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

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker
Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, Yanjun Gao

LARK — Language, Reasoning, Knowledge

Anschutz

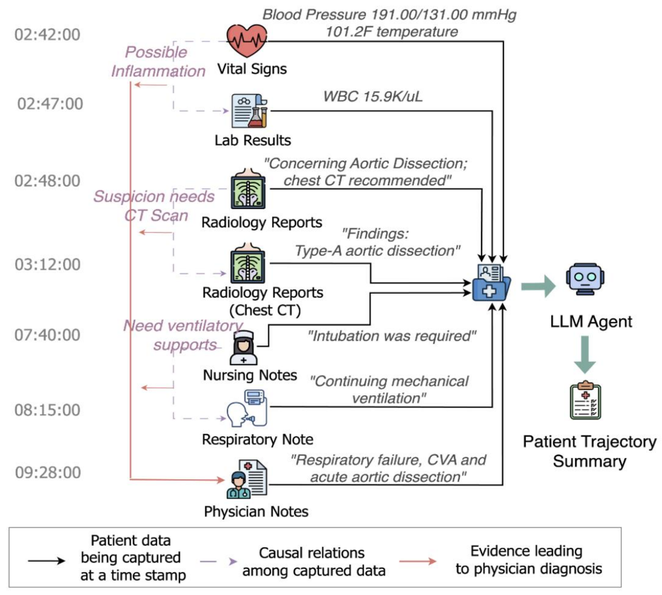
University of Colorado Boulder

Northeastern University

AMIA — INFORMATICS PROFESSIONALS. LEADING THE WAY.

Introduction

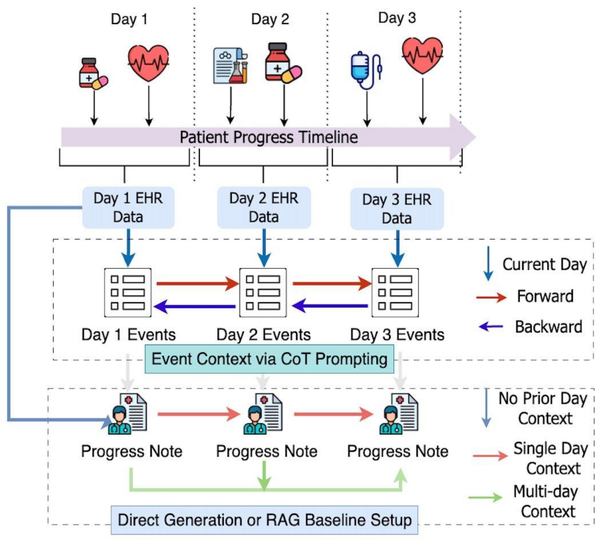
- **Electronic Health Records (EHRs)** are *longitudinal, multi-modal, and time-dependent*,
- containing structured data and unstructured notes.
- Clinicians face **information overload** when reviewing multi-day patient histories.
- **LLMs** show promise for summarization and clinical reasoning but struggle with
- **temporal consistency** and **causal understanding**.
- Existing evaluations focus on **static, single-visit summaries**, not longitudinal reasoning.
- Goal: assess **open-source LLMs' ability** to synthesize, summarize, and predict across
- **time-evolving patient records**.



Contributions

- First systematic evaluation of **temporal reasoning in LLM-based EHR summarization**.
- Introduce **longitudinal task reformulations** for Discharge Summarization, A&P Generation, and Diagnosis Prediction.
- Compare **Direct Generation**, **RAG**, and **Chain-of-Thought (CoT)** reasoning paradigms

Methods



- **Datasets:**
  - *MIMIC-III*: ICU patient notes + structured data for summarization tasks.
  - *EHRShot*: structured-only longitudinal dataset for diagnosis prediction.
- **Models:** Mistral-7B, Llama3-8B, Qwen2.5-7B, DeepSeek-R1-32B, Llama2-13B.
- **Approaches:**
  - *Direct Generation*: standard summarization.
  - *RAG*: retrieve relevant segments to reduce "lost-in-the-middle."
  - *CoT*: extract temporally ordered daily events before generation.
- **Evaluation:**
- Automated metrics: ROUGE-L, BERTScore, F1.
- LLM-as-Judge (PDSQI-9 dimensions: accuracy, completeness, usefulness, clarity, synthesis).
- Expert clinical review for factuality and relevance.

Chain-of-Thought Prompt for 48h Discharge Summary Event Extraction

DISCHARGE EVENT EXTRACTION TASK Analyze the following data from the final 48 hours of the hospital stay and identify key clinical events that are most relevant for summarizing the course of treatment and informing discharge planning. {chronology_text} Only include events that reflect:

1. Significant changes in symptoms or status (e.g., improvements, worsening, new findings).
2. Clinically important test results (especially abnormal values that lead to certain treatments).
3. Major treatments or interventions (e.g., medication changes, procedures, escalation/de-escalation).
4. Care team decisions that indicate readiness for discharge or change in care goals.
5. Events linked to the final diagnosis or that inform follow-up care.

Response Format: ### Day X Key Events ### - [Time]: [Description] (Reasoning)

Example Model Output

Day 1 Key Events ### - **2159-03-12 08:00**: Stable Vital Signs (BP 120/70 mmHg, HR 88 bpm, RR 18 breaths/min, Temp 98.6°F). This indicates overall stability and readiness for discharge.

Results

Approach	Task	CIT	ACC	THR	USE	ORG	CMP	SUC	SYN
Direct Gen	DS	1.00±0	1.82±1.21	1.42±0.64	1.70±0.93	2.48±0.99	4.08±0.75	2.96±0.92	3.5±0.71
	A&P	2.13±0.89	3.19±1.15	2.66±0.87	3.51±0.88	3.46±0.86	3.84±0.8	3.03±0.92	3.76±0.66
RAG	DS	1.06±0.24	1.68±0.88	1.52±0.79	1.66±0.94	2.4±1.01	3.92±0.78	2.78±0.76	3.5±0.71
	A&P	2.35±1.02	3.96±0.98	3.46±0.76	4.07±0.81	3.71±0.9	4.11±0.74	3.00±0.91	3.61±0.7

Table 7: Results of the PDSQI-9 evaluation on discharge summarization (DS) and Assessment and Plan generation (A&P). Abbreviations: CIT = citations, ACC = accuracy_extractive, THR = thoroughness, USE = usefulness, ORG = organization, CMP = comprehensibility, SUC = succinctness, SYN = synthesis_abstraction. Definitions of scoring criteria could be found in (Croxford et al., 2025).

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**±3.79	68.16±6.91	16.60±3.39	72.01±6.04	15.15±3.53	72.17±4.51	17.15±5.60	68.57±6.16	15.48±3.60	**73.78**±4.50	15.60±3.60	72.16±4.63
Single-Day	26.11±11.68	71.63±10.86	**33.16**±10.68	80.10±8.93	20.74±4.19	78.01±4.95	29.50±12.46	74.23±7.79	22.09±5.51	78.80±4.96	25.42±8.08	**79.27**±5.37
Multi-Day	23.32±13.05	72.41±9.47	**32.42**±12.14	**80.42**±8.81	21.60±5.23	78.41±4.81	27.40±10.08	73.55±8.46	21.42±6.81	78.96±5.42	25.36±7.50	80.28±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.

- **Retrieval-Augmented Generation (RAG)** consistently improves **factuality, thoroughness, and overall usefulness** compared to direct generation.
- **Chain-of-Thought (CoT)** increases interpretability but offers limited gains in summary quality.
- **Assessment & Plan (A&P)** summaries outperform **Discharge Summaries (DS)** — structured, focused inputs lead to more coherent outputs.
- Across models, **temporal reasoning remains limited**: LLMs often misorder events or omit key timeline details.
- **Quantitative metrics (ROUGE-L, BERTScore, F1)** show moderate gains with RAG; absolute scores remain below human-level summaries.
- **Expert evaluation** confirms RAG reduces hallucination and improves clinical relevance, but redundancy and missing causal links persist.
- Overall, **RAG + A&P** provides the most balanced and reliable longitudinal performance.

		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

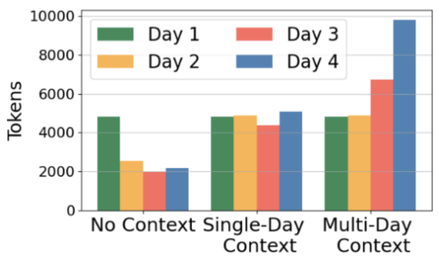
Table 6: Results on the binary diagnosis prediction task using EHRShot data (macro-average).

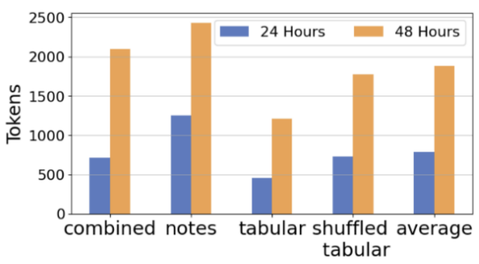
Setting	Model	ROUGE-L	BERTScore
Direct Gen	Mistral	16.28 ±12.83	65.23 ±13.38
	Llama3	11.82 ±14.42	55.35 ±22.48
	Qwen	15.28 ±5.38	63.98 ±13.50
	DeepSeek	13.51 ±5.04	63.64 ±13.86
	Llama2	15.34 ±5.90	63.82 ±11.44
RAG	Mistral	15.04 ±2.34	65.89 ±5.69
	Llama3	15.90 ±2.00	64.38 ±4.81
	Qwen	**17.91** ±2.35	65.25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

Discussion

- **Longer context inputs** enhance information coverage but not temporal reasoning—models recall facts, not timelines.
- **RAG** improves factual grounding but disrupts narrative coherence due to context fragmentation.
- **CoT** reasoning helps explainability but doesn't strengthen accuracy without explicit temporal modeling.
- **Low citation scores** reveal weak evidence attribution—LLMs summarize confidently but without source traceability.
- **Expert feedback** highlights recurring temporal gaps and omission of evolving diagnoses.
- Findings expose a gap between **data integration** and **causal reasoning** in current LLMs.
- Indicates need for **timeline-aware training** and **causality-grounded reasoning objectives** for clinical summarization.





Conclusion

- Long contexts ≠ better reasoning
- RAG reduces hallucination but fragments narrative flow
- CoT adds interpretability, not factual precision
- Key gap/future work: causal, temporally grounded LLMs for clinical use

Acknowledgements

This work is supported by U.S. National Library of Medicine R00 LM014308