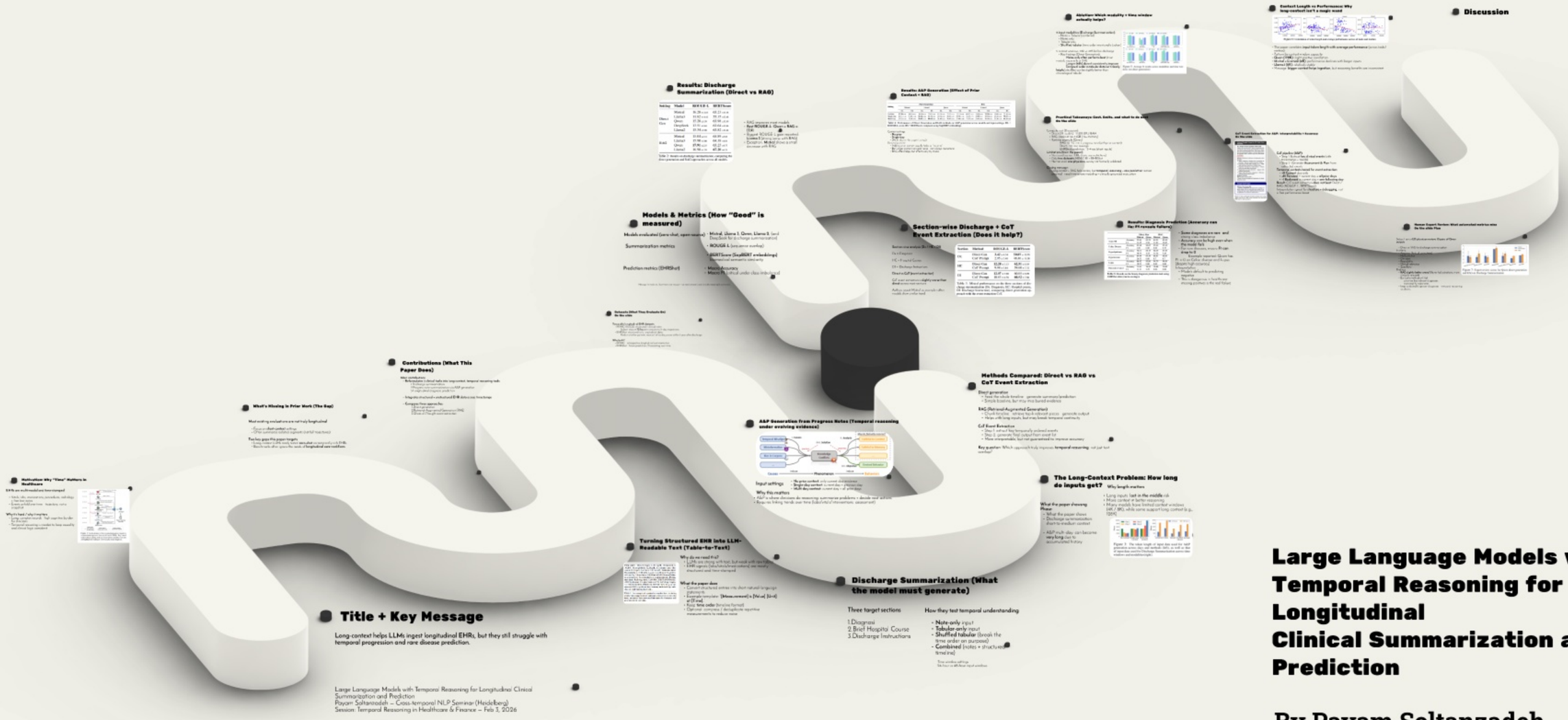


Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction

By Payam Soltanzadeh

Title + Key Message

Long-context helps LLMs ingest longitudinal EHRs, but they still struggle with temporal progression and rare disease prediction.

Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction
Payam Soltanzadeh — Cross-temporal NLP Seminar (Heidelberg)
Session: Temporal Reasoning in Healthcare & Finance — Feb 3, 2026

Motivation: Why "Time" Matters in Healthcare

EHRs are multi-modal and time-stamped

- Vitals, labs, medications, procedures, radiology + free-text notes
- Events unfold over time trajectory, not a snapshot

Why it's hard / why it matters

- Long, complex records high cognitive burden for clinicians
- Temporal reasoning is needed to keep causality and clinical logic consistent

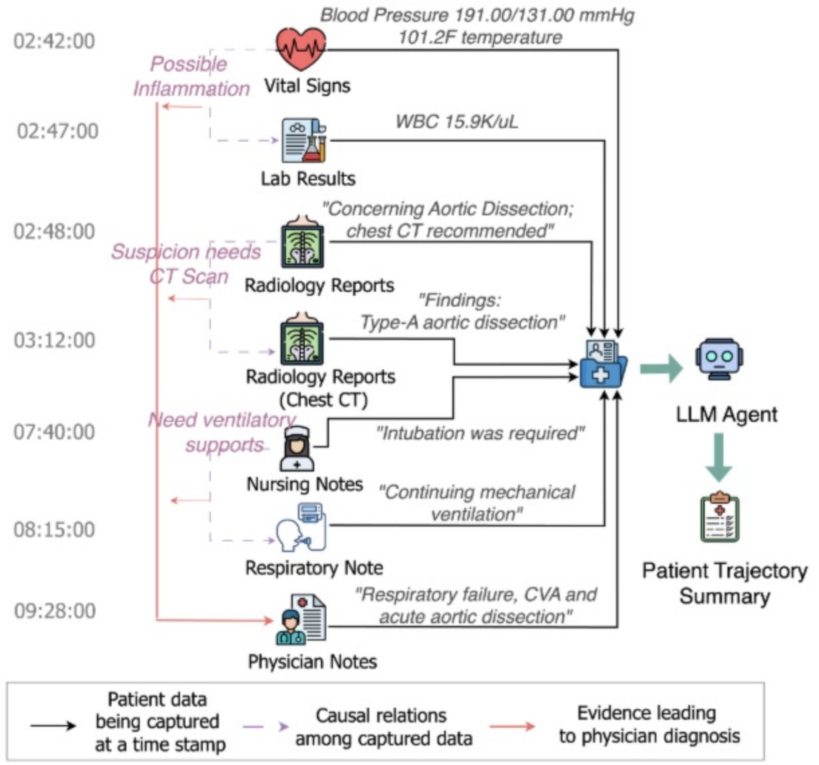


Figure 1: An illustration of the longitudinal patient trajectory summarization process from multi-modal EHRs. Key causal relationships among medical observations and interventions are highlighted, leading to the final physician diagnosis.

What's Missing in Prior Work (The Gap)

Most existing evaluations are not truly longitudinal

- Focus on **short-context** settings
- Often summarize isolated segments (not full trajectories)

Two key gaps this paper targets

- Long-context LLMs rarely tested **zero-shot** on temporally rich EHRs
- Benchmarks often ignore the needs of **longitudinal care workflows**

Contributions (What This Paper Does)

Main contributions

- **Reformulates 3 clinical tasks into long-context, temporal reasoning tasks**
  1. Discharge summarization
  2. Progress note summarization via A&P generation
  3. Longitudinal diagnosis prediction

- **Integrates structured + unstructured EHR data across timestamps**

- **Compares three approaches:**
  1. Direct generation
  2. Retrieval-Augmented Generation (RAG)
  3. Chain-of-Thought event extraction

Datasets (What They Evaluate On)
On the slide

Two public longitudinal EHR datasets

- MIMIC-III (ICU): structured + clinical notes
  Subset: stays **> 72 hours** to ensure multi-day trajectories
- EHRShot: structured-only longitudinal data
  Predict whether patients receive 1 of six diagnoses within 1 year after discharge

Why both?

- MIMIC retrospective longitudinal summarization
- EHRShot future prediction / forecasting over time

Turning Structured EHR into LLM-Readable Text (Table-to-Text)

First entry: Urea Nitrogen is 26 mg/dL. Hematocrit is 34.20%. Hemoglobin is 12.20 g/dL. 12 minutes later: Glucose is 214 mg/dL. Lactate is 2.40 mmol/L. **3 minutes later:** Hemoglobin is 13.00 g/dL. Lactate is 2.40 mmol/L. pO2 is 441 mm Hg. **1 hour later**: 100.00 ml of 0.9% Normal Saline is administered. Neosynephrine-k is administered. **49 minutes later:** Radiology note: 12:48 PM CHEST (PORTABLE AP) Clip Reason: line placement, r/o PTx Admitting Diagnosis: HEAD BLEED; MEDICAL CONDITION: 68-year-old man s/p MVA significant head trauma, intubated s/p r subclavian triple lumen placement...

Table 1: An example of a patient's compiled data, including relative timestamp (minutes and hours since previous recorded data), structured data converted into narrative format as well as unstructured note data.

Why do we need this?

- LLMs are strong with text, but weak with raw tables
- EHR signals (labs/vitals/medications) are mostly structured and time-stamped

What the paper does

- Convert structured entries into short natural-language statements
- Example template: "**[Measurement] is [Value] [Unit] at [Time].**"
- Keep **time order** (timeline format)
- Optional: compress / deduplicate repetitive measurements to reduce noise

Discharge Summarization (What the model must generate)

Three target sections

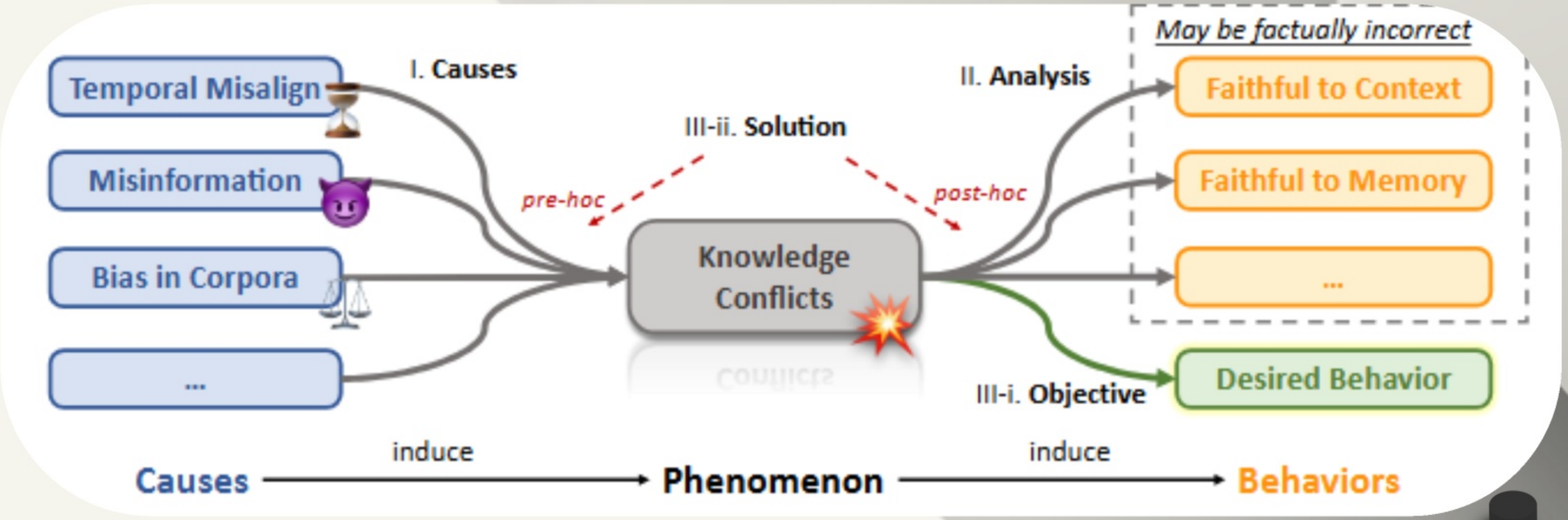
1. Diagnosi
2. Brief Hospital Course
3. Discharge Instructions

How they test temporal understanding

- **Note-only** input
- **Tabular-only** input
- **Shuffled tabular** (break the time order on purpose)
- **Combined** (notes + structured timeline)

Time window settings
24-hour vs 48-hour input windows

A&P Generation from Progress Notes (Temporal reasoning under evolving evidence)



Input settings

- **No prior context:** only current day evidence
- **Single-day context:** current day + previous day
- **Multi-day context:** current day + all prior days

Why this matters

- A&P is where clinicians do reasoning: summarize problems + decide next actions
- Requires linking trends over time (labs/vitals/interventions assessment)

The Long-Context Problem: How long do inputs get?

Why length matters

- Long inputs **lost-in-the-middle** risk
- More context ≠ better reasoning
- Many models have limited context windows (4K / 8K), while some support long-context (e.g., 128K)

What the paper showsng Phase:

- What the paper shows
- Discharge summarization: short-to-medium context

- A&P multi-day: can become **very long** due to accumulated history

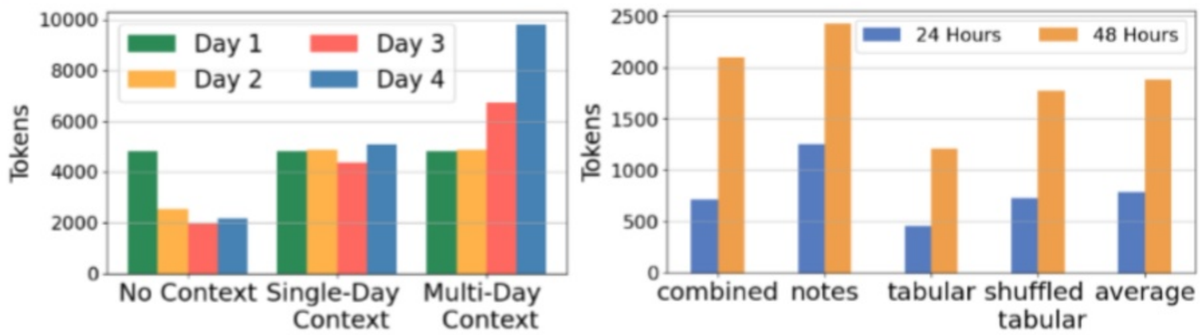


Figure 3: The token length of input data used for A&P generation across days and methods (left), as well as that of input data used for Discharge Summarization across time windows and modalities(right).

Methods Compared: Direct vs RAG vs CoT Event Extraction

Direct generation

- Feed the whole timeline generate summary/prediction
- Simple baseline, but may miss buried evidence

RAG (Retrieval-Augmented Generation)

- Chunk timeline retrieve top-k relevant pieces generate output
- Helps with long inputs, but may break temporal continuity

CoT Event Extraction

- Step 1: extract key temporally ordered events
- Step 2: generate final output from event list
- More interpretable, but not guaranteed to improve accuracy

Key question: Which approach truly improves **temporal reasoning** not just text overlap?

Models & Metrics (How "Good" is measured)

Models evaluated (zero-shot, open-source)

- **Mistral, Llama 3, Qwen, Llama 2**, (and DeepSeek for discharge summarization)

Summarization metrics

- **ROUGE-L** (sequence overlap)
- **BERTScore (SapBERT embeddings)** biomedical semantic similarity

Prediction metrics (EHRShot)

- **Macro Accuracy**
- **Macro F1** (critical under class imbalance)

Message: In medicine, fluent text is not enough — we need semantic and clinically meaningful correctness.

Results: Discharge Summarization (Direct vs RAG)

Setting	Model	ROUGE-L	BERTScore
Direct Gen	Mistral	16.28 ±12.83	65.23 ±13.38
	Llama3	11.82 ±14.42	55.35 ±22.48
	Qwen	15.28 ±5.38	63.98 ±13.50
	DeepSeek	13.51 ±5.04	63.64 ±13.86
	Llama2	15.34 ±5.90	63.82 ±11.44
RAG	Mistral	15.04 ±2.34	65.89 ±5.69
	Llama3	15.90 ±2.00	64.38 ±4.81
	Qwen	**17.91** ±2.35	65.25 ±4.77
	Llama2	16.98 ±1.59	**67.20** ±6.21

Table 2: Results on discharge summarization, comparing the direct generation and RAG approaches across all models.

- RAG improves most models
- **Best ROUGE-L: Qwen + RAG = 17.91**
- Biggest ROUGE-L gain reported: **Llama 3** (strong jump with RAG)
- Exception: **Mistral** shows a small decrease with RAG

Section-wise Discharge + CoT Event Extraction (Does it help?)

Section-wise analysis (Dx / HC / DI)

Dx = Diagnosis

HC = Hospital Course

DI = Discharge Instructions

Direct vs CoT (event extraction)

CoT event extraction is **slightly worse than direct** across most sections

Authors report Mistral as example; other models show similar trend

Section	Method	ROUGE-L	BERTScore
Dx	Direct Gen	**3.42** ± 4.34	**50.07** ± 10.78
	CoT Prompt	2.95 ± 3.48	48.46 ± 11.00
HC	Direct Gen	**12.28** ± 3.13	**62.51** ± 5.95
	CoT Prompt	9.98 ± 3.84	59.60 ± 7.32
DI	Direct Gen	**12.07** ± 3.90	60.05 ± 9.09
	CoT Prompt	10.83 ± 4.58	**60.52** ± 7.86

Table 3: Mistral performance on the three sections of discharge summarization (Dx: Diagnosis, HC: Hospital course, DI: Discharge Instruction), comparing direct generation approach with the event extraction CoT.

Results: A&P Generation (Effect of Prior Context + RAG)

Setting	*Direct Generation*						*RAG*					
	Mistral		Llama3		Qwen		Mistral		Llama3		Qwen	
	RL	BS	RL	BS	RL	BS	RL	BS	RL	BS	RL	BS
No Prior	**17.70**$_{\pm 3.79}$	68.16$_{\pm 6.91}$	16.60$_{\pm 3.39}$	72.01$_{\pm 6.04}$	15.15$_{\pm 3.53}$	72.17$_{\pm 4.51}$	17.15$_{\pm 5.60}$	68.57$_{\pm 6.16}$	15.48$_{\pm 3.60}$	**73.78**$_{\pm 4.50}$	15.60$_{\pm 3.60}$	72.16$_{\pm 4.63}$
Single-Day	26.11$_{\pm 11.68}$	71.63$_{\pm 10.86}$	**33.16**$_{\pm 10.68}$	80.10$_{\pm 8.93}$	20.74$_{\pm 4.19}$	78.01$_{\pm 4.95}$	29.50$_{\pm 12.46}$	74.23$_{\pm 7.79}$	22.09$_{\pm 5.51}$	78.80$_{\pm 4.96}$	25.42$_{\pm 8.08}$	**79.27**$_{\pm 5.37}$
Multi-Day	23.32$_{\pm 13.05}$	72.41$_{\pm 9.47}$	**32.42**$_{\pm 12.14}$	**80.42**$_{\pm 8.81}$	21.60$_{\pm 5.23}$	78.41$_{\pm 4.81}$	27.40$_{\pm 10.08}$	73.55$_{\pm 8.46}$	21.42$_{\pm 6.81}$	78.96$_{\pm 5.42}$	25.36$_{\pm 7.50}$	80.28$_{\pm 4.68}$

Table 4: Performance of Direct Generation and RAG methods on A&P generation across models and input settings. RL = ROUGE-L score; BS = BERTScore computed using SapBERT embeddings.

Context settings:

- **No prior**
- **Single-day**
- (Multi-day in the paper's setup)

General pattern:

- Adding prior context usually helps vs "no prior"
- But longer context can add noise not always monotonic
- RAG often helps, but effects vary by model

Results: Diagnosis Prediction (Accuracy can lie; F1 reveals failure)

		Direct Gen		RAG	
		Mistral	Qwen	Mistral	Qwen
Acute MI	Accuracy	52.84	62.72	63.03	65.50
	F1	33.45	2.58	11.83	16.87
Celiac Disease	Accuracy	62.96	96.05	95.04	95.30
	F1	2.60	0.00	0.00	9.52
Hyperlipidemia	Accuracy	58.02	68.15	66.09	63.03
	F1	30.33	1.53	12.74	7.45
Hypertension	Accuracy	64.44	69.38	68.81	68.81
	F1	26.53	6.06	8.7	8.7
Lupus	Accuracy	66.42	95.06	94.79	94.81
	F1	10.53	0.00	0.00	0.00
Pancreatic Cancer	Accuracy	73.83	78.52	78.86	79.05
	F1	13.11	2.25	0.00	0.00

Table 6: Results on the binary diagnosis prediction task using EHRShot data (macro-average).

- **Some diagnoses are rare and** strong class imbalance
- **Accuracy can be high even when the model fails**
- For rare diseases, macro **F1 can drop to 0**

Example reported: Qwen has F1 = 0 on Celiac disease and Lupus (despite high accuracy)

Interpretation

- Models default to predicting negative
- This is dangerous in healthcare: missing positives is the real failure

Ablation: Which modality + time window actually helps?

4 input modalities (Discharge Summarization):

- Notes + Tabular (combined)
- Notes only
- Tabular only
- **Shuffled tabular** (time order intentionally broken)

2 context windows: 24h vs 48h before discharge

- Key findings (Direct Generation):

Notes-only often performs best (most models, especially at 24h)

Longer (48h) doesn't consistently improve

Temporal order in tabular data isn't clearly helpful: shuffled can be slightly better than chronological tabular

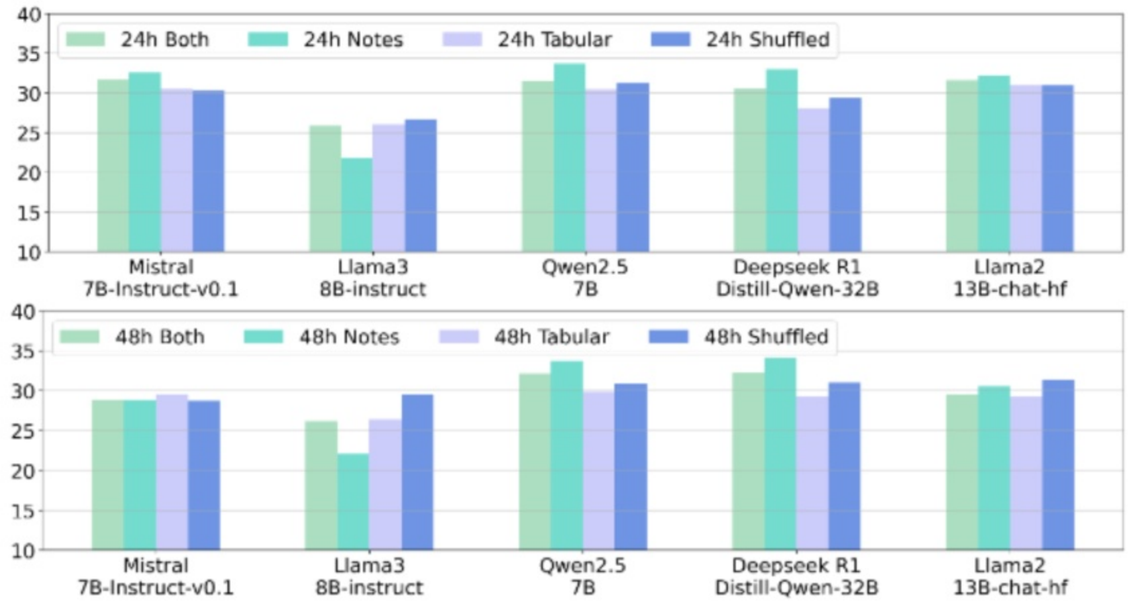


Figure 5: Average f1 results across modalities and time windows (on direct generation)

CoT Event Extraction for A&P: Interpretability ≠ Accuracy

On the slide

Chain-of-Thought Prompt for ICU Daily Event Extraction

ICU DAILY EVENT EXTRACTION TASK
Analyze this structured ICU data by identifying critical clinical events, paying special attention to numerical values and their progression. Only report values showing meaningful change or clinical significance. For repeated values, mention only those demonstrating changes.
{Input Text}
Identify (with direct references to data points when possible):

1. Major symptoms or changes (new, worsening, improving) – Specify relevant numerical changes.
2. Critical test results (labs, imaging, etc.) – Highlight significant abnormal or normal values.
3. Important treatments or interventions – Clearly link to the preceding clinical data.
4. Significant care team decisions – Support with relevant clinical data.
5. Major medical decisions or diagnoses – Reference pertinent clinical observations.

Response Format:
Day X Key Events
- [Time] | [Event Description] (Explanation for identifying this event)

Example Model Output

Day 3 Key Events
- **2178-02-11 09:50:00** | Blood pressure readings of 88/46 mmHg (This low blood pressure reading is concerning and may require additional fluid resuscitation or adjustment of vasoactive medications.)

Figure 4: Chain-of-Thought prompting template and example output used for ICU daily event extraction from structured EHR data for A&P generation.

CoT pipeline (A&P):

- Step 1: Extract **key clinical events** (with timestamps + trends)
- Step 2: Generate **Assessment & Plan** from extracted events

Temporal contexts tested for event extraction:

- **+0 Current** day only
- **+N Forward** = current day + **all prior days**
- **-1 Backward** = current day + **one following day**

Result: CoT event extraction **does not beat** Direct / RAG (ROUGE-L, BERTScore)

Interpretation: great for **structure + debugging,** not a free performance boost

Context Length vs Performance: Why long-context isn't a magic wand

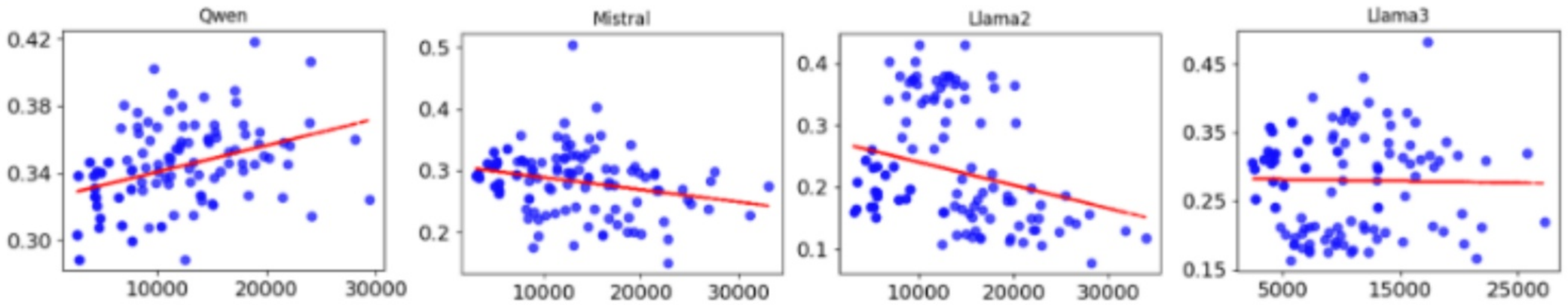


Figure 6: Correlation of token length and average performance across all tasks and metrics

- The paper correlates **input token length with average performance** (across tasks/metrics)
- Pattern by context window capacity:
- **Qwen (128K):** slight positive correlation
- **Mistral + Llama2 (4K):** performance declines with longer inputs
- **Llama3 (8K):** relatively stable
- Message: **bigger context helps ingestion**, but reasoning benefits are inconsistent

Human Expert Review: What automated metrics miss
On the slide Plan

Setup: 1 senior ED **physician reviews 10 pairs of Qwen outputs**

- Direct vs RAG for discharge summarization

6 criteria:

- Accuracy (factual correctness)
- Hallucination
- Omission
- Readability
- Clinical relevance
- Specificity

Findings:

- **RAG slightly better overall** (fewer hallucinations; more clinically relevant)
- But both methods still fail:
  prioritize less relevant diagnoses
  oversimplify trajectories
- keep outdated/disproven diagnoses temporal reasoning weakness

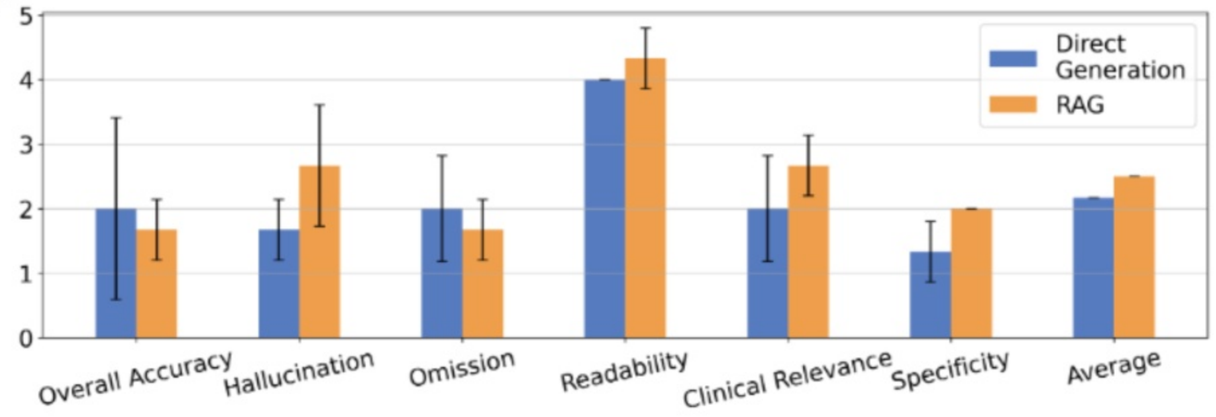


Figure 7: Expert review scores for Qwen direct generation and RAG on Discharge Summarization

Practical Takeaways: Cost, limits, and what to do next
On the slide

Compute cost (Discussion):

- Direct (7B models): ~15GB GPU RAM
- RAG: peaks at 32–33GB (≈2× memory)
- Runtime example (Qwen):

  RAG: 18–20 min / progress note (with prior context)
  Direct: ~10 min average
  EHRShot prediction: ~3–4 sec (short inputs)

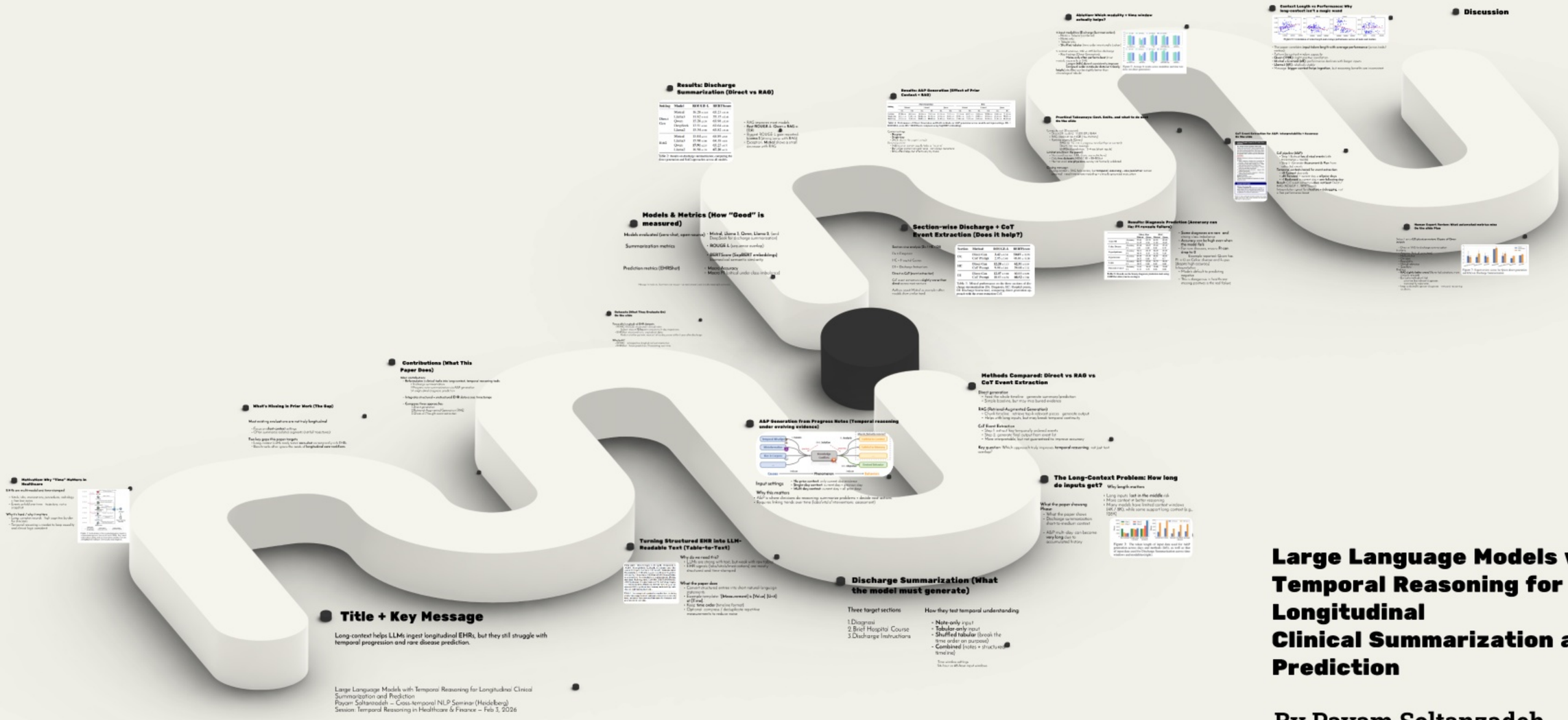
Limitations (from the paper):

- No closed-source LLMs (data use restrictions)
- Only **two datasets:** MIMIC-III + EHRShot
- Human eval: **one physician,** survey not formally validated

Closing message:

- Long-context + RAG help access, but **temporal reasoning** + **rare prediction** remain unsolved need time-aware modeling + clinically grounded evaluation

Discussion



Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction

By Payam Soltanzadeh

Take this with you. Revisit anytime.

Missed something? Want to explore further?
Scan or click below to open this presentation.
Anytime, anywhere.

View presentation