SEMINAR

TIMER: Temporal Instruction Modeling and Evaluation

For Longitudinal Clinical Records

Presenter: Yuzhen He

02.12.2025

Agenda

The Challenge

- Multiple-Visit temporal reasoning
- The Bias Problem
- Existing Benchmark Gaps

The TIMER Solution

- TIMER-Bench Framework
- TIMER-Instruct Methodology
- Experimental Results & Impact

|— 🏥Visit 1 —|——— 🏥Visit 2 ———|——— 🏥Visit 3 ————|——— 🏥Visit 4 ——|

2.2018 … 7.2019 … 1.2021 … 10.2023

Context: The Longitudinal Challenge

- **Promise of LLMs:** Growing utility in medical tasks (USMLE, knowledge retrieval).
- **The Reality:** Clinical records span years, not just single visits.
- **Cognitive Load:** Physicians analyze thousands of entries across time.
- **Critical Gap:** Current models struggle to reason over temporal dependencies across multiple visits.

|── 🏥Visit 1 ──|──── 🏥Visit 2 ─────|──── 🏥Visit 3 ──────|──── 🏥Visit 4 ───|

2.2018 … 7.2019 … 1.2021 … 10.2023

Problem: Temporal Evaluation Gaps

Recency Bias

Existing benchmarks focus on recent notes or events.

Limits understanding of model performance across full patient timelines.

Single-Visit Focus

Datasets like MIMIC-Instr rely on ICU stays (avg 7.2 days).

Fails to capture chronic disease management or long-term care planning.

Benchmark Comparison

Benchmark	Avg Time Span	Multi-Visit?	Limitations
MIMIC-Instr	7.2 days	No	Restricted to short ICU episodes; Notes only
MedAlign	3,895 days	Yes	**Recency Bias:** Instructions focus heavily towards end of timeline
TIMER-Bench	1,295 days	Yes	**Balanced:** Explicit temporal evidence; Structured + Unstructured data

The TIMER Framework

Temporal **I**nstruction **M**odeling and **E**valuation for **R**ecords

Component 1: TIMER-Bench

(Q, A, T)

Explicit Evidence

Includes date-time evidence tuples (Q, A, T) to ground responses.

Multi-Timepoint

Evaluates reasoning across non-contiguous visits.

Benchmark Generation Pipeline

Input

Aggregated patient visits (XML format).

LLM Generation

Prompt model for Q&A with Time Evidence.

Grounding

Extract Evidence Tuples {T1...Tn}.

Filtering

Select questions requiring multi-visit synthesis.

Clinical Validation

Validated by 3 clinicians on relevance, accuracy and complexity.

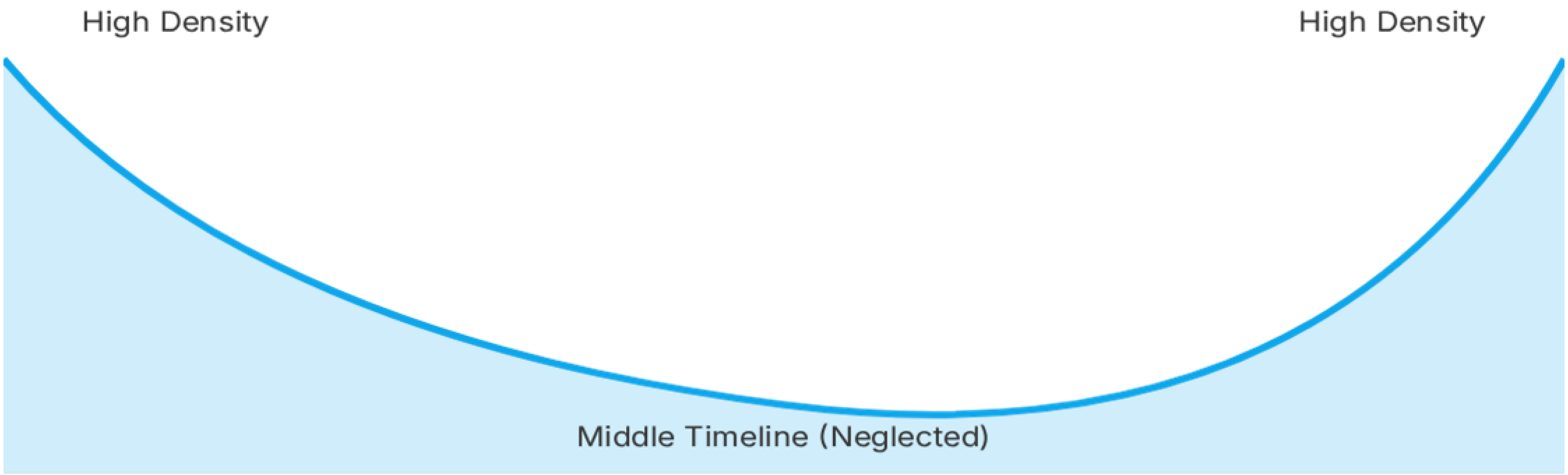
95/100	98/100	80/100
Clinical Relevance	Factual Accuracy	Reasoning Complexity

Component 2: TIMER-Instruct

Methodology for Temporal Instruction Tuning

The "Lost-in-the-Middle" Phenomenon

Analysis of model-generated instructions revealed a default bias:



Models focus on edges (Start/End) and overlook the middle period.

Tuning Strategies: Temporal Distribution

Recency-Focused

Concentrates instructions in the last quartile.

(Mimics human data)

Edge-Focused

Higher density at start and end of timeline.

(Natural LLM bias)

Uniform

Balanced coverage across all relative positions.

(The TIMER Approach)

Experimental Setup

Base Model

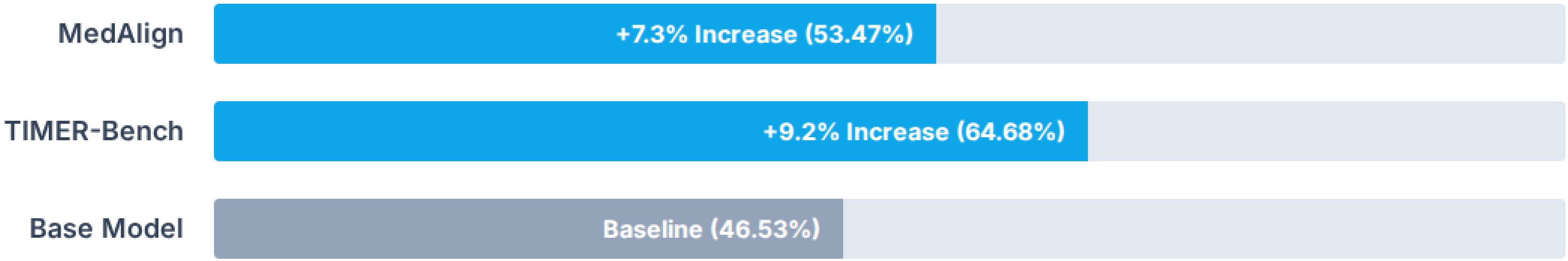
- Llama-3.1-8B-Instruct
- Context Window: 16K tokens
- Training: LoRA (6 epochs)

Datasets & Metrics

- **Training:** 5,000 synthetic pairs.
- **Evaluation:** MedAlign (Human) & TIMER-Bench.
- **Metrics:** LLM-Judge (Correctness/Completeness), ROUGE-L.

Results: Performance Improvement

Comparison of Llama-3.1-8B Base vs. TIMER-Instruct Tuned



Head-to-Head: TIMER vs. Baselines

Baseline Model	MedAlign Win %	TIMER-Bench Win %
Meditron-7B	**+83.10%**	**+95.02%**
MedAlpaca	**+72.80%**	**+86.41%**
MedLM-Large	**+27.80%**	**+52.49%**
Llama-3.1-8B Base	**+23.80%**	**+17.67%**

*Values indicate additional win margin by TIMER-Instruct.

Make time visible.

Use the whole timeline.

Q&A

Thank you for your attention