AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow

Huizi Yu, MS[1], Jiayan Zhou, PhD[2], Lingyao Li, PhD[1], Themistocles L. Assimes, MD[2], Danielle S. Bitterman, MD[3,4], Xin Ma, PhD[5], Lizhou Fan, PhD[1,3,4]

1. University of Michigan, Ann Arbor, MI, USA
2. Stanford University, Stanford, CA, USA
3. Artificial Intelligence in Medicine Program, Mass General Brigham, Boston, MA, USA
4. Harvard Medical School, Boston, MA, USA
5. School of Control Sciences and Engineering, Shandong University, Ji'nan, Shandong, China

MOTIVATION

- Simulated patient systems play a crucial role in medical training and evaluation.
- Challenges with current simulated patient systems include limited intelligence, lack of diverse patient profiles, and trustworthiness concerns.
- Large Language Models offer an opportunity to enhance realism and effectiveness [1].

METHOD

- We developed an LLM-powered simulated patient system ***AIPatient*** incorporating the **AIPatient Knowledge Graph (KG)** and **Reasoning RAG** agentic workflow.

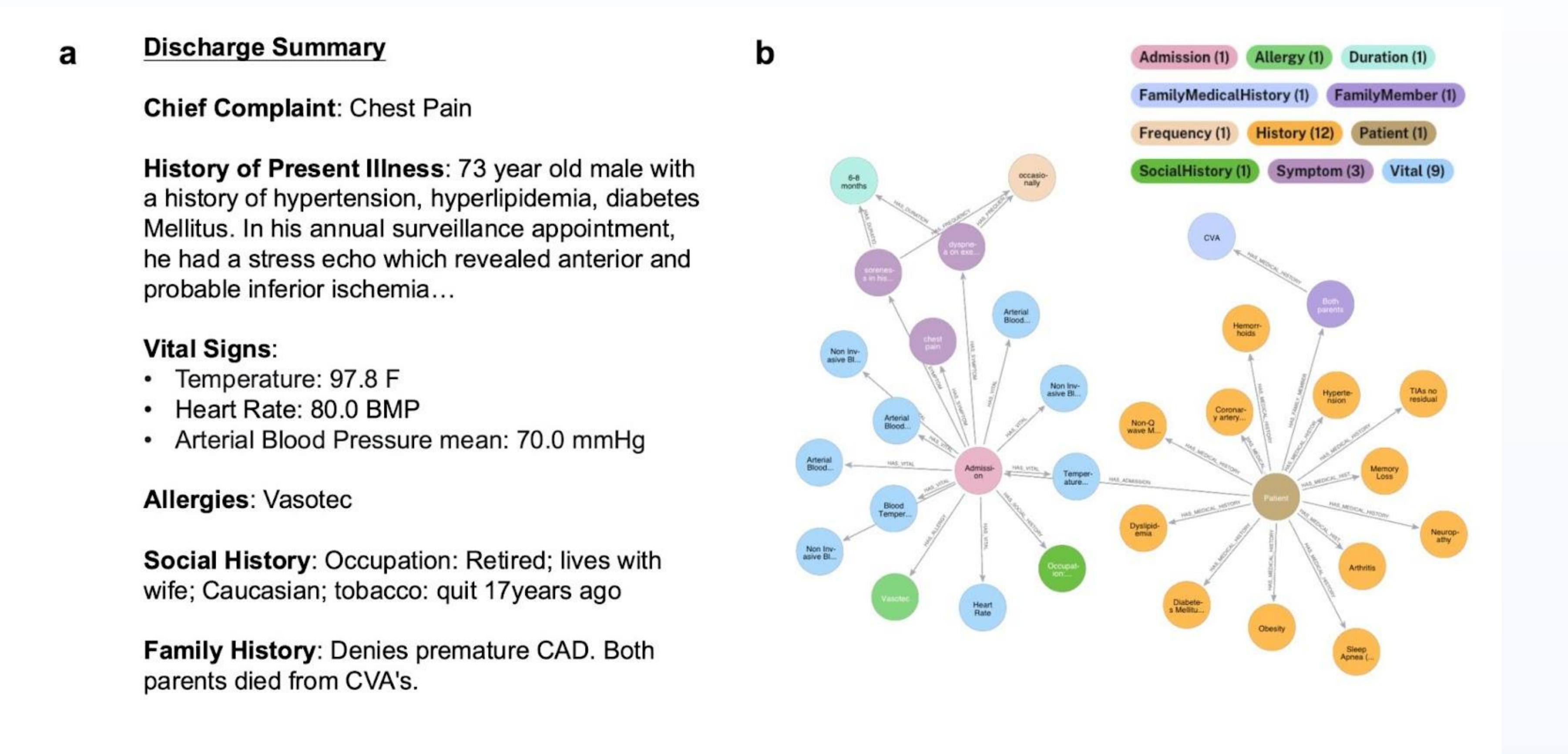

Figure 1: **Data transformation of MIMIC-III EHRs** [2] from **(a)** raw discharge notes (with extracted entities) to **(b)** constructed AIPatient Knowledge Graph (through NER).

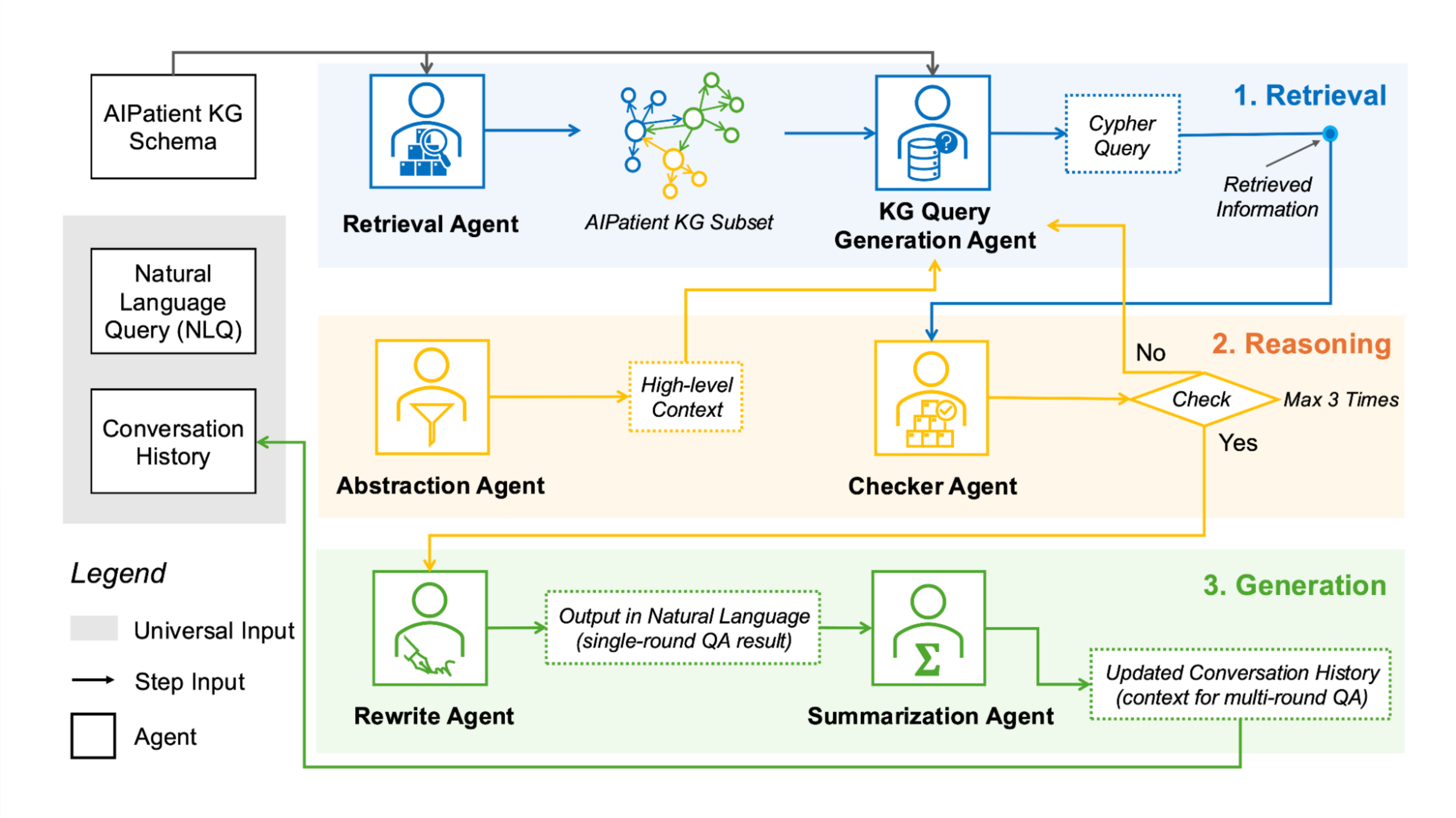

Figure 2. Reasoning RAG agentic workflow is the AIPatient system's processing backbone, comprising three key stages: retrieval, reasoning, and generation. It first retrieves relevant information from the knowledge graph, then applies contextual reasoning to reduce hallucinations, and finally generates natural language responses based on conversation continuity and tailored to the perceived patient personality.

EVALUATION FRAMEWORK

Table 1. Evaluation framework

Performance aspect	Evaluation dimension	Evaluation by	Metrics
Effectiveness	Knowledgebase validity (NER)	Medical doctors	F1
	QA accuracy (conversation)	Researchers	Accuracy
	Readability	Algorithm	Flesch Reading Ease, Flesch-Kincaid Grade Level
Trustworthiness	Robustness (system)	Researchers	Accuracy, ANOVA
	Stability (personality)	Researchers	Accuracy, ANOVA

RESULTS

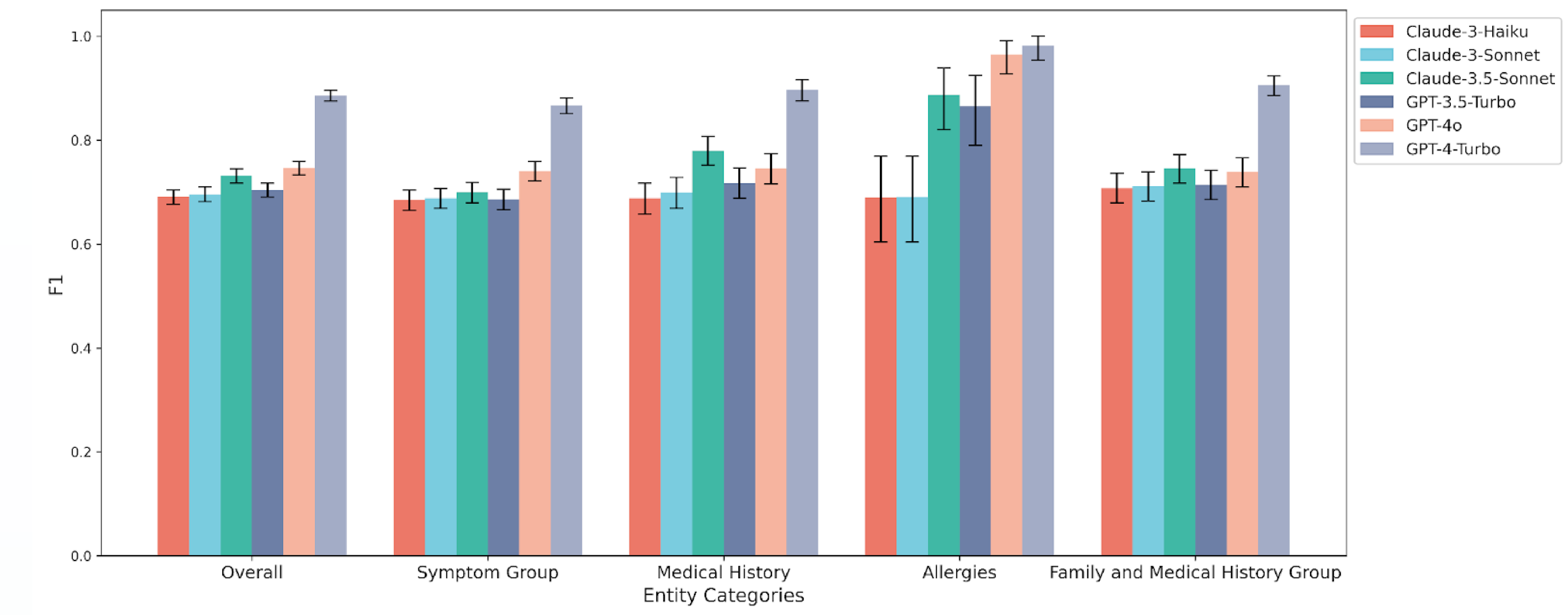

Figure 3: Comparison Knowledgebase validity across LLMs; F1 scores for different LLMs across medical entity categories. GPT-4 Turbo achieves the highest overall performance, particularly in Allergies and Medical History, while Claude models show lower scores in these categories. Whiskers indicate 95% credible intervals from 10,000 bootstrap iterations.

Table 2 Ablation Studies Result[1] of QA Accuracy by Medical Category

Few Shot	Retrieval Agent	Abstraction Agent	Overall	Symptom Group	Medical History	Family and Social History
✓	✓	✓	**94.15%**[3]	**91.20%**	**87.10%**	**85.56%**
✓	✓		92.60%	89.68%	83.87%	78.89%
✓		✓	93.80%	90.48%	83.87%	**85.56%**
✓			92.94%	90.48%	69.35%	82.22%
	✓	✓	81.41%	85.71%	25.81%	60.00%
	✓		81.93%	84.92%	27.42%	58.89%
		✓	83.13%	87.20%	30.65%	64.44%
Only with *KG Query Generation Agent*			82.62%	88.80%	25.81%	60.00%
Without *Reasoning RAG* & Without *AIPatient KG*			68.94%	64.29%	53.45%	13.33%

Table 2: Accuracy across different AI agent configurations; The highest accuracy (94.15%) is achieved with Few Shot Learning, Retrieval, and Abstraction Agents, showing their combined effectiveness. The Retrieval Agent improves performance across all categories, while the Abstraction Agent particularly enhances Family and Social History (85.56%).

Readability, Robustness and Stability:

- Median **Flesch Reading Ease = 77.23**, suitable for medical trainees.
- Robustness confirmed with non-significant variation in accuracy (ANOVA $p > 0.1$).
- Stability confirmed across **32 personality types** (ANOVA $p = 0.799$).

Out-of-Distribution (OOD) Generalization:

- AIPatient performs well on **CORAL dataset (oncology reports)** with QA accuracy of **81.04%**.

CONCLUSION

Key Insights:

- The multi-agent AI framework significantly improves simulated patient realism and intelligence.
- Reasoning RAG enhances accuracy, reliability, and patient interaction fidelity.
- AI-generated patient simulations can support medical education, clinical decision-making, and model evaluation.

Limitations & Future Work:

- Expanding the patient dataset to include diverse demographics.
- Reducing computational cost and improving response speed.
- Ethical considerations and user feedback collection.

REFERENCE

[1] Li, Y., Zeng, C., Zhong, J., Zhang, R., Zhang, M., & Zou, L. (2024). Leveraging large language model as simu-lated patients for clinical education. In arXiv [cs.CL]. arXiv.

[2] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.