypLoRA	State	Ranking	Radius	S@10↑	S@15↑	S@25↑	Stat-F1↑	sMAPE↓	Avg↑
–	–	–	–	49.98	57.64	66.42	84.14	47.69	64.55
✓	–	–	–	55.07	63.41	72.94	85.67	39.07	69.27
✓	–	✓	✓	<u>55.53</u>	<u>63.86</u>	<u>73.48</u>	85.74	<u>38.10</u>	<u>69.65</u>
✓	✓	–	–	55.51	63.80	73.43	<u>85.81</u>	38.24	69.64
–	✓	✓	✓	43.98	50.51	57.90	64.85	39.77	54.31
✓	✓	✓	✓	**57.40**	**66.13**	**76.28**	**86.11**	**30.64**	**71.48**

6 ANALYSIS

6.1 PATIENT STATE LEVEL ANALYSIS

Figure 4 shows that TheraWorld's mean hyperbolic radius increases monotonically with SOFA severity, from 0.82 in the SOFA 0 group to 1.30 in the SOFA$\geq$3 group. This trend suggests that TheraWorld learns to place more severe patient states farther from the origin, making the radial distance an interpretable indicator of clinical severity. By moving these states toward regions with larger representational capacity in hyperbolic space, TheraWorld can better organize diverse high-severity patient states than Euclidean LoRA adaptation. This advantage is reflected in the prediction gains. Compared with LoRA, TheraWorld achieves increasingly larger improvements as SOFA severity grows. For example, ΔS@25 rises from $+1.11$ in the SOFA 0 group to $+4.44$ in the SOFA$\geq$3 group, with similar upward trends observed for S@10 and S@15. These results support our hypothesis that hyperbolic state modeling better represents the rapidly expanding state space of clinically complex patients, thereby yielding larger prediction gains in higher-severity cases. More experiment details are provided in Appendix G.1.

6.2 TOKEN LEVEL ANALYSIS

Figure 5 compares the last-layer hidden state norms across token frequency groups for the base model, LoRA, and TheraWorld. The base model and LoRA both peak at the Rare Word group and then decline for Very Rare Words, suggesting their limited capacity to further distinguish the most specialized terms. In contrast, TheraWorld continues to increase from Rare Words to Very Rare Words, reaching the largest norm among all models. This indicates that HypLoRA encourages the model to allocate more representational space to low-frequency clinical concepts.

6.3 ABLATION STUDY

Table 4 studies the contribution of each TheraWorld component on Qwen3-8B. HypLoRA brings a clear improvement over the base model, showing the effectiveness of hyperbolic adaptation. Adding the radial losses or the state consistency loss individually each yields further gains, indicating complementary geometric supervision. The full model achieves the best results across all metrics, indicating that jointly modeling patient-state consistency and severity-aware radius is important for accurate prediction. Critically, applying the same supervision objectives using Euclidean representation degrades performance, demonstrating that the auxiliary losses require hyperbolic geometry to be effective. More hyperparameter experiments are provided in Appendix G.3.

6.4 CASE STUDY

Figure 7 shows a representative ICU case where TheraWorld more accurately predicts next-round lab values than LoRA and DeepSeek-V4-Flash. Its hyperbolic trajectory also moves outward as the SOFA score increases from 4 to 5, indicating that the learned radius captures worsening clinical severity. This case suggests that TheraWorld improves prediction accuracy while providing an interpretable view of patient-state progression. More details are provided in Appendix G.4, and counterfactual simulations comparing original versus withheld treatment and trajectory predictions at extended time horizons are provided in Appendix G.5.

7 CONCLUSION

In this work, we investigated whether hyperbolic geometry can serve as a principled latent space for treatment-driven clinical trajectory simulation. We introduced TheraWorld, a hyperbolic foundation model adaptation framework combining HypLoRA, hyperbolic state consistency regularization, and severity-aware radial regularization. Experiments across multiple model scales show that TheraWorld achieves state-of-the-art performance, with the largest gains concentrated in the most critical patients. Furthermore, the learned hyperbolic radius reflects patient severity in a statistically significant and clinically interpretable manner, demonstrating that geometric inductive bias is not merely a representational choice but a functional driver of prediction quality. Our findings suggest that aligning latent space geometry with the intrinsic structure of clinical data is a promising and underexplored direction for building more accurate and interpretable medical world models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with cleaning and formatting our de-identified clinical dataset, and to provide feedback on research methodology and experimental design. We have not used generative AI tools to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to propose or refine hypotheses, or to interpret our experimental results, and the remaining tasks requiring disclosure are not applicable to this work. In addition, we used generative AI tools to assist with writing and debugging software code, to draft and polish portions of the manuscript, and to create or revise figures. All AI-assisted outputs were reviewed and verified by the authors. Dataset cleaning and formatting scripts were inspected against the source MIMIC-IV records, methodology feedback was evaluated by the authors before adoption, all generated code was tested for correctness, and every figure and manuscript passage was checked and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The data used in this study are derived from MIMIC-IV, which is available under its established credentialed-access and data-use requirements. We comply with the corresponding terms of use and do not redistribute the original patient records. Any released code, processed data specifications, or model artifacts will follow the applicable licenses and access restrictions of the underlying datasets and backbone models. While TheraWorld demonstrates strong predictive performance, its outputs are probabilistic approximations of clinical trajectories and should not be interpreted as ground-truth forecasts or treatment recommendations. The model is trained on ICU populations from a single institution, and predictions may not generalize to patient demographics or care practices underrepresented in MIMIC-IV. We further note that hyperbolic radius, while correlating with severity, is a learned geometric proxy and should not be conflated with validated clinical severity scores. Any application of TheraWorld beyond research settings requires prospective validation, human oversight, and compliance with local clinical and regulatory standards.

REPRODUCIBILITY STATEMENT

We provide a complete specification of our method, data, and experimental protocol to facilitate reproduction. The construction of our treatment-round-based dataset is described in Appendix A, covering the source MIMIC-IV tables, round partitioning rules, SOFA-based round selection, subject-level separation across splits, dataset statistics, and the exact training and inference prompt templates. Formal definitions of all evaluation metrics are given in Appendix E. Training configurations, including the distributed setup, optimization hyperparameters, baseline specifications, and computational budgets, are listed in Appendix F. The Euclidean control experiment, non-neural baselines, statistical significance tests, and hyperparameter sensitivity analysis are reported in Appendices G.6 and G.3. The data are derived from MIMIC-IV, which is distributed under credentialed access and cannot be redistributed. We will release our code, training scripts, and processed-data specifications upon acceptance.

REFERENCES

Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The patient is not a moving document: A world model training paradigm for longitudinal ehr, 2026. URL https://arxiv.org/abs/2601.22128.

Anthropic. Claude Haiku 4.5 model card, 2025. URL https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint Embedding Predictive Architecture

for Vision-Language. In *International Conference on Learning Representations*, 2026. URL https://mlanthology.org/iclr/2026/chen2026iclr-vljepa/.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, towards medical complex reasoning with LLMs, 2024a.

Weize Chen, Xu Han, Yankai Lin, Kaichen He, Ruobing Xie, Jie Zhou, Zhiyuan Liu, and Maosong Sun. Hyperbolic pre-trained language model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3101–3112, 2024b. doi: 10.1109/TASLP.2024.3407575.

William Youngwoo Chung, Calvin Yeung, Hansen Jin Lillemark, Zhuowen Zou, Xiangjian Liu, and Mohsen Imani. Geometric priors for generalizable world models via vector symbolic architecture, 2026. URL https://arxiv.org/abs/2602.21467.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://www.alphaxiv.org/abs/deepseek-v4.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by modeling context-aware disease trajectories in latent space. *ArXiv*, abs/2512.08029, 2025. URL https://api.semanticscholar.org/CorpusID:283712049.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 19337–19359, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3132d0405fabe24b2a7b6cd7ba9de6b5-Paper-Conference.pdf.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures, 2025. URL https://arxiv.org/abs/2509.14252.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2025. URL https://arxiv.org/abs/2405.02957.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Pony Ma, Guanhua Chen, and Heng Ji. From word to world: Can large language models be implicit text-based world models?, 2026. URL https://arxiv.org/abs/2512.18832.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P. Menden. Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, 8(1):588, 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02004-3. URL https://doi.org/10.1038/s41746-025-02004-3.

Meta AI. Llama 4: The next generation of Meta's open foundation models, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Marko Valentin Micic and Hugo Chu. Hyperbolic deep learning for chinese natural language understanding, 2018. URL https://arxiv.org/abs/1812.10408.

MiniMax AI. MiniMax M2: An efficient model for the agentic era, 2025. URL https://www.minimax.io/news/minimax-m2.

Hans Moen, Vishnu Raj, Andrius Vabalas, Markus Perola, Samuel Kaski, Andrea Ganna, and Pekka Marttinen. Towards modeling evolving longitudinal health trajectories with a transformer-based deep learning model. *Annals of Epidemiology*, 111:30–43, 2025. ISSN 1047-2797. doi: https://doi.org/10.1016/j.annepidem.2025.08.025. URL https://www.sciencedirect.com/science/article/pii/S1047279725002406.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. Ehrworld: A patient-centric medical world model for long-horizon clinical trajectories, 2026. URL https://arxiv.org/abs/2602.03569.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

OpenAI. GPT-5 system card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Sarang Patil, Zeyong Zhang, Yiran Huang, Tengfei Ma, and Mengjia Xu. Hyperbolic large language models, 2025. URL https://arxiv.org/abs/2509.05757.

Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. Ehr2path: Scalable modeling of longitudinal patient pathways from multimodal electronic health records. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2506.04831.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning

with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1. URL https://doi.org/10.1038/s41746-018-0029-1.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Sarthak Srivastava and Kathy Wu. Hypervlm: Hyperbolic space guided vision language modeling for hierarchical multi-modal understanding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2368–2379, 2025. doi: 10.1109/ICCVW69036.2025.00247.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, Jinyang Tai, Jixiang Hong, Kai Lu, Linzhuang Sun, Peidong Guo, Qian Ma, Rihui Xin, Shihui Yang, Shusen Zhang, Yichuan Mo, Zheng Liang, Zhishou Zhang, Hengfu Cui, Zuyi Zhu, and Xiaochuan Wang. Baichuan-m2: Scaling medical capability with large verifier system, 2025. URL https://arxiv.org/abs/2509.02208.

Jean-Louis Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, Camelia K. Reinhart, Peter M. Suter, and Lambert G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. doi: 10.1007/BF01709751.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Menglin Yang, Ram B, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 29613–29648. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2a711847bb13b3c55ad74853b7efd70a-Paper-Conference.pdf.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model: Generative simulation of tumor evolution for treatment planning, 2025c. URL https://arxiv.org/abs/2506.02327.

Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. FineMedLM-o1: Enhancing medical knowledge reasoning ability of LLM from supervised fine-tuning to test-time training, 2025.

Zeyu Zhang, Danning Li, Ian Reid, and Richard Hartley. Geoworld: Geometric world models, 2026. URL https://arxiv.org/abs/2602.23058.

A Dataset Construction and Statistics

A.1 Source Data and Extraction

We extract patient trajectories from MIMIC-IV by combining seven complementary tables that capture both structured clinical events and unstructured hospitalization context. Laboratory trajectories are derived from `labevents` (∼15.0M records), including test values, labels, specimen types, categories, and units. Microbiology findings are collected from `microbiologyevents` (∼336K records), which provide organism-level and antibiotic-susceptibility information. Treatment-related events are obtained from `emar` (∼16.4M records) and `procedures_icd` (∼246K records), corresponding to medication administrations and ICD-coded procedures, respectively. Finally, `patients`, `admissions`, and `discharge` are used to supplement each trajectory with demographic information, admission metadata, and discharge-summary descriptions. Table 5 summarizes the four event types used in our trajectory representation.

Table 5: Event types and representative examples in the constructed clinical trajectories.

Event Type	**Example**
`lab`	Potassium, 3.5, mEq/L
`microbiology`	URINE CULTURE, E. COLI, CIPROFLOXACIN, S
`emar`	Vancomycin, Administered
`procedure`	0DB68ZX, Excision of Stomach

A.2 Event Timeline and Round Construction

For each hospital admission, we integrate events from all source tables into a unified temporal trajectory. Events are ordered by `charttime`; for procedure events without precise timestamps, `chartdate` is used as a fallback. Each event is assigned an `event_type` label while preserving its original clinical attributes. We further attach demographic information and admission-level metadata to construct one structured longitudinal record for each admission. This initial integration yields 236,475 admission records from 122,054 patients.

To improve data quality and ensure that each trajectory contains meaningful clinical dynamics, we apply three filtering criteria. First, we remove outlier admissions with length of stay above the 90th percentile, or with total event counts outside the 10th–90th percentile range. Second, we exclude admissions without medication administration records, since such cases provide limited information about treatment-conditioned state transitions. Third, we remove admissions with fewer than four clinical segments, where a segment denotes a maximal consecutive span of observation-only rounds, intervention-only rounds, or a single mixed round. After filtering, the cohort contains 139,406 admissions from 81,065 patients.

We then normalize the event trajectories by standardizing timestamps, removing events with missing temporal information, and deduplicating repeated entries. Deduplication is performed using source-specific identifiers, which removes approximately 7.03M duplicate events.

To model treatment-conditioned patient-state evolution, we organize each event timeline into a sequence of clinical rounds. Events sharing the same timestamp are first grouped together. We then split the timeline whenever an observation follows previously observed interventions, so that post-intervention measurements are separated from the pre-intervention context. Observation events include laboratory and microbiology records, while intervention events include medication administrations and procedures. This design reflects the temporal ordering of clinical care: therapeutic actions precede subsequent physiological measurements, and the model is trained to predict subsequent observations conditioned on prior patient context and interventions.

Overall, this procedure produces approximately 860K clinical rounds, of which 93% contain valid prediction targets. In the final curated dataset, target rounds are concentrated in the early trajectory, with 80.2% appearing within rounds 1–5 and the maximum target round index reaching 19.

Table 6: SOFA component scoring criteria for the three laboratory-based organ systems used in this work.

Component	Parameter	0	1	2	3	4
$\text{SOFA}_{\text{coag}}$	Platelet (K/μL)	$\geq$150	100–149	50–99	20–49	$<$20
$\text{SOFA}_{\text{liver}}$	Bilirubin (mg/dL)	$<$1.2	1.2–1.9	2.0–5.9	6.0–11.9	$\geq$12.0
$\text{SOFA}_{\text{renal}}$	Creatinine (mg/dL)	$<$1.2	1.2–1.9	2.0–3.4	3.5–4.9	$\geq$5.0

A.3 Training Format

We formulate each training instance as a three-role chat example to align the data with instruction-tuned LLMs. The system message specifies basic patient information, including age and sex, together with the task definition. The user message provides the structured clinical context, including all historical events from previous rounds, prediction targets with test metadata but without outcome values, schema descriptions for each event section, and the SOFA-derived severity information from the previous round. Historical events are organized by time and represented as structured rows covering laboratory tests, microbiology results, medications, and procedures. The assistant message contains the target output, where laboratory and microbiology predictions are represented as structured rows grouped by time.

We further normalize label variants to reduce unnecessary surface-form ambiguity. For example, short aliases such as H, I, and L are mapped to their full forms, Hemolysis Index, Icterus Index, and Lipemia Index. Since samples without microbiology targets account for 99.3% of the training set, we remove microbiology-related instructions and fields when they are not applicable. This avoids introducing redundant empty structures and keeps the training format focused on the prediction targets present in each sample.

A.4 SOFA-Based Round Selection and Curation

We use the SOFA score (Vincent et al., 1996) as an interpretable estimate of organ dysfunction for each prediction round. Since our analysis focuses on laboratory-grounded severity, we consider three SOFA components that can be directly derived from laboratory measurements. For each retained sample, we construct an organ-state vector from the most recent laboratory observations in the history, i.e., the round immediately preceding the prediction target. Specifically, we compute SOFA component scores for coagulation, liver, and renal dysfunction using platelet count, total bilirubin, and creatinine, respectively. Each component is assigned an integer score from 0 to 4 according to the standard SOFA criteria, as shown in Table 6. The resulting organ-state vector is defined as $\mathbf{v}_t = [\text{SOFA}_{\text{coag}}, \text{SOFA}_{\text{liver}}, \text{SOFA}_{\text{renal}}]$, with a maximum aggregate score of 12. This vector provides an interpretable severity axis for both severity-conditioned evaluation and geometric analysis of patient representations.

A round is retained as a prediction target only when its immediately preceding round contains valid values for all three components, ensuring that each sample is associated with a complete severity baseline. This criterion is satisfied by 50,680 out of 663,057 candidate rounds, corresponding to approximately 7.6% of the full pipeline output.

From this candidate pool, we curate the final dataset with subject-level separation across splits to avoid patient leakage. The final dataset is split into training, validation, and test sets, containing 10,001,302, and 996 samples, respectively. Each sample includes the previous round's SOFA component values and aggregate score, providing a clinically grounded severity signal for severity-conditioned modeling and geometric analysis.

A.5 Dataset Statistics

Demographics. Table 7 summarizes the patient population. The cohort is predominantly middle-aged and elderly (51.8% aged 60+), consistent with ICU epidemiology. Each patient contributes 2.1 samples on average, and each admission contributes 1.6.

Table 7: Patient demographics (n = 8,559 with available demographics).

Dimension	Statistic
Age range	18–91 (mean 59.5, median 60)
Age 18–44	1,582 (18.5%)
Age 45–59	2,541 (29.7%)
Age 60–74	2,648 (30.9%)
Age 75+	1,788 (20.9%)
Male	4,467 (52.2%)
Female	4,092 (47.8%)

Event volume. Table 8 reports per-sample event statistics. Laboratory tests dominate the historical context, with microbiology, medication, and procedure events providing additional clinical signals organized across time groups (mean 22.0 per sample).

Lab diversity. The dataset covers **387 unique lab labels** across 3 categories and 8 body fluid sources. Category distribution: Chemistry 465,935 (57.1%), Hematology 329,379 (40.4%), Blood Gas 20,644 (2.5%). Fluid sources: Blood 792,062 (97.1%), Urine 20,372 (2.5%), with Ascites, Pleural, CSF, Joint Fluid, Other Body Fluid, and Stool comprising the remaining $\sim$3,500 (0.4%). Each sample requires prediction of **19.2 lab tests** on average (median 21, max 223).

Medication and procedure coverage. The dataset covers **1,212 unique medications**, predominantly supportive ICU care (saline flushes, heparin, insulin, analgesics). Procedure events (9,972 total) encode ICD-10 surgical interventions and contribute to round partitioning alongside medication administrations.

Data diversity summary. Coverage of 387 lab labels, 53 organisms, 26 antibiotics, and 1,212 medications across the 11,299 samples exposes models to the rich clinical patterns of real ICU environments. The broad distribution across age, sex, and severity levels ensures population representativeness. The current pipeline uses three laboratory-based SOFA components; extending to the full six-component score and incorporating additional context such as admission chief complaints is left to future work.

Table 8: Per-sample event statistics.

Event Type	Total	Mean	Median	Max
Laboratory (lab)	815,958	72.2	57	329
Microbiology	16,665	1.5	0	68
Medication (emar)	320,790	28.4	21	196
Procedure	9,972	0.9	0	14
Time groups	–	22.0	17	124

B Pilot Experiments

B.1 Patient State Level Experiment

A core hypothesis of this work is that severe patient states are not simply characterized by larger scalar SOFA scores, but also by more diverse and heterogeneous combinations of organ dysfunctions. If this hypothesis holds, it provides a principled motivation for allocating greater representational capacity to more severe states, such as through higher-rank adaptations or larger latent spaces. We examine this hypothesis on the training set ($N = 10{,}001$ patient-rounds) by quantifying organ-state pattern diversity across SOFA severity strata.

For each patient sample, we define an **organ-state pattern** as the combination of three SOFA component scores: $\mathbf{v} = (\text{SOFA}_{\text{coag}}, \text{SOFA}_{\text{liver}}, \text{SOFA}_{\text{renal}})$. Samples are grouped by their total SOFA-

based severity score, and we measure how many distinct organ-dysfunction configurations appear within each group.

Table 1 reports three complementary statistics. **Unique** denotes the number of distinct organ-state patterns observed in each severity group, directly measuring the richness of clinically distinguishable organ-dysfunction configurations. e^H denotes the effective number of commonly observed patterns, providing a distribution-aware estimate of pattern diversity. **Density** normalizes the number of unique patterns by group size, measuring the number of distinct configurations per 100 patients.

As shown in Table 1, patient-state diversity increases consistently with severity. The number of unique organ-state patterns grows from 4 in the Healthy group to 40 in the Severe group, while the effective number of patterns e^H increases from 2.6 to 20.7. After normalizing by group size, the density of distinct patterns rises even more sharply, from 0.05 to 17.94 patterns per 100 patients. These results indicate that severe patients occupy a substantially larger and more heterogeneous organ-dysfunction configuration space, supporting the need for a representation geometry with sufficient capacity to separate diverse high-acuity states.

B.2 Token Level Experiment

To examine whether clinical vocabulary exhibits a long-tailed structure, we conduct a token-level frequency analysis on the training corpus. We first collect all word types appearing in the structured clinical trajectories and count their occurrence frequencies across the dataset. Each distinct word type is then assigned to one of four frequency groups: Common Words ($\geq$50K occurrences), Uncommon Words (10K–50K occurrences), Rare Words (100–10K occurrences), and Very Rare Words (1–100 occurrences). For each group, we report the number of distinct word types, which reflects how many unique clinical terms are represented at that frequency level.

As shown in Figure 3, the vocabulary distribution is strongly long-tailed. Only 32 distinct word types appear more than 50K times, while the number of distinct types increases to 122 for uncommon words, 759 for rare words, and 1,988 for very rare words. This indicates that most clinical concepts occur infrequently, even though they account for a large portion of the vocabulary space.

This pattern has an important implication for representation learning. Frequent words are relatively few and repeatedly observed, whereas rare and very rare clinical terms are numerous, sparse, and often more specialized. A flat representation space may therefore struggle to allocate sufficient separation capacity to these low-frequency medical concepts. This motivates our use of hyperbolic adaptation, where the exponentially expanding space can provide more room for organizing rare and specialized clinical terms.

C Hyperbolic Space: A Brief Overview

Hyperbolic space is a Riemannian manifold of constant *negative* sectional curvature. Intuitively, it curves *away* from itself at every point, causing distances to stretch exponentially as one moves outward from any center. This stands in contrast to Euclidean space, which is flat and has zero curvature.

C.1 The Lorentz Hyperboloid Model

The most computationally convenient realization of hyperbolic space embeds it as the upper sheet of a hyperboloid in $\mathbb{R}^{n+1}$, equipped with the Lorentzian inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{L}} = -u_0 v_0 + \sum_{i=1}^{n} u_i v_i. \tag{8}$$

Hyperbolic space of curvature $-1/\kappa$ is then the set

$$\mathbb{H}^n_\kappa = \left\{ \mathbf{z} \in \mathbb{R}^{n+1} \mid \langle \mathbf{z}, \mathbf{z} \rangle_{\mathcal{L}} = -\kappa,\ z_0 > 0 \right\}, \tag{9}$$

with origin $\mathbf{o} = (\sqrt{\kappa}, \mathbf{0})$. The geodesic distance between two points $\mathbf{z}, \mathbf{w} \in \mathbb{H}^n_\kappa$ is

$$d_\kappa(\mathbf{z}, \mathbf{w}) = \sqrt{\kappa}\, \operatorname{arccosh}\left(-\frac{\langle \mathbf{z}, \mathbf{w} \rangle_{\mathcal{L}}}{\kappa} \right). \tag{10}$$

Points far from the origin have large z_0, and the $\operatorname{arccosh}$ grows slowly: hyperbolic space provides exponentially increasing volume as the radius grows, enabling low-distortion embeddings of tree-like structures.

C.2 Exponential and Logarithmic Maps

Since the hyperboloid is curved, one cannot add or subtract points directly. The *exponential map* $\exp_{\mathbf{z}}^{\kappa}$ moves from the flat tangent space at $\mathbf{z}$ onto the manifold, following a geodesic for arc-length $\|\mathbf{v}\|$:

$$\exp_{\mathbf{z}}^{\kappa}(\mathbf{v}) = \cosh\Big(\frac{\|\mathbf{v}\|}{\sqrt{\kappa}}\Big)\,\mathbf{z} + \sqrt{\kappa}\sinh\Big(\frac{\|\mathbf{v}\|}{\sqrt{\kappa}}\Big)\frac{\mathbf{v}}{\|\mathbf{v}\|}. \tag{11}$$

Its inverse, the *logarithmic map* $\log_{\mathbf{z}}^{\kappa}$, projects a manifold point $\mathbf{w}$ back into the tangent space at $\mathbf{z}$, returning the unique tangent vector pointing toward $\mathbf{w}$ whose norm equals $d_\kappa(\mathbf{z}, \mathbf{w})$. Together they allow Euclidean operations (e.g. linear transformations, gradient steps) to be performed locally in tangent space and then mapped back to the manifold.

C.3 Geometric Properties and Curvature

The defining feature of hyperbolic space is *exponential volume growth*: the volume of a geodesic ball of radius r scales as $e^{(n-1)r}$ rather than r^n. Geometrically, this means there is "more room" near the boundary than near the center: the space fans out faster than any polynomial.

This behavior is governed by the sectional curvature $K = -1/\kappa$. The Jacobi equation, which describes how neighboring geodesics spread apart, yields transverse deviation growing as $\cosh(\sqrt{|K|}\,t)$ in hyperbolic space, versus linearly in flat space or oscillating on the sphere. Faster geodesic spreading implies greater volume growth and thus greater capacityfor separating points.

A direct consequence visible in elementary geometry: in the hyperbolic plane, the interior angles of a geodesic triangle sum to *less* than π,

$$\alpha + \beta + \gamma = \pi - \frac{A}{\kappa}, \tag{12}$$

where A is the triangle's area. The angular deficit $\pi - (\alpha + \beta + \gamma)$ grows with triangle size: a concrete, observable signature of negative curvature that requires no ambient space to measure.

D HypLoRA: Derivation and Implementation Details

D.1 HypLoRA Forward Pass

Given a frozen linear layer with input activations $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, the HypLoRA update proceeds in three steps.

Mapping to the hyperboloid. We normalize $\mathbf{x}$ to unit direction and scale by a learned tangent norm σ, then map to $\mathbb{H}^{d_{\text{in}}}_{\kappa}$ via the exponential map at the origin:

$$\mathbf{v} = \sigma \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \qquad \mathbf{x}_{\mathcal{H}} = \exp_{\mathbf{o}}^{\kappa}(\mathbf{v}) \tag{13}$$

Low-rank transformation on the hyperboloid. A low-rank map $\mathbf{BA}$ is applied to the spatial components, and the time component is reconstructed to keep the result on the manifold:

$$\mathbf{y}_{\mathcal{H}} = \begin{bmatrix} \sqrt{\|\mathbf{BA}\,\mathbf{x}_{\mathcal{H}}^{(1:)}\|_2^2 + \kappa} \\ \mathbf{BA}\,\mathbf{x}_{\mathcal{H}}^{(1:)} \end{bmatrix}, \quad \begin{matrix} \mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}, \\ \mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}. \end{matrix} \tag{14}$$

Mapping back to Euclidean space. The adapted point is projected back via the logarithmic map at the origin:

$$\Delta\mathbf{W}_{\text{hyp}}\mathbf{x} = \log_{\mathbf{o}}^{\kappa}(\mathbf{y}_{\mathcal{H}}) \tag{15}$$

D.2 Curvature Parameterization

The curvature is parameterized as $\kappa = \text{softplus}(\phi) + \epsilon$, where ϕ is a learnable scalar initialized so that $\kappa \approx 1$. This lets the model discover the task-appropriate curvature during training while guaranteeing $\kappa > 0$.

D.3 State Consistency Projection Head

Hidden states $\mathbf{h} \in \mathbb{R}^{d_{\text{model}}}$ are projected to a 256-dimensional tangent space via $\tilde{\mathbf{h}} = \text{LayerNorm}(\mathbf{W}_{\text{proj}}\mathbf{h})$, then mapped to $\mathbb{H}^{256}_1$ via the exponential map. $\mathbf{W}_{\text{proj}}$ is initialized with small Gaussian noise ($\sigma = 0.001$) and zero bias to avoid early instability. Tangent norm clipping at $\tau = 8.0$ is applied before the exponential map for numerical stability.

D.4 Consistency Loss Warmup

The state consistency weight is linearly warmed up over $W = 30$ steps,

$$\lambda_t = \lambda_s \cdot \min(t/W, 1), \tag{16}$$

with $\lambda_s = 0.001$, to prevent gradient explosion from the randomly initialized projection head.

E Evaluation Metrics

E.1 Numerical Precision Metrics

Threshold-Based Success Rate (S@k). The success rate at threshold k measures the fraction of predictions whose relative error falls within $k\%$ of the ground truth:

$$\text{S@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\frac{|p_i - g_i|}{|g_i|} \times 100\% \leq k\right], \tag{17}$$

where $\mathbf{1}[\cdot]$ is the indicator function. We report S@10, S@15, and S@25, corresponding to increasingly permissive clinical tolerability thresholds. A prediction is considered clinically acceptable at threshold k if it approximates the true lab value within a margin that would not alter clinical interpretation. Higher values indicate better performance.

Symmetric Mean Absolute Percentage Error (sMAPE). sMAPE measures the average relative deviation between predicted and ground-truth lab values in a scale-invariant manner, making it suitable for comparing prediction quality across lab tests with vastly different numerical ranges. Formally, given N prediction–ground-truth pairs (p_i, g_i), sMAPE is defined as:

$$\text{sMAPE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|p_i - g_i|}{(|p_i| + |g_i|)/2} \times 100\%. \tag{18}$$

Unlike standard MAPE, sMAPE is symmetric with respect to over- and under-prediction, and remains well-defined when ground-truth values are close to zero. Lower values indicate better performance.

E.2 Clinical Interpretability Metrics

Clinical Status F1 (Stat-F1). Raw numerical accuracy does not capture whether a prediction correctly identifies the clinical status of a lab result, i.e., whether the predicted value falls within or outside the established normal range. To measure this clinically actionable dimension, we map each predicted and ground-truth value to a binary label $\ell \in \{0, 1\}$, where $\ell = 1$ denotes an abnormal result, using reference ranges for 23 common lab tests. Clinical Status F1 is then defined as the harmonic mean of precision and recall over the abnormal class:

$$\text{Stat-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{19}$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{20}$$

and TP, FP, and FN denote true positives, false positives, and false negatives with respect to the abnormal class, respectively. This metric captures whether pathological deviations are correctly identified, which is the clinically actionable signal most relevant to treatment decisions. Higher values indicate better performance.

F Implementation Details

F.1 Baseline Details

We evaluate against medical-domain models (MedGemma-4B-IT and MedGemma-27B-IT (Sellergren et al., 2026), HuatuoGPT-o1-8B (Chen et al., 2024a), FineMedLM-o1 (Yu et al., 2025), and Baichuan-M2-32B (Team et al., 2025)), open-source general-purpose models (Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3.5-122B-A10B, and Qwen3-235B-A22B (Yang et al., 2025a), MiniMax-M2.7 (MiniMax AI, 2025), DeepSeek-V4-Flash (DeepSeek-AI, 2026), Llama-4-Maverick-17B (Meta AI, 2025), and GLM-5.1 (GLM et al., 2024)), and proprietary models (GPT-5-Mini (OpenAI, 2025) and Claude-Haiku-4.5 (Anthropic, 2025)). We further include LoRA and full SFT applied to Qwen3-4B, Qwen3-8B, and Qwen3-14B (Yang et al., 2025a) as fine-tuning baselines.

F.2 Distributed Training.

Training is performed on 8 NVIDIA A100 GPUs using `torch.distributed.run` with the NCCL backend. The training data are sharded across GPUs with `DistributedSampler`, and both the main model and the auxiliary state-loss head are wrapped with `DistributedDataParallel`. We enable gradient checkpointing to reduce memory consumption, allowing larger models and longer input sequences to be trained at the cost of additional computation. We also use FlashAttention-2 (Dao, 2023) to improve attention efficiency during training. All experiments use bfloat16 mixed precision. The total computational budget is approximately 18 GPU-hours for the 4B model, 34 GPU-hours for the 8B model, and 62 GPU-hours for the 14B model.

F.3 Optimization

All models are trained for three epochs using the AdamW optimizer (Loshchilov & Hutter, 2019) with a base learning rate of 1×10^{-4} and zero weight decay. We use a cosine learning rate schedule with a warmup ratio of 0.1, decaying the learning rate to 0 after warmup. The HypLoRA curvature parameters, ϕ and σ, are placed in a separate optimizer group with a $10\times$ learning-rate multiplier and zero weight decay, as these geometric parameters benefit from faster adaptation. We employ gradient accumulation over 8 micro-batches of size 2, yielding an effective batch size of 16. To stabilize training, gradient clipping with a maximum norm of 1.0 is applied before each optimizer step, which helps mitigate the large initial gradient spike caused by random HypLoRA initialization combined with the language modeling loss.

G Supplement Experiments

G.1 Patient State Level Analysis

To complement the main-text results, we provide additional experiments across multiple model scales and further examine whether the learned hyperbolic geometry aligns with clinical severity and yields larger gains for more severe patients.

Geometric organization by severity. Figure 6 reports the mean hyperbolic radius of patient representations at the last prompt token, stratified by SOFA-lab3 severity group. Across all three model scales, the radius increases monotonically with SOFA severity: in the 14B model, mean radius rises from 0.823 for SOFA$= 0$ patients to 1.304 for SOFA≥ 3, a relative increase of 58.4%. This ordering is consistent across 4B and 8B variants, confirming that the geometric trend is robust to scale.

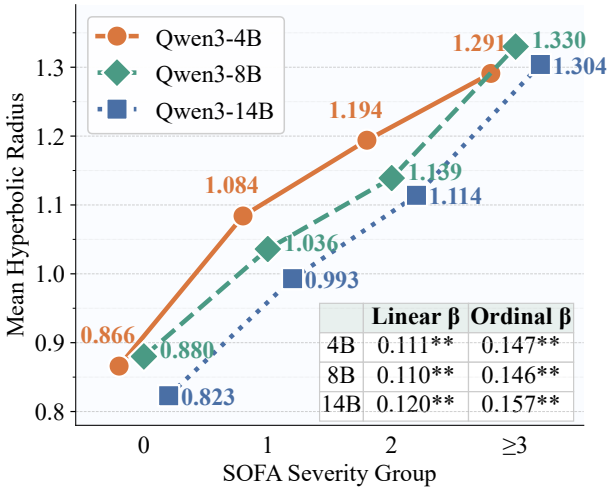


Figure 6: Mean hyperbolic radius by SOFA severity group across Qwen3 model scales. Radius increases monotonically with severity ($\beta > 0$, $p < 10^{-17}$ for all models). Inset table reports linear and ordinal regression coefficients (** $p < 0.01$).

Both linear and ordinal regression slopes are statistically significant at $p < 10^{-17}$ for all model sizes, with ordinal β slightly exceeding linear β at every scale (e.g., 0.157 vs. 0.120 for 14B), indicating that the radius–severity relationship is better captured by the ordered group structure than by a strictly linear SOFA scale. This result validates the severity-aware radial regularization: the pairwise ranking loss successfully couples the hyperbolic geometry to clinical acuity, organizing the latent space along a clinically interpretable severity axis without supervision beyond the SOFA labels already embedded in the training data.

Severity-conditioned prediction gains. Table 9 disaggregates the improvement of TheraWorld over LoRA by ground-truth SOFA group. Across model sizes and thresholds, the largest gains consistently appear in the SOFA≥ 3 group. At 14B scale, TheraWorld outperforms LoRA by $+2.98$, $+3.45$, and $+4.44$ on S@10, S@15, and S@25 respectively for the most critical patients, compared to gains of $+0.54$, $+0.72$, and $+1.11$ for the mildest group. The 4B model shows the same directional pattern even more sharply: despite slight degradation on mild and moderate patients, it achieves gains of up to $+3.77$ on the SOFA≥ 3 group, suggesting that the geometric inductive bias is especially valuable for critical states even when overall model capacity is limited.

Note on severity grouping. The severity grouping used in this analysis differs from that in the pilot experiment because the two analyses are conducted on different data scopes. The pilot experiment analyzes the full dataset, where the original severity thresholds provide sufficient samples in each group. In contrast, the analysis in Figure 6 and Table 9 is conducted on the test set only. Since the test set is much smaller, applying the same thresholds would leave very few samples in the most severe group, resulting in highly imbalanced groups and unstable estimates. We therefore adopt a coarser severity partition for the test-set analysis to ensure more reliable and statistically meaningful comparisons.

G.2 Token Level Analysis

As a more detailed counterpart to the token-level analysis in the main text, we report the same results with full group-level statistics in Table 10. Specifically, we compare the base model (Qwen3-8B), its standard LoRA-adapted variant, and TheraWorld. For each model, we process training-set samples with a full forward pass and extract the hidden states from the last Transformer layer. Tokens are grouped into four frequency-based categories: Common Word, Uncommon Word, Rare Word, and Very Rare Word. These groups reflect progressively lower token frequencies and are used to examine how different adaptation methods allocate representational capacity across frequent and long-tail

Table 9: S@k improvement over LoRA (Δ) by ground-truth SOFA severity group. Δ = TheraWorld − LoRA. Bold values indicate the severity group with the largest Δ.

	Δ by GT SOFA Group			
Model	SOFA=0	SOFA=1	SOFA=2	SOFA$\geq$3
Qwen3-4B				
S@10	−0.43	−0.18	−1.91	**+2.72**
S@15	−0.96	−0.70	−2.44	**+3.77**
S@25	−0.86	−0.85	−1.49	**+3.70**
Qwen3-8B				
S@10	+0.23	+0.22	+0.09	**+0.79**
S@15	+0.67	+0.31	+0.25	+0.29
S@25	+1.14	+0.38	+0.90	**+1.65**
Qwen3-14B				
S@10	+0.54	+1.39	+1.10	**+2.98**
S@15	+0.72	+1.10	+1.42	**+3.45**
S@25	+1.11	+1.08	+2.79	**+4.44**

Table 10: Token-level representation statistics across frequency-based word groups. Group 1 to Group 4 denote Common Word, Uncommon Word, Rare Word, and Very Rare Word, respectively, corresponding to progressively lower token frequencies and increasingly long-tail medical concepts. For each group, we report the average token frequency and the ℓ_2 norm of last-layer hidden states for the base model, LoRA, and TheraWorld.

Model	**Group**	**Frequency**	**Norm**
Base Model	1	4,258.5 [2,919∼6,231]	112.71 [99.94∼126.81]
	2	934.5 [2,82∼3,417]	113.51 [90.31∼153.75]
	3	266.7 [226∼328]	127.44 [103.56∼161.50]
	4	89.1 [1∼174]	118.48 [94.06∼147.75]
LoRA	1	4,258.5 [2,919∼6,231]	110.14 [98.31∼123.50]
	2	934.5 [282∼3,417]	112.33 [88.81∼152.75]
	3	266.7 [226∼328]	126.24 [101.69∼161.12]
	4	89.1 [1∼174]	116.97 [92.56∼144.50]
TheraWorld	1	2,956.3 [1,301∼6,231]	113.518 [98.000∼127.560]
	2	801.0 [3∼3,417]	115.211 [84.380∼161.000]
	3	135.9 [2∼1,499]	131.550 [92.120∼183.250]
	4	46.0 [46∼46]	136.880 [136.880∼136.880]

medical tokens. For each group, we report the average token frequency and the ℓ_2 norm of the corresponding last-layer hidden states.

Table 10 shows that the base model and standard LoRA exhibit similar trends. Both models assign relatively larger norms to rare medical terms in Group 3, but their norms decrease for the most specific Group 4 tokens. For example, the base model increases from 112.71 in Group 1 to 127.44 in Group 3, but drops to 118.48 in Group 4; LoRA follows the same pattern, decreasing from 126.24 in Group 3 to 116.97 in Group 4. This suggests that standard Euclidean adaptation does not consistently preserve additional representational separation for the rarest and most specialized clinical terms.

In contrast, TheraWorld exhibits a monotonic increase in hidden-state norm as token frequency decreases and medical specificity increases. Its norm rises from 113.52 in Group 1 to 115.21 in Group 2,131.55 in Group 3, and 136.88 in Group 4. This indicates that HypLoRA encourages the model to allocate larger representational capacity to low-frequency specialized clinical concepts, consistent with the motivation of using hyperbolic geometry to better separate long-tailed medical

Table 11: Hyperparameter study on Qwen3-8B for the lab prediction task. Bold indicates the best result in each column and underline the second-best, computed across all configurations listed. The $r = 16$ row corresponds to the default configuration used throughout the paper, with λ_s=0.05, λ_P=0.002, λ_B=0.0001, and λ_T=0.02. The radial loss weights are varied one at a time from this default while the remaining weights stay fixed.

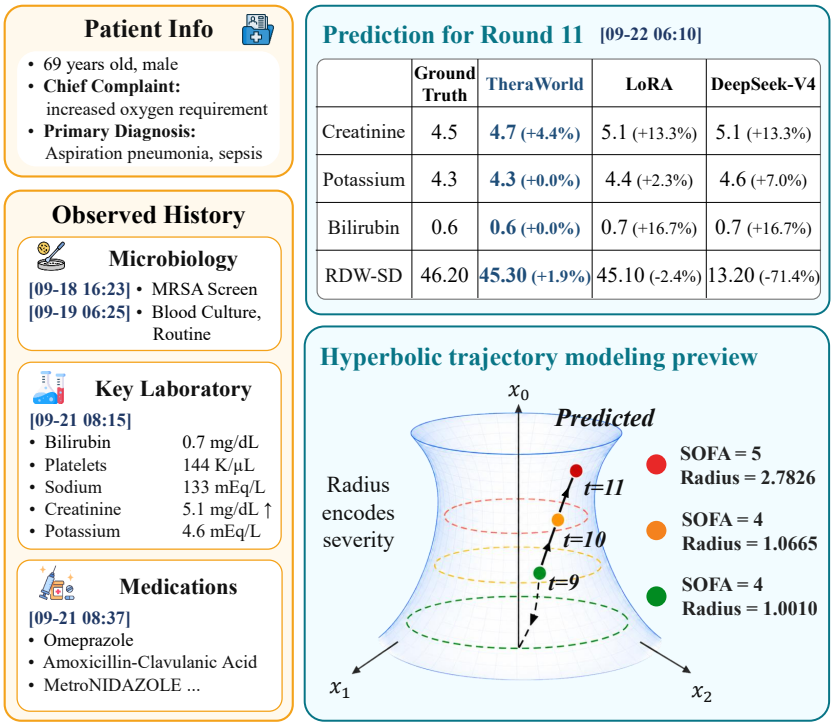
Hyperparameter	**S@10↑**	**S@15↑**	**S@25↑**	**sMAPE↓**	**Stat-F1↑**	**Avg↑**
Base (no LoRA)	49.98	57.64	66.42	47.69	84.14	64.55
LoRA Rank						
$r = 8$	57.24	65.88	76.03	30.57	86.19	71.33
$r = 16$	57.40	66.13	76.28	30.64	86.11	71.48
$r = 32$	57.57	<u>66.40</u>	76.35	31.47	<u>86.37</u>	71.67
$r = 64$	**57.80**	**66.64**	**76.71**	30.97	86.36	**71.88**
State Loss Weight λ_s ($r = 16$)						
$\lambda_s = 0.00$	55.53	63.86	73.48	38.10	85.74	69.65
$\lambda_s = 0.01$	<u>57.63</u>	66.29	76.38	30.60	86.19	71.62
$\lambda_s = 0.05$	57.40	66.13	76.28	30.64	86.11	71.48
$\lambda_s = 0.10$	57.39	66.08	76.20	30.72	86.23	71.48
Radial Loss Weights ($r = 16$)						
$\lambda_P = 0$	56.90	65.74	75.79	31.72	86.03	71.11
$\lambda_P = 0.004$	57.45	66.22	76.29	30.66	86.29	71.56
$\lambda_B = 0$	57.45	66.11	76.11	30.92	86.13	71.45
$\lambda_B = 0.0005$	57.57	66.33	76.51	**30.22**	86.32	<u>71.68</u>
$\lambda_T = 0$	57.47	66.33	<u>76.54</u>	<u>30.24</u>	**86.39**	<u>71.68</u>
$\lambda_T = 0.01$	57.62	66.31	76.37	30.47	86.19	71.62
$\lambda_T = 0.04$	57.21	65.96	76.00	31.12	86.18	71.34

vocabulary. Overall, these results suggest that TheraWorld not only improves downstream prediction performance, but also induces a more clinically meaningful organization of token representations in the final hidden space.

G.3 Hyperparameter Experiment

Table 11 presents the hyperparameter study of LoRA configurations on Qwen3-8B. Overall, all LoRA-based variants substantially outperform the base model without LoRA, demonstrating the effectiveness of parameter-efficient adaptation for lab prediction. For the LoRA rank, performance improves consistently as the rank increases from $r = 8$ to $r = 64$, with $r = 64$ achieving the best average score of 71.88 and the strongest results on S@10, S@15, and S@25. This suggests that a higher-rank adaptation provides greater capacity to capture clinically relevant state-transition patterns. For the state consistency loss, removing the loss term ($\lambda_s = 0.00$) leads to a clear performance drop, especially in sMAPE and threshold-based accuracy metrics, indicating that preserving patient-state structure is important for stable prediction. Among different loss weights, $\lambda_s = 0.01$ achieves the best average performance, while larger weights show slightly lower scores, suggesting that a moderate constraint is beneficial, whereas overly strong regularization may limit task-specific adaptation. Overall, the results show that both sufficient LoRA capacity and an appropriately weighted state consistency loss are crucial for improving lab prediction performance.

The radial loss weights behave differently, with performance remaining within a narrow band as each weight is varied in isolation from its default. Setting any single radial weight to zero changes sMAPE by at most 1.08 points, and the full grid spans only 1.50 points, in contrast to the 7.46-point degradation incurred by removing the state consistency term. This suggests that the radial objectives primarily shape the geometry of the representation space rather than carry predictive accuracy, and that the method is not sensitive to their precise calibration.



	Ground Truth	TheraWorld	LoRA	DeepSeek-V4
Creatinine	4.5	**4.7 (+4.4%)**	5.1 (+13.3%)	5.1 (+13.3%)
Potassium	4.3	**4.3 (+0.0%)**	4.4 (+2.3%)	4.6 (+7.0%)
Bilirubin	0.6	**0.6 (+0.0%)**	0.7 (+16.7%)	0.7 (+16.7%)
RDW-SD	46.20	**45.30 (+1.9%)**	45.10 (-2.4%)	13.20 (-71.4%)

Figure 7: Case study of TheraWorld for treatment-driven patient-state prediction. Given patient information and multi-source clinical history, TheraWorld predicts laboratory values for Round 11 and is compared with LoRA and DeepSeek-V4-Flash. The lower panel visualizes the corresponding hyperbolic trajectory, where the outward movement and increasing radius indicate worsening clinical severity.

G.4 Case Study

Figure 7 presents a representative case illustrating TheraWorld's prediction and geometric interpretation of patient-state evolution. The input consists of multi-source clinical context, including patient demographics, microbiology findings, key laboratory observations, and medication history. Given this observed history, the model predicts laboratory values for Round 11 and compares them with LoRA and DeepSeek-V4-Flash. TheraWorld produces predictions closest to the ground truth for most indicators, accurately matching potassium and bilirubin, while also reducing the creatinine error compared with both baselines. These results suggest that TheraWorld better captures short-term physiological changes from heterogeneous clinical evidence.

Beyond numerical prediction, the lower-right panel visualizes the learned hyperbolic patient trajectory. The patient's representation moves outward along the hyperbolic manifold from earlier rounds to the predicted state at $t = 11$, with the radius increasing from 1.0010 to 2.7826 and the corresponding SOFA score rising to 5. This outward movement indicates worsening clinical severity, consistent with our design that hyperbolic radius encodes patient acuity. Overall, this case shows that TheraWorld not only improves laboratory prediction accuracy, but also provides an interpretable geometric view of disease progression.

G.5 Counterfactual Simulations

To examine whether TheraWorld's predictions depend on the administered treatment rather than patient history alone, we simulate the same patient under two treatment conditions: the original observed treatment sequence and a withheld-treatment scenario in which medications and procedures are removed from the input. Table 12 reports predicted lab values for both conditions across eight indicators spanning liver function and hematology. All eight indicators shift in the clinically expected direction under withheld treatment, with liver injury markers (ALT, AST) and albumin showing the

largest changes, suggesting that the model has internalized treatment-physiology dependencies from training data.

Table 12: Predicted lab values for the same patient under the original treatment sequence and a withheld-treatment condition.

Lab	Original	No Treatment
Hematology: White Blood Cells	3.6	3.7
Hematology: Hematocrit	34.5	34.7
Hematology: Hemoglobin	11.3	11.4
Hematology: Red Blood Cells	3.98	4.04
Chemistry: Alkaline Phosphatase	75	77
Chemistry: Alanine Aminotransferase	185	193
Chemistry: Aspartate Aminotransferase	77	80
Chemistry: Albumin	3.8	3.9

We further evaluate time-conditioned prediction by querying the model at three future time horizons beyond the current round. Table 13 reports predictions at +6h, +24h, and +48h for seven indicators. Creatine Kinase MB decreases markedly over 48 hours, glucose and potassium normalize progressively, and urea nitrogen declines steadily, producing trajectories that are clinically plausible given the patient's condition.

Table 13: Time-conditioned predictions at +6h, +24h, and +48h beyond the current treatment round.

Lab	Original	+6h	+24h	+48h
Chemistry: Creatine Kinase MB	128	108	102	10
Chemistry: Creatinine	1.6	1.6	1.6	1.4
Chemistry: Glucose	145	127	127	105
Chemistry: Magnesium	2.4	2.7	2.6	2.1
Chemistry: Potassium	5.3	4.6	4.5	4.3
Chemistry: Sodium	136	137	137	137
Chemistry: Urea Nitrogen	70	68	68	61

G.6 Extended Main Results

Table 14 extends the main results table with two additions: statistical significance annotations for API-based models, and non-neural baselines including gradient boosted regression trees (GBRT) and random forests (RF). Locally trained models use greedy decoding and produce deterministic outputs, so no confidence interval is reported for these. API-based models use stochastic generation; we report mean $\pm$ std over three runs and annotate significance relative to TheraWorld-14B using a z-test. TheraWorld-14B achieves statistically significant improvements over all API models on S@10, S@15, and S@25 ($z > 3$). TheraWorld-8B substantially outperforms both tree-based baselines, with a sMAPE reduction exceeding 0.10 (25% relative improvement), showing that the natural-language generation setting is not a barrier to accurate numerical prediction.

H Prompt

H.1 Training Prompt

We present the prompt template used during supervised fine-tuning in Table 15. The training prompt consists of a system message and a user message; the assistant answer (a structured JSON prediction) serves as the training target and is masked from the loss computation over prompt tokens.

The system message describes the task and patient demographics, while the user message provides structured clinical history (medications, procedures, prior lab results), target lab items to predict, and optional prior-round SOFA severity scores.

Table 14: Extended main results with non-neural baselines and statistical significance. API-based models report mean$_{\pm\text{std}}$ over three runs; significance is against TheraWorld-14B via z-test (*** $p < 0.001$, * $p < 0.05$). Locally trained models use greedy decoding and non-neural baselines use deterministic inference produce, so no variance is reported.

Model	S@10↑	S@25↑	sMAPE↓	Avg↑
Non-Neural Baselines				
GBRT	47.53	63.85	41.97	62.31
RF	47.63	64.05	41.10	62.44
Closed-Source Models				
GPT-5-Mini	$55.28_{\pm 0.61}$ ***	$73.62_{\pm 0.48}$ ***	$32.89_{\pm 0.89}$ ***	–
Claude-Haiku-4.5	$56.00_{\pm 0.08}$ ***	$75.47_{\pm 0.08}$ ***	$31.17_{\pm 0.26}$ ***	–
Open-Source General-Purpose Models				
Qwen3.5-122B	$56.54_{\pm 0.61}$ *	$75.36_{\pm 0.13}$ ***	$30.70_{\pm 0.12}$ ***	–
DeepSeek-V4-Flash	$54.32_{\pm 0.32}$ ***	$72.54_{\pm 1.14}$ ***	$35.75_{\pm 3.75}$	–
GLM-5.1	$55.08_{\pm 0.51}$ ***	$72.45_{\pm 0.34}$ ***	$38.25_{\pm 0.68}$ ***	–
Ours TheraWorld				
TheraWorld-8B	<u>57.40</u>	<u>76.28</u>	<u>30.64</u>	<u>71.48</u>
TheraWorld-14B	**58.01**	**76.81**	**30.14**	**72.10**

H.2 Inference Prompt

We present the prompt template used during inference in Table 16. The system message and user message follow the same structure as training.

Table 15: The prompt template used for supervised fine-tuning.

Prompt Template for SFT Training

System Message:

```
You are given patient information available before the target
round, chronological prior clinical events, and the next round
target metadata. Predict the requested lab numeric values. Rows
follow the schemas in the user message; trailing missing fields are
omitted. Return only valid JSON matching the assistant schema.

Patient Information
---
Subject ID: {subject_id}
Admission ID: {hadm_id}
Age: {age}
Sex: {sex}
```

User Message:

```
{
  "task": "predict_next_round_lab_values",
  "schemas": {
    "history_lab": ["category", "fluid", "label", "value", "unit"],
    "history_medication": ["medication", "event"],
    "history_procedure": ["code", "title"],
    "target_lab": ["category", "fluid", "label", "unit"],
    "assistant_lab": ["category", "fluid", "label", "value", "unit"]
  },
  "patient": {
    "subject_id": {subject_id},
    "hadm_id": "{hadm_id}",
    "age": {age},
    "sex": "{sex}"
  },
  "history": [
    {
      "round_index": 0,
      "groups": [
        {
          "time": "{time}",
          "medications": [["{medication}", "{event}"], ...]
        },
        {
          "time": "{time}",
          "lab": [["{category}", "{fluid}", "{label}", {value},
"{unit}"], ...]
        }
      ]
    }, ...
  ],
  "target_round_index": {target_round_index},
  "targets": [
    {
      "time": "{time}",
      "lab": [["{category}", "{fluid}", "{label}", "{unit}"], ...]
    }
  ],
  "previous_round_sofa_lab3": {
    "total": {score}, ...
  }
}
```

Table 16: The prompt template used during inference.

Prompt Template for Inference

System Message:

```
You are given patient information available before the target
round, chronological prior clinical events, and the next round
target metadata. Predict the requested lab numeric values. Rows
follow the schemas in the user message; trailing missing fields are
omitted. Return only valid JSON matching the assistant schema.

Patient Information
---
Subject ID: {subject_id}
Admission ID: {hadm_id}
Age: {age}
Sex: {sex}
```

User Message:

```
{
  "task": "predict_next_round_lab_values",
  "schemas": { ... },
  "patient": { ... },
  "history": [ ... ],
  "target_round_index": {target_round_index},
  "targets": [
    {
      "time": "{time}",
      "lab": [
        ["{category}", "{fluid}", "{label}", "{unit}"]
      ]
    }
  ],
  "previous_round_sofa_lab3": {
    "total": {score},
    ...
  }
}
```