RECO: Recovery Companion

A chatbot-based solution to monitor heart failure patients after hospital discharge

Logo generated by AI

Agenda

1. The Problem Space
2. MVP Demonstration
3. Technical Approach and Model Evaluation
4. Technical Takeaways and Future Roadmap
5. Conclusion

The Problem Space

Heart Failure: A Problem

Heart failure is a growing clinical and economic problem, with hospital readmissions being a major burden

U.S. HF Prevalence (2012)

1 in 33 ↑ 46% (2012-30F)

U.S. HF Direct Medical Costs (2012)

$21B ↑ 152% (2012-30F)

U.S. HF Direct Medical Costs (2030F)

$53B

Hospitalization Contribution to Direct Medical Costs

49-73%

Patients Readmitted Within 30 Days

1 in 5

Sources: *Journal of the American College of Cardiology* 79.17 (2022): e263-e421; *J Manag Care Spec Pharm.* 2022 Feb; 28(2): 10.18553/jmcp.2022.28.2.157.

Issues With Post-Discharge Monitoring

Experts highlight that challenges with monitoring and gaps in existing solutions drive hospital readmissions

Monitoring patients post-discharge is challenging

Exacerbation detection
Challenges in early detection of worsening symptoms and vitals to avoid readmission

Medication management
Difficulty in ensuring patients follow prescribed schedules

Resource constraints
Limited resources to monitor patients post-discharge

Current solutions do not adequately address these challenges

Current solutions
Predominantly based on forms or rule-based chatbots

Limitations

- Unintuitive
- Low patient engagement
- Resource-intensive
- Poor systems integration

Result: Reactive (vs. proactive) management contributing to readmissions

The Solution: RECO

A generative AI-powered chatbot to routinely check on patients' health and report to their doctors

Target User

RECO targets high-risk patients and their healthcare providers

The Patient

Recently discharged patients aged 40 and older at risk of HF complications and readmissions

The HCP

Cardiologists in outpatient settings who manage these high-risk HF patients

Our Mission

Empowering heart failure patients and healthcare providers through scalable, generative AI-driven patient monitoring

$770 - $1,870

Estimated Annual U.S. Cost Savings per Patient (2023)

$850m - $2.1bn

Estimated Total U.S. Annual Cost Savings (2023)

The MVP

MVP Overview

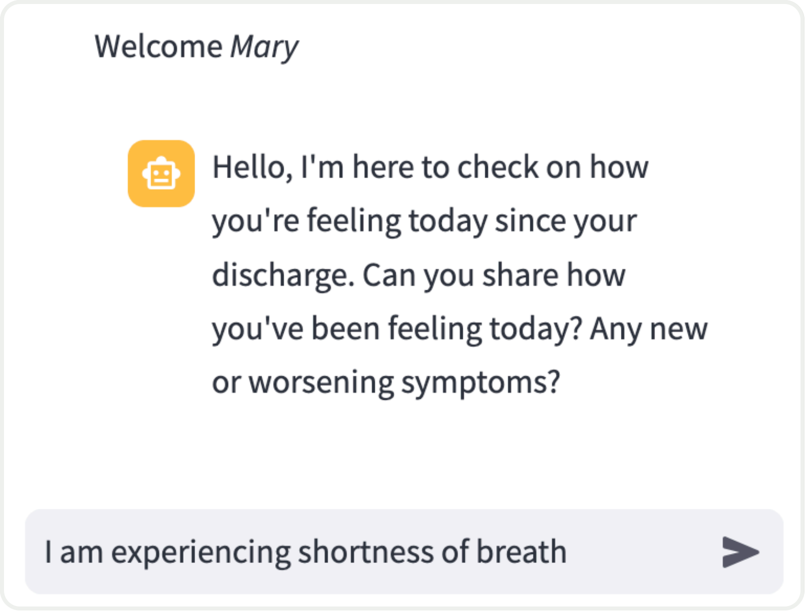
A medical chatbot and summarization engine for improved patient monitoring

Benefits for Patients

- Convenient
- Engaging
- Supportive

Benefits for Doctors

- Efficient
- Comprehensive
- Proactive



RECO Demo

User Feedback

“I would prefer this over filling out forms!”

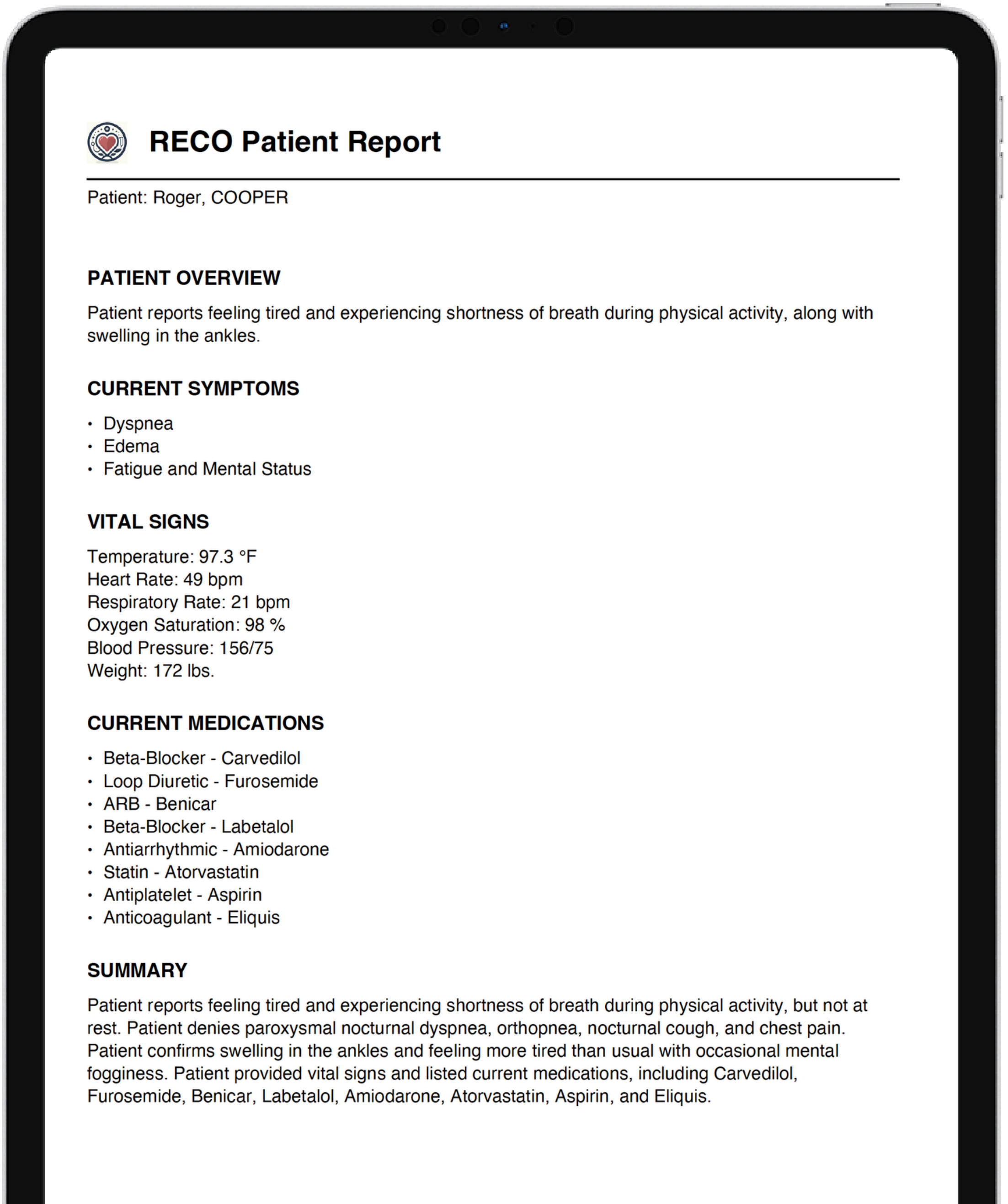
Patient

“Real-time monitoring through the chatbot allows us to provide personalized care, improving our heart failure patients' quality of life”

Cardiologist

“Easier and more straightforward than a doctor’s visit”

Patient



RECO Patient Report

Patient: Roger, COOPER

PATIENT OVERVIEW

Patient reports feeling tired and experiencing shortness of breath during physical activity, along with swelling in the ankles.

CURRENT SYMPTOMS

- Dyspnea
- Edema
- Fatigue and Mental Status

VITAL SIGNS

Temperature: 97.3 °F
Heart Rate: 49 bpm
Respiratory Rate: 21 bpm
Oxygen Saturation: 98 %
Blood Pressure: 156/75
Weight: 172 lbs.

CURRENT MEDICATIONS

- Beta-Blocker - Carvedilol
- Loop Diuretic - Furosemide
- ARB - Benicar
- Beta-Blocker - Labetalol
- Antiarrhythmic - Amiodarone
- Statin - Atorvastatin
- Antiplatelet - Aspirin
- Anticoagulant - Eliquis

SUMMARY

Patient reports feeling tired and experiencing shortness of breath during physical activity, but not at rest. Patient denies paroxysmal nocturnal dyspnea, orthopnea, nocturnal cough, and chest pain. Patient confirms swelling in the ankles and feeling more tired than usual with occasional mental fogginess. Patient provided vital signs and listed current medications, including Carvedilol, Furosemide, Benicar, Labetalol, Amiodarone, Atorvastatin, Aspirin, and Eliquis.

User Impact and Interpretability

The summary PDF outputs are easily interpretable and addresses key pain points of HF patient monitoring

Key Benefits

- **Improved Data Accuracy**

  Minimizes human error in data interpretation and recording

- **Enhanced Decision-making**

  Provides concise, relevant information for quicker, better-informed decisions

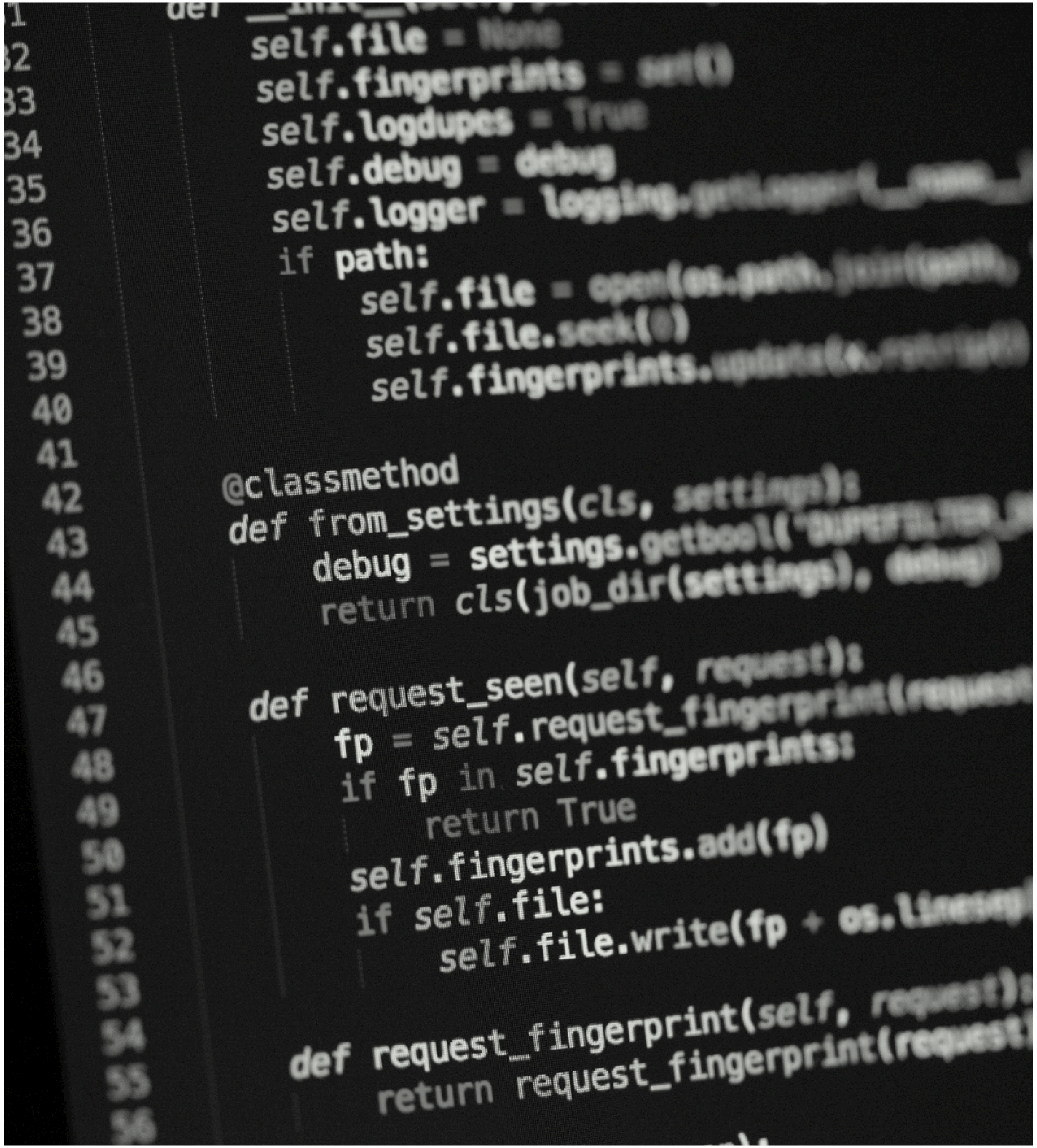
- **Consistency**

  Ensures uniform reporting format for easier comparison and analysis

- **Scalability**

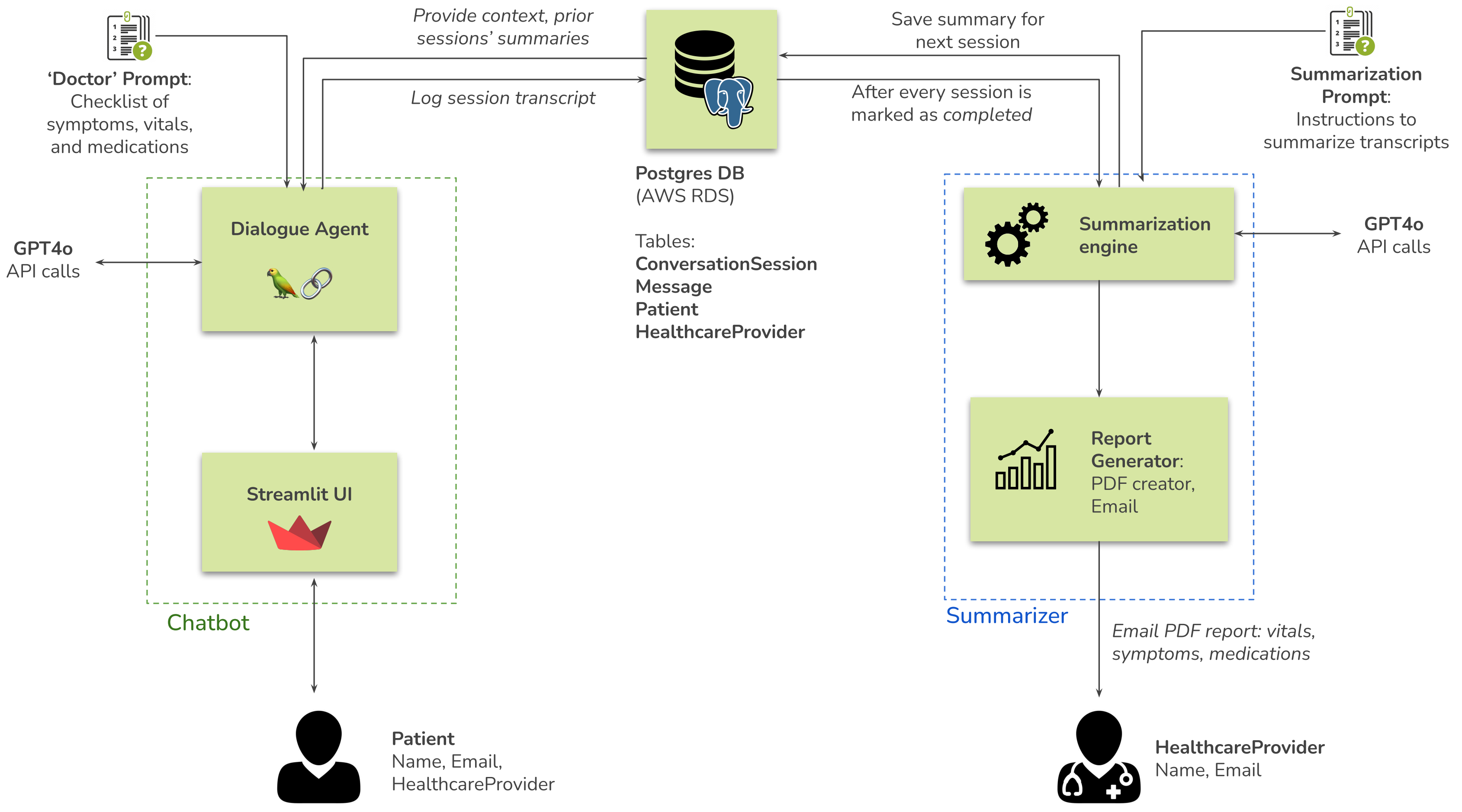
  Enables management of larger patient volumes without overburdening HCPs

Technical Approach and Evaluation

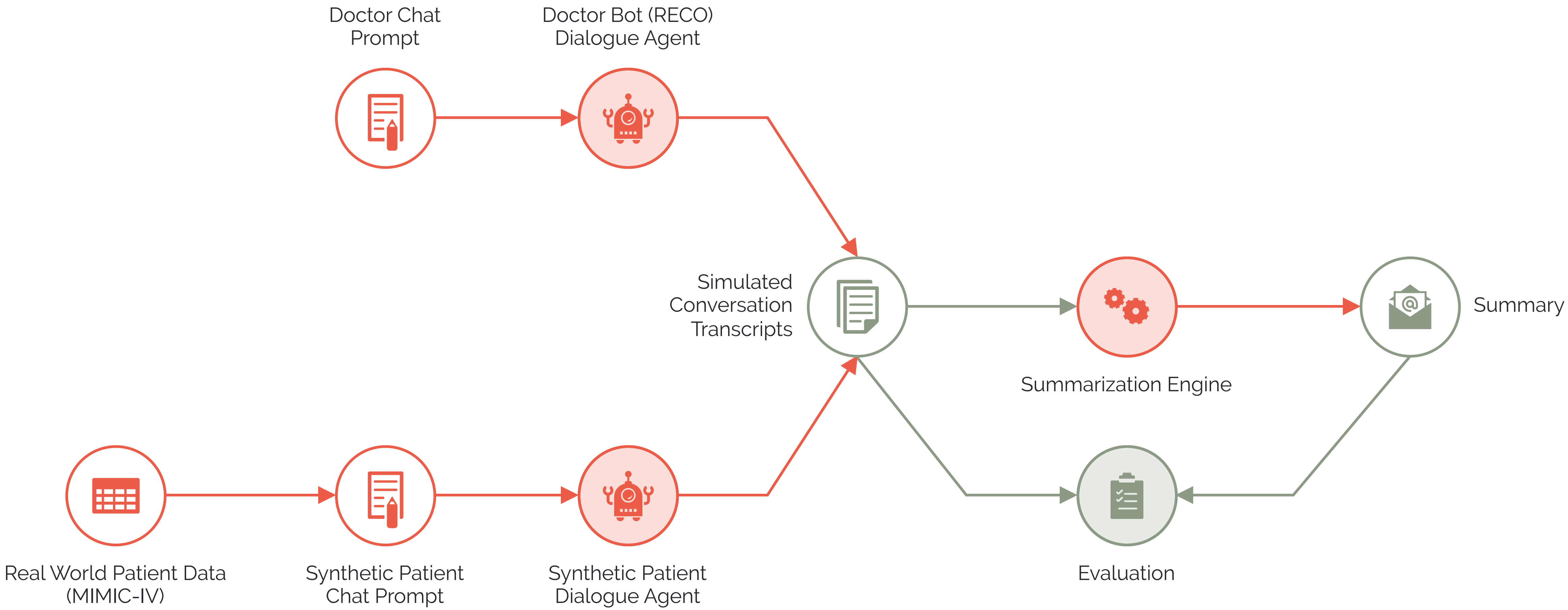


Overall Architecture

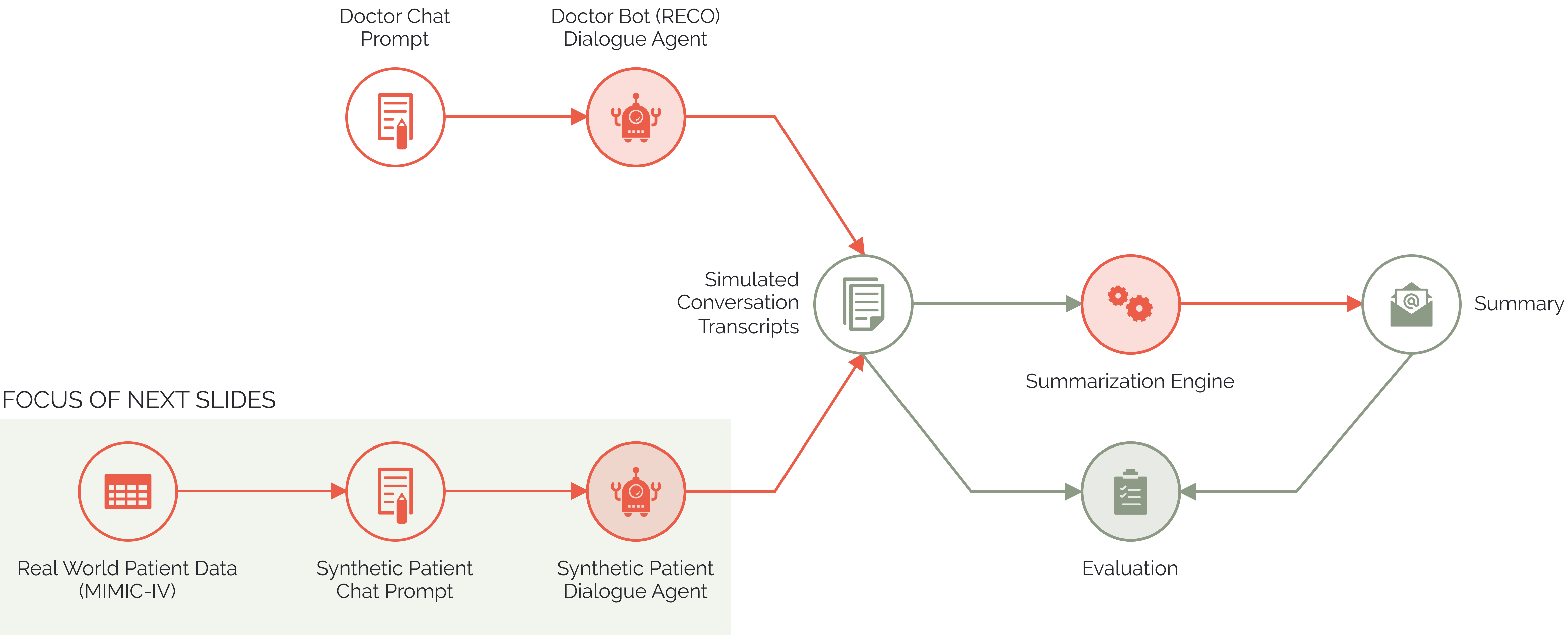
Our architecture integrates a UI, chatbot agent, database, and summarizer



Modeling Approach: Overview

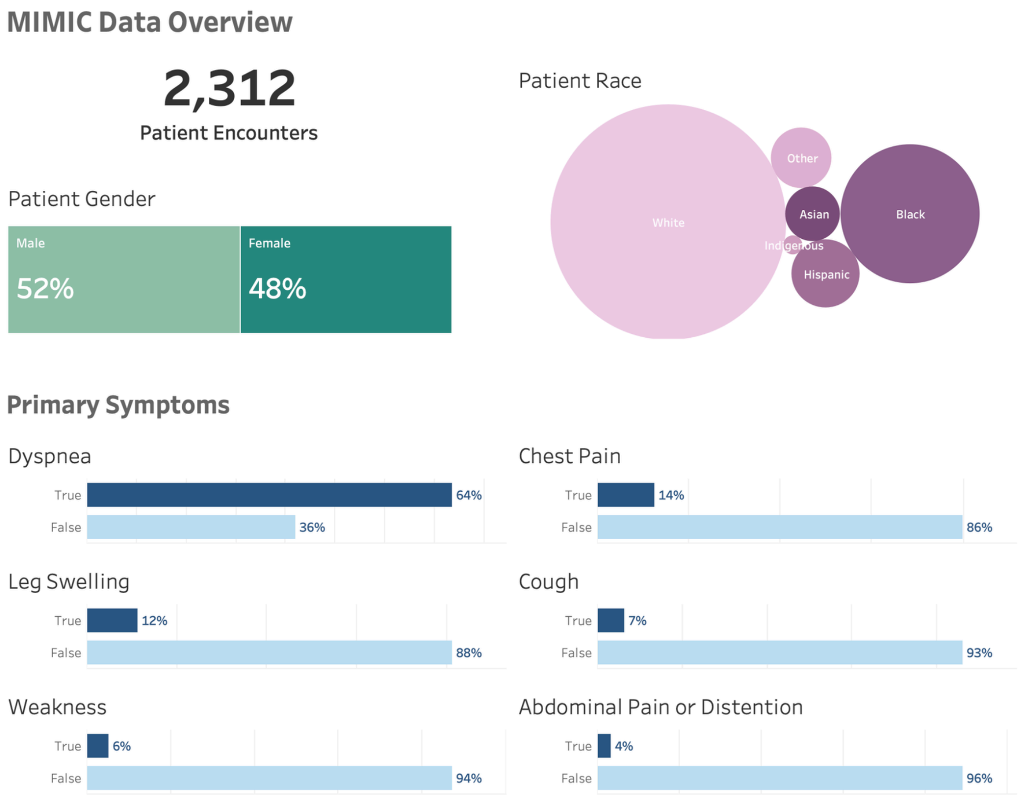


Synthetic Patients



The Dataset: MIMIC-IV

We used real world data from MIMIC-IV to create our synthetic patients



- Large, publically-available database
- De-identified patient data including symptoms, vitals and medications
- Heart failure patients selected for clinical appropriateness for our application

Source: Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2). *PhysioNet*. https://doi.org/10.13026/6mm1-ek67.

Synthetic Patient Generation

Realistic patient prompts derived from MIMIC-IV data drive our synthetic patient Dialogue Agents

Extract Patient Data From MIMIC-IV → **Feed to Patient Prompt Templates** (Cooperative and Reluctant Personas) → **Feed to Synthetic Patient Dialogue Agent**

Demographics

- **Age:** 91
- **Race:** White
- **Marital Status:** Widowed

Clinical Characteristics

- **Symptoms:** Respiratory Distress
- **Medications:** Carvedilol, Furosemide

Vitals

- **Temperature:** 97.4
- **Heart Rate:** 82 bpm
- **Respiratory Rate:** 22 bpm
- **O2 Saturation:** 98%
- **Blood Pressure:** 126/74 mmHg
- **Weight:** 93 Lbs

Example: Cooperative Patient Prompt Template

You are {name}, a patient who has been discharged after a hospital stay for heart failure. You are reporting your symptoms for a routine check-in with your doctor. Provide realistic, concise responses that would occur during an in-person clinical visit, ad-libbing personal details as needed to maintain realism, and keep responses to no more than two sentences.

Use the input during the conversation:

- **Age**: {Age}
- **Race**: {Race}
- **Marital Status**: {Marital Status}
- **Symptoms**: {Symptoms}
- **Medications**: {Medications}
- **Temperature**: {Temperature}
- **Heart Rate**: {Heart Rate}
- **Respiratory Rate**: {Respiratory Rate}
- **O2 Saturation**: {O2 Saturation}
- **Weight**: {Weight}

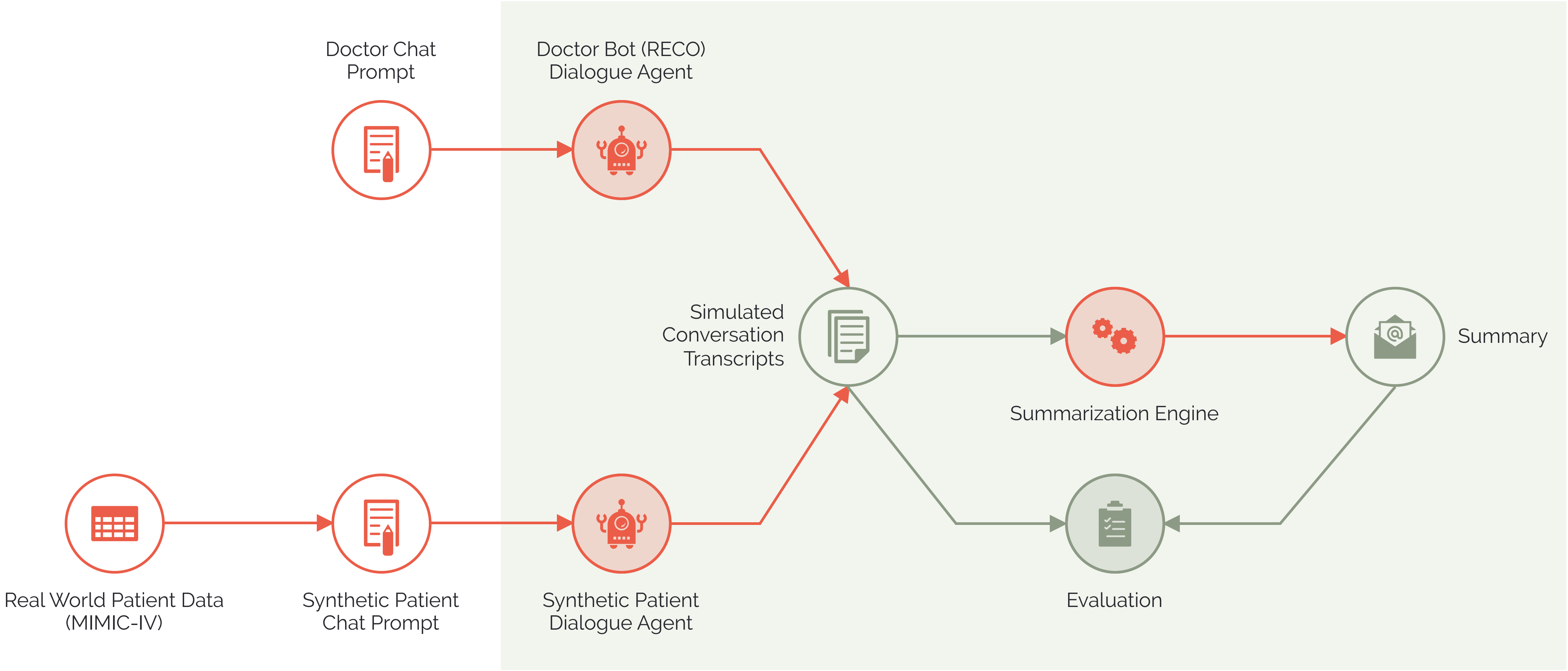
Synthetic Patient Validation

Dialogue generated by synthetic patients were validated using four criteria

Criteria for Assessing Synthetic Patients

- **Plain Language**
  Patient uses plain language
- **Consistency**
  Patients are consistent about their symptoms
- **Factual Accuracy**
  Patients do not invent information that contradicts the prompt (no confabulations)
- **Flow**
  Patients allow the doctor to ask questions and do not take over the conversation

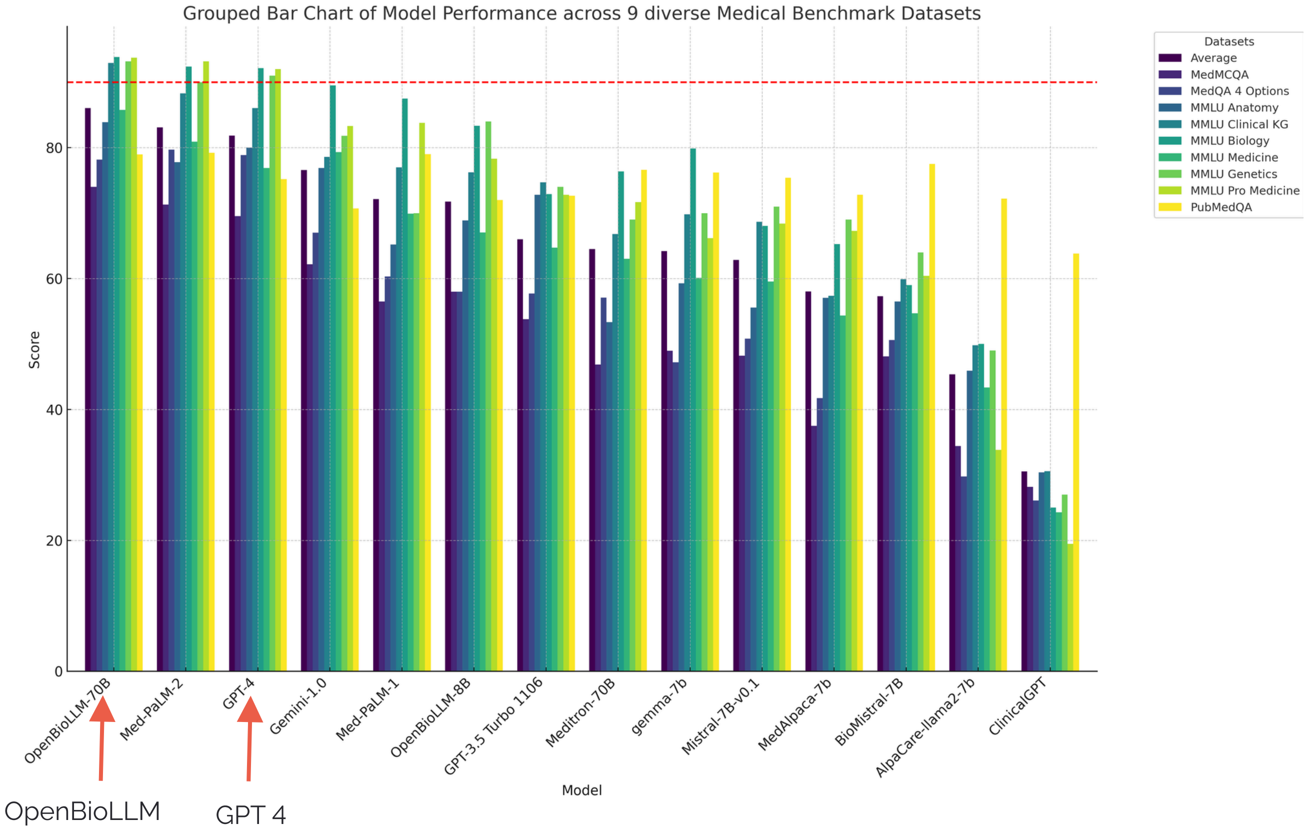
Model Selection, Evaluation and Prompt Engineering



FOCUS OF NEXT SLIDES

Model Selection

Dialogue Agents and the summarization engine are powered by OpenAI GPT given several key advantages over OpenBioLLM



OpenBioLLM (Llama3-based)

- ✔ Stronger biomedical understanding
- ✖ Requires expensive hosting

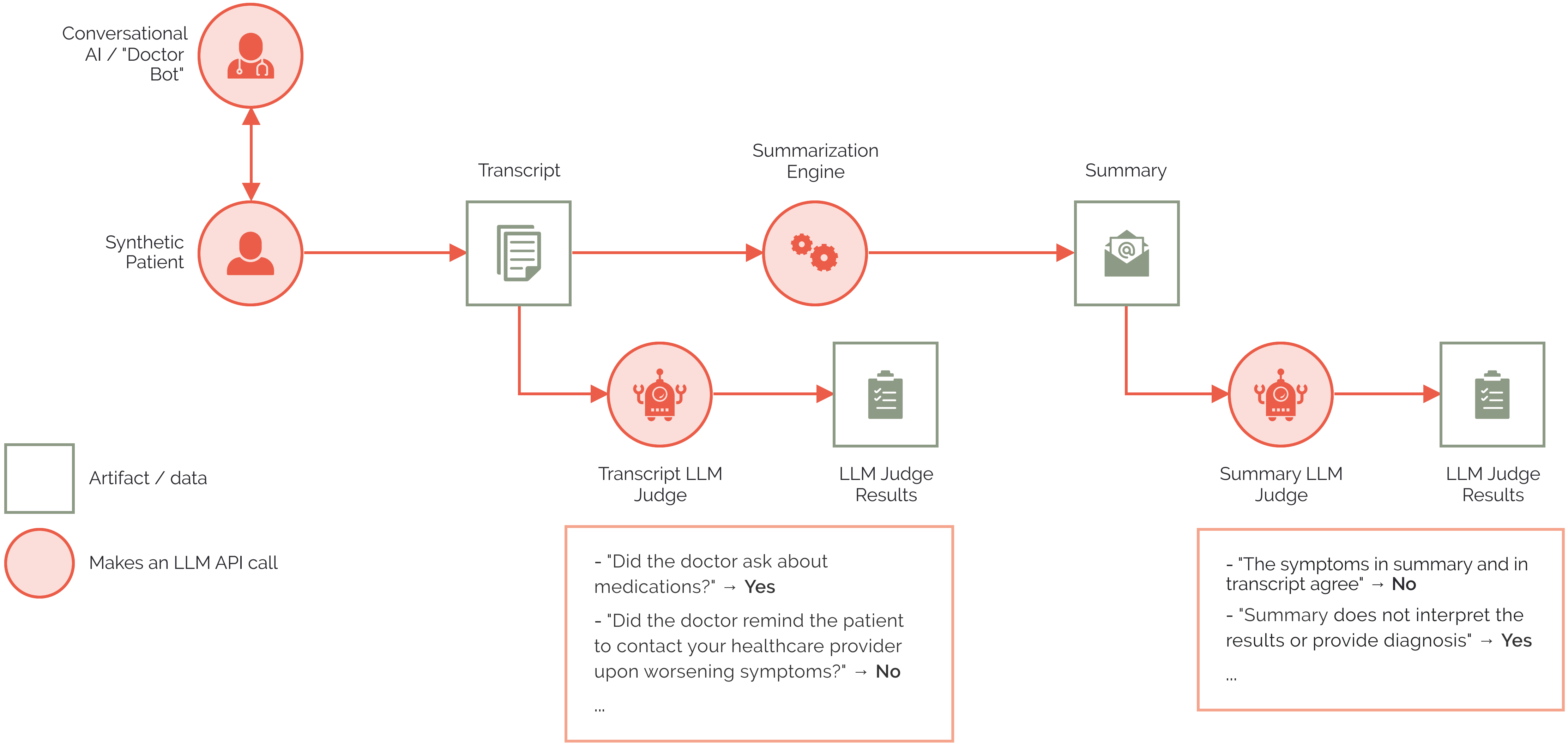
OpenAI's GPT 4o/4o-mini/3.5

- ✔ Longer context window
- ✔ Affordable
- ✖ Slightly inferior performance to OpenBioLLM in the biomedical domain

Source: Pal, Ankit, and Malaikannan Sankarasubbu. 'OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences.' Hugging Face repository, 2024. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

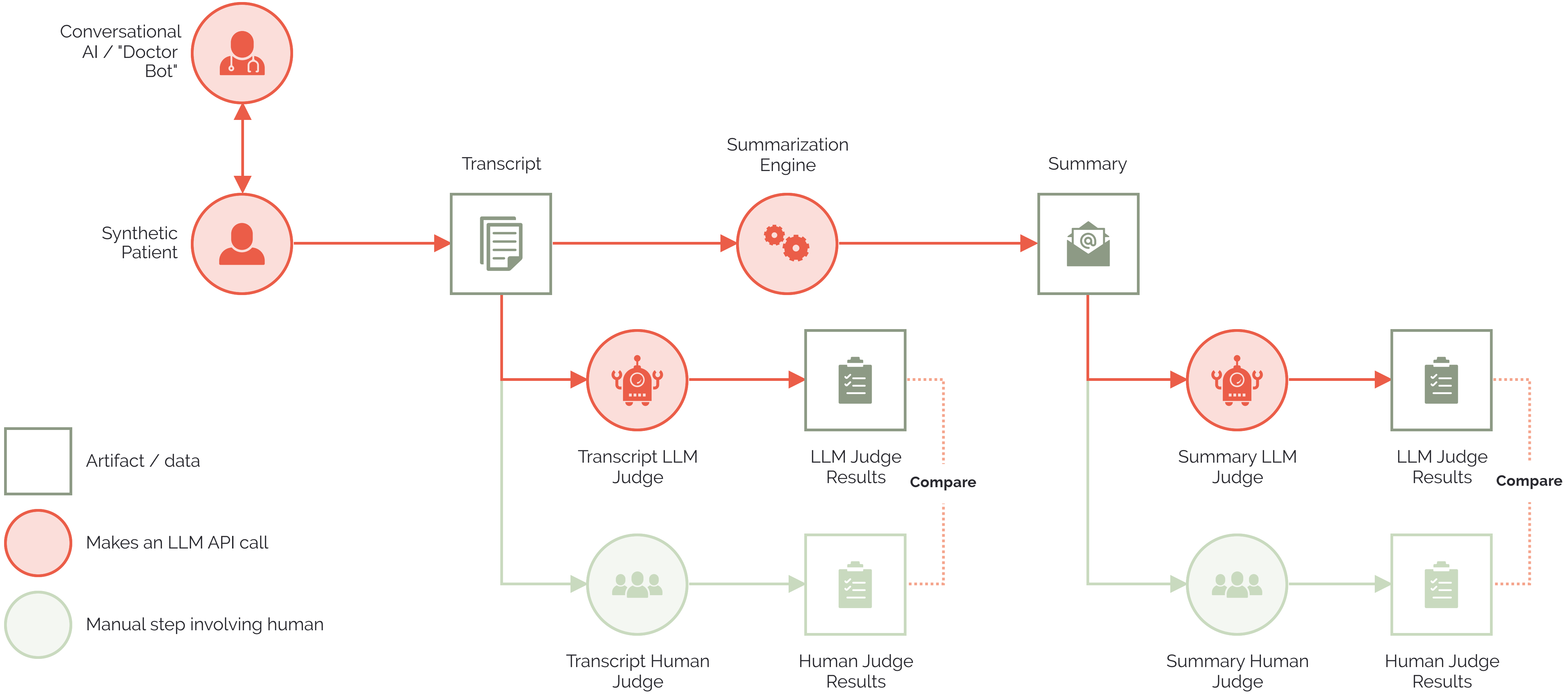
Performance Evaluation

We employed **LLM-as-a-judge** to quickly evaluate our Doctor Bot and Summarization Engine



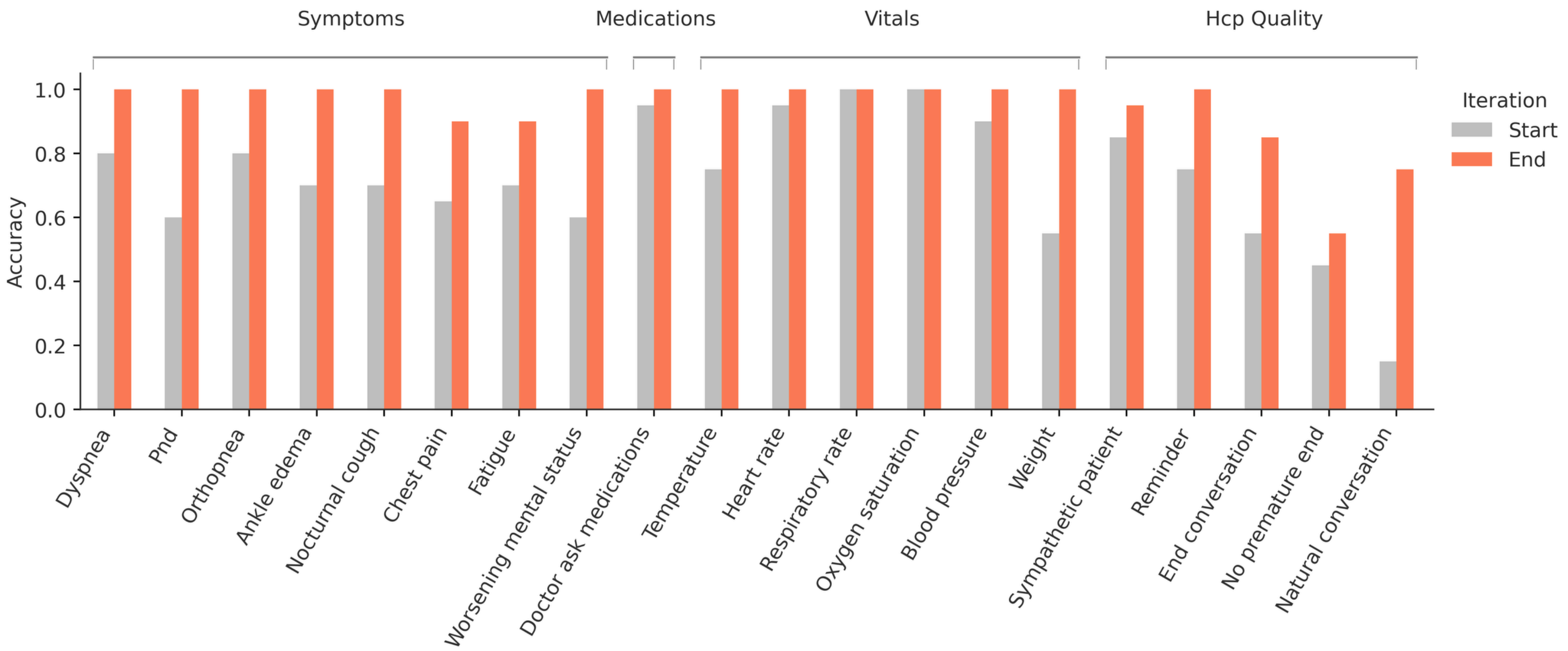
Performance Evaluation

To validate our **LLM-as-a-judge**, we compared it against human judgement, then iterate until they agree



After iterating on its prompt, the **Transcript LLM Judge** is able to meet the golden baseline of human judgement on most criteria questions.

Binary classification accuracy (yes/no)

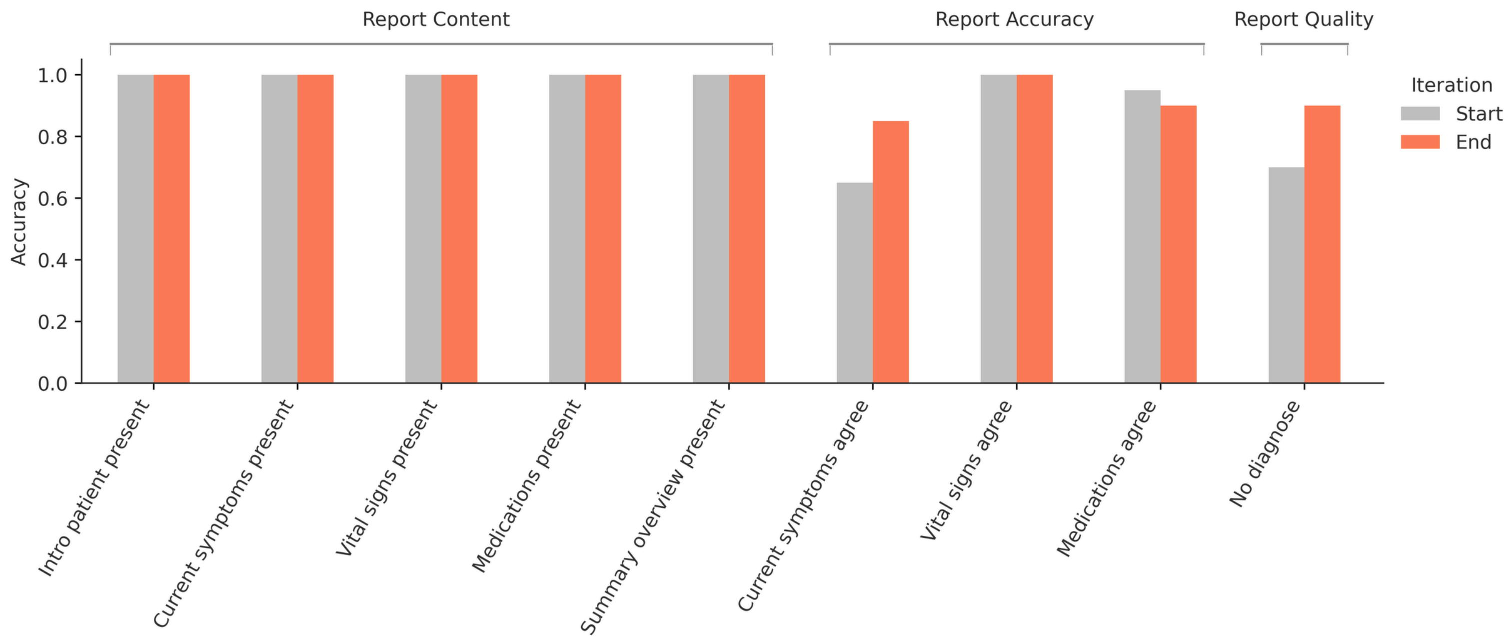


Iteration improvements made to Transcript LLM Judge

- Asking for a **reasoning** before "yes/no" judgement, instead of just "yes/no"
  - Before: pnd,0
  - After: pnd,"Doctor did not ask about PND in the conversation",0
- Ask to evaluate whether "doctor bot" was successful in **getting a patient confirmation** on symptoms (penalize "doctor bot" for not pushing for information)

Likewise, the **Summary LLM Judge** aligns well with the human judgement golden baseline after some iterations. With these two LLM judges validated, we have an *automated* way of evaluating our chatbot + summarization engine.

Binary classification accuracy (yes/no)



Iteration improvements made to Summary LLM Judge

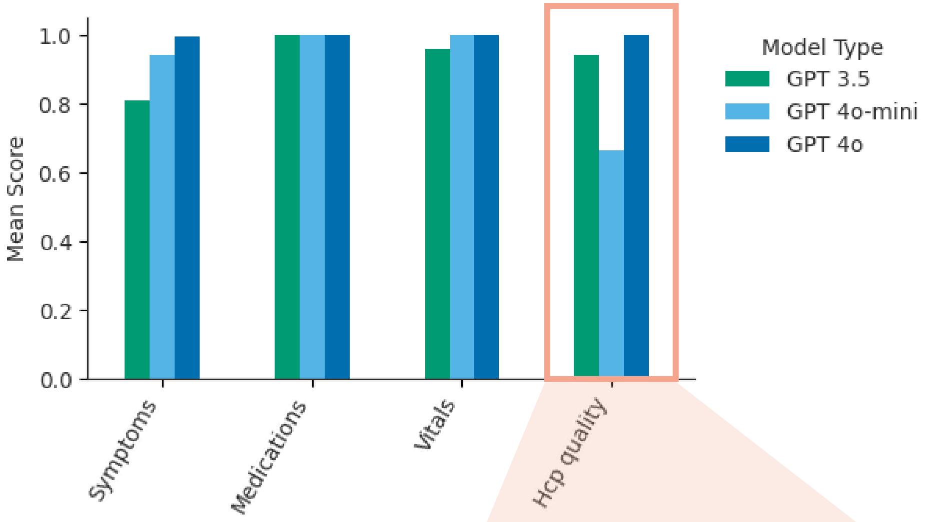
- Asking for a **reasoning** before "yes/no" judgement.
  - Before: vital_signs_agree,0
  - After: vital_signs_agree,"Heart rate in Summary is 130, but in Transcript it's 131",0
- **Split into 4 sub-evaluations**, revealing a single summary section to its relevant criteria at a time to prevent confusion. E.g. symptoms agreement should be checked against "current symptoms" section.
- Clarified **the meanings of "symptoms agreement" and "diagnosis,"** and add examples of edge cases and exemptions.

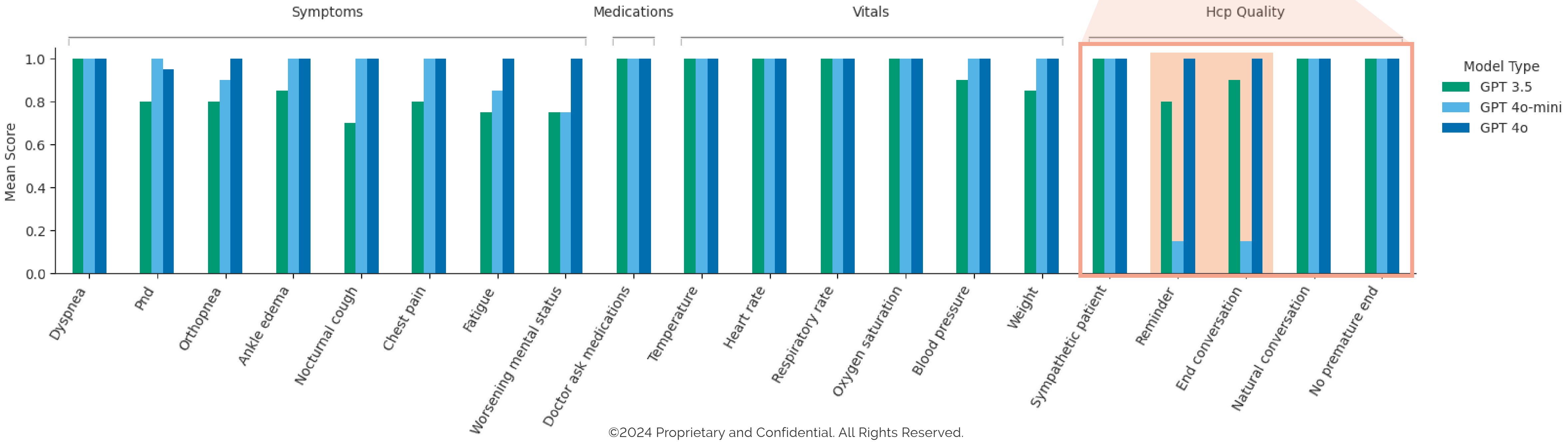
Improvement 1

GPT 4o outperforms GPT 3.5 and GPT 4o-mini

While GPT 4o-mini showed promise in its affordability, it lacked the ability to give a reminder and end the conversation.

GPT 4o has a near-perfect score and is picked for subsequent experiments.





Improvement 2

The doctor bot prompt was modified to better handle a "**reluctant patient**" persona

Example improvement: few-shot learning

Aim to follow the good examples and avoid the bad examples when probing or redirecting:

<examples_probing_reassurance>

- Example 1:

    Patient: I'm not really sure about the swelling. I mean, I sometimes feel a bit strange in my legs, but I don't want to say it's swelling without looking.

    Doctor:

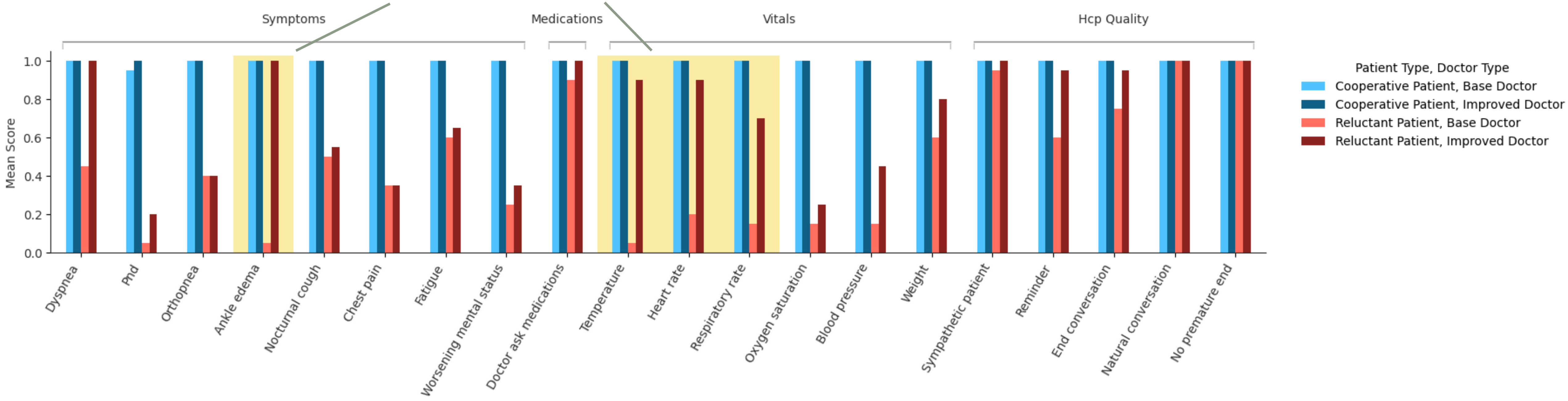
    - Good response: **Can you check and let me know if there's any visible swelling right now?**

    - Bad response: I understand your concern. Let's move on. Can you tell me if you've experienced any coughing at night? </examples_probing_reassurance>

- Example 2: ...

</examples_sufficient_insufficient>

Drastic improvements observed in ankle edema and some vitals



Key Innovations and Future Roadmap

Key Innovations

Clinically informed generative AI application for heart failure patients

Generated High Quality Synthetic Patient Profiles

Using de-identified medical data for realistic synthetic profiles let us safely develop and test our application

Established Scalable LLM Evaluation Framework

A novel approach that rigorously analyzed results as we improved our application's performance

Developed Clinically Informed End-to-End Monitoring Solution

Empowering heart failure patients and their doctors with comprehensive monitoring

Future Roadmap

Multiple Patient Interactions

Track and report patients' progress over time using context awareness

Enhanced Medication Management

Record initial medications and dosages, then ask patient to verify compliance over time

EHR Integration

Integrate the platform with EHR systems to better integrate RECO into clinical workflows

Our Mission

Empowering heart failure patients and healthcare providers through scalable, generative AI-driven patient monitoring


Thank You

Appendix

Market Opportunity

The estimated U.S. Serviceable Addressable Market for RECO is ~$130 million, growing at 2-3% p.a.

Est. U.S. Market Size for Heart Failure Monitoring Chatbot (2023)	Data	Source
Total Heart Failure Patients (# of Patients)	7.2m	J Card Fail. 2023 Oct; 29(10): 1412–1451.
Estimated Hospitalization Rate (% of Patients)	61% (Based on estimated 2018 hospitalization rate)	Clinicoecon Outcomes Res. 2023 Feb 24:15:139-149. eCollection 2023; JAMA Cardiol. 2021 Aug; 6(8): 1–5; Circulation. 2021 Feb 23;143(8):e254-e743.
Total Addressable Market (# of Hospitalized Patients)	**4.4m**	**Estimate**
Eligible for RECO (% Hospitalized Patients Readmitted in 30 Days)	25% (30-day readmission rates: 21-22%)	J Manag Care Spec Pharm. 2022 Feb; 28(2): 10.18553/jmcp.2022.28.2.157.
Serviceable Addressable Market (# of Hospitalized Patients at Risk of Readmission)	**1.1m**	**Estimate**
Price per Patient (USD / year)	$120	Estimate, corresponding to 6-15% of expected cost savings per patient.
Serviceable Addressable Market (USD)	**~$130m**	**Estimate**
Annual Growth Rate, 2023-30F (%)	**2-3%**	**Solely based on forecasts for increased prevalence of heart failure patients**

Cost Savings from Telemonitoring

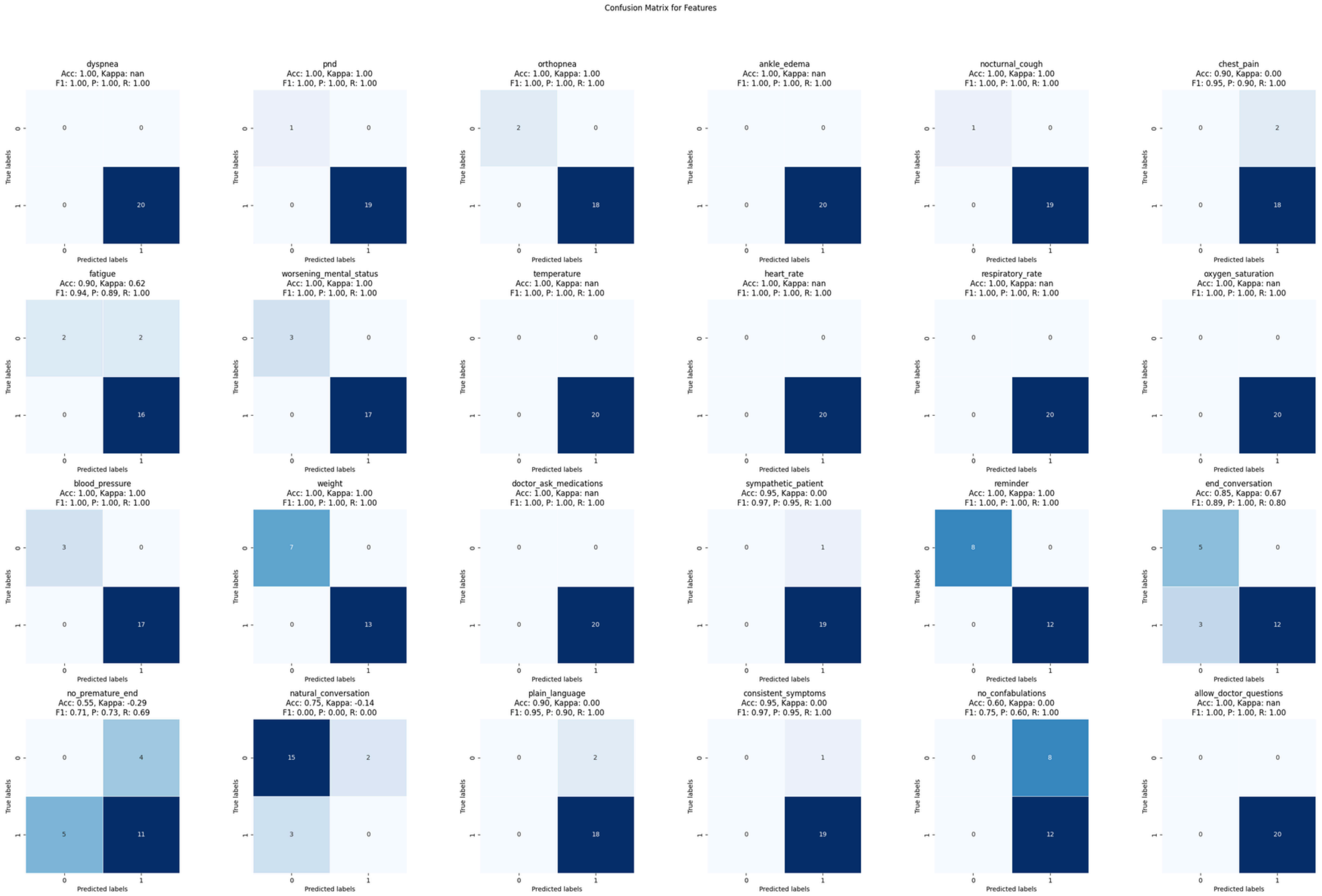
Estimated annual cost savings from avoided hospitalizations range from $770 - $1,870 per patient

Estimated Annual Cost Savings from Avoided Hospitalizations (2018, 23)

	2018	2023	Notes
Number of Heart Failure Patients	6m	6.7m	Sources: Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association; Descriptive Epidemiology and Outcomes of Patients with Short Stay Hospitalizations for the Treatment of Congestive Heart Failure in the US - PMC
Number of Heart Failure Hospitalizations	4.98m	Est. 5.56m	Source: Burden of hospitalization for heart failure in the United States: a systematic literature review - PMC. Note: Estimate assuming constant ratio of heart failure patients to hospitalizations in 2018 and 2023
Costs per Hospitalization	Est. $6.2k - $13.1k	Est. $7.1k - $15.0k	Source: Burden of hospitalization for heart failure in the United States: a systematic literature review - PMC. Note: Estimates based on 2012 Medicare figures of $5400-$11437 with annual healthcare inflation rates from https://www.usinflationcalculator.com/ applied.
Total Hospitalization Costs	$31.0bn - $65.6bn	$39.3bn - $83.3bn	Calculation: Number of Heart Failure Hospitalizations * Costs per Hospitalization
% Reduction in Hospitalizations from Telemonitoring	15%	15%	Source: Telemonitoring for heart failure: a meta-analysis | European Heart Journal | Oxford Academic
Total Reduction in Hospitalization Costs	**$4.7bn - $9.8bn**	**$5.9bn - $12.5bn**	**Calculation: Total Hospitalization Costs * % Reduction in Hospitalizations**
Reduction in Hospitalization Costs per Patient	**$770 - $1,640**	**$880 - $1,870**	**Calculation: Total Reduction in Hospitalization Costs * Number of Heart Failure Patients**

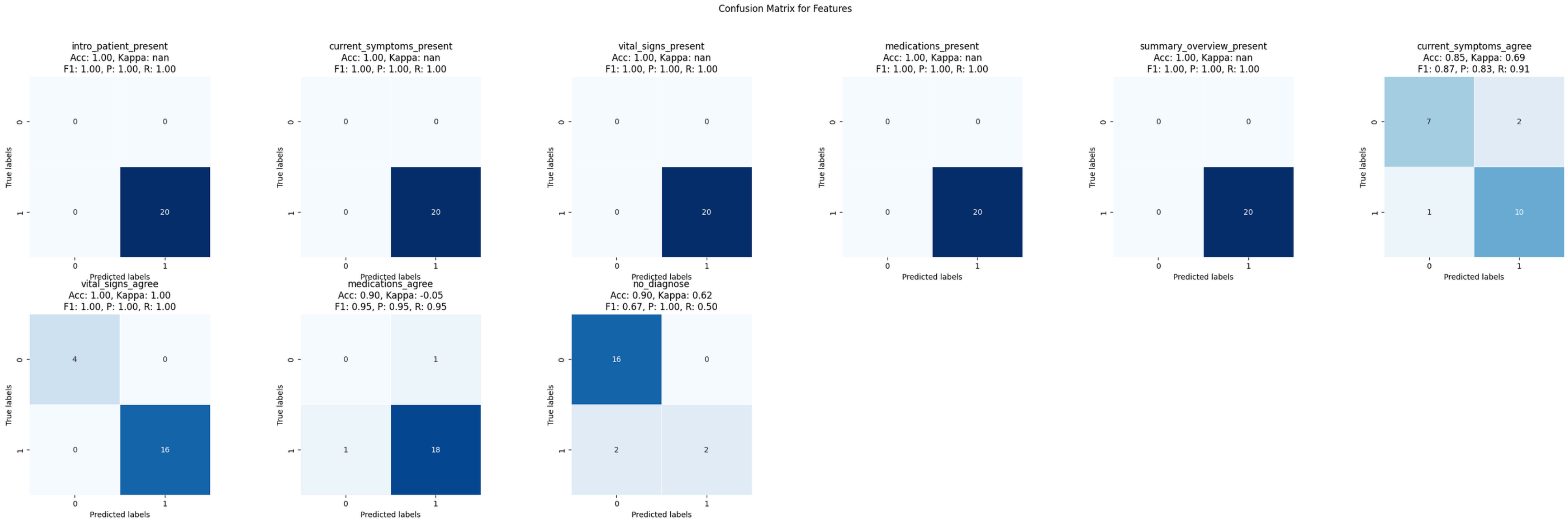
Confusion matrices for Transcript LLM Judge

Compared to ground truth of human judgement, after improvements made



Confusion matrices for Summary LLM Judge

Compared to ground truth of human judgement, after improvements made



Schema for PostgreSQL Tables