

KAIST AI기술설명회 2026

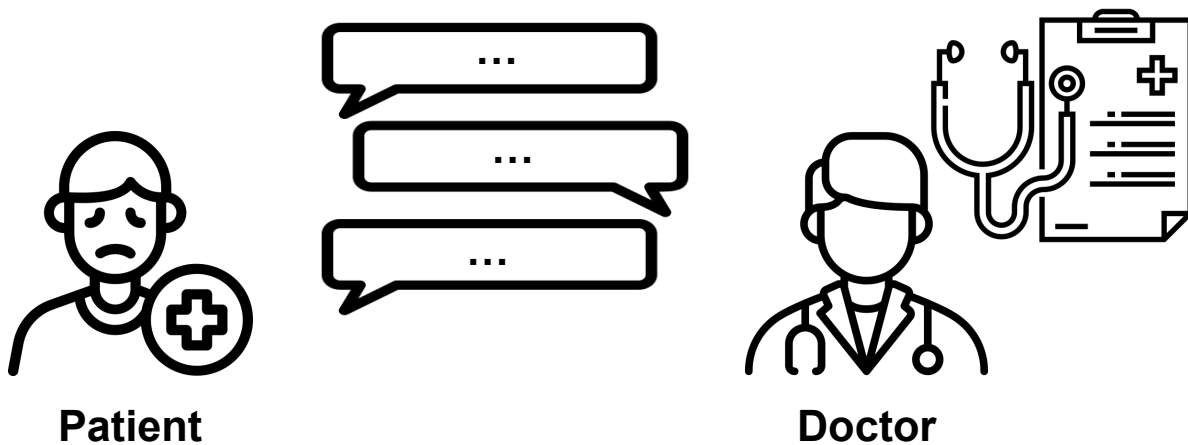
PatientSim: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions

KAIST 김재철시대학원 최윤재 교수님 연구실
경다운

Introduction

Motivation

- In clinical settings, doctors engage in multi-turn, context-aware conversations to gather patient information.
- Training or evaluating doctor LLMs in such settings requires realistic patient interaction systems.



Introduction

Motivation

- However, existing approaches have limitations:
 - 1) Mainly focus on evaluating Doctor LLMs, while neglecting the performance of Patient LLMs.
 - 2) Limited diversity in patient personas
 - Often provide only basic patient information, lacking detailed persona.
 - In clinical settings, doctor-patient interactions are deeply influenced by the patient's persona, such as emotional state or language proficiency.
- To train and evaluate Doctor LLMs effectively, we must develop realistic, reliable Patient LLMs that simulate rich, diverse, and human-like interactions.

Introduction

Problem Definition

- Current constraints for clinical simulation
 - 1. Data completeness issues
 - In real settings, doctors have access to comprehensive test / imaging data as they needed.
 - However, providing all possible data for every potential question in a simulation is impractical.
⇒ Limit interactions to patient's verbal response (symptoms, history) without test results.
 - 2. Inability to simulate longitudinal patient state changes
 - Predicting how treatment affects the patient between sessions is too complex for this project.
⇒ Focus on the single-session interaction (first-time visit scenarios)
- Current focus:
 - **Differential Diagnosis based on initial consultation in an Emergency Department (ED)**
 - Doctors rely on subjective information based on the patient's verbal descriptions (e.g., symptoms).
 - Test results may be unavailable, due to time constraints in the initial consultation.

Methodology

Patient profile construction

- We construct the patient profile based on the real-world clinical database, MIMIC-IV, ED and MIMIC-Note.

Real-World Clinical Data

| subject_id | gender | ... | pain | chief complaint |
|------------|--------|-----|------|-----------------|
| 5736479 | F | ... | 10 | dyspnea |
| ... | ... | ... | ... | ... |
| 5821254 | F | ... | 0 | nu... |

MIMIC-IV & ED

5736479
Service: NEUROLOGY
Allergies: ...
Chief Complaint: SOB, cough
History of Present Illness: ...

MIMIC-Note



```
{
  "subject_id": 5736479,
  "age": 77,
  "gender": "F",
  "marital_status": "married",
  "family_medical_history": "...",
  "medical_history": "...",
  "chief_complaint": "dyspnea, ...",
  "pain": 10,
  "medication": "...",
  ...
}
```

Profile #1

Methodology

Patient profile construction

- Target Diseases
 - Selection criteria:
 - High clinical prevalence and significance in ED settings
 - Sufficient representation in MIMIC-ED for robust model development
 - Distinct, differentiable symptom profiles suitable for diagnosis through history-taking alone
 - Final Disease Set
 - Myocardial Infarction (MI), Urinary Tract Infection (UTI), Pneumonia, Intestinal Obstruction, Cerebral Infarction

Methodology

Patient personality setting

- 1. Personality trait
 - Neutral (straightforward communication)
 - Impatient
 - Overanxious
 - Distrustful
 - Overly positive
 - Verbose (talkative)

- 2. Language proficiency
 - A scale describing language ability across three levels, following the CEFR framework
: A (Beginner), B (Intermediate), C (Proficient)

Methodology

Patient personality setting

- 3. Medical history recall level
 - Indicates the patient's ability to recall their medical history.
 - Patients may struggle to remember details, such as the names of medications or surgeries.
 - Categorized into two levels: high, low
- 4. Level of cognitive confusion
 - Describes the patient's emotional state at the time of admission.
 - Patients admitted to the ED often experience dazedness, regardless of their typical personality.
 - Categorized into two levels: highly confused, normal

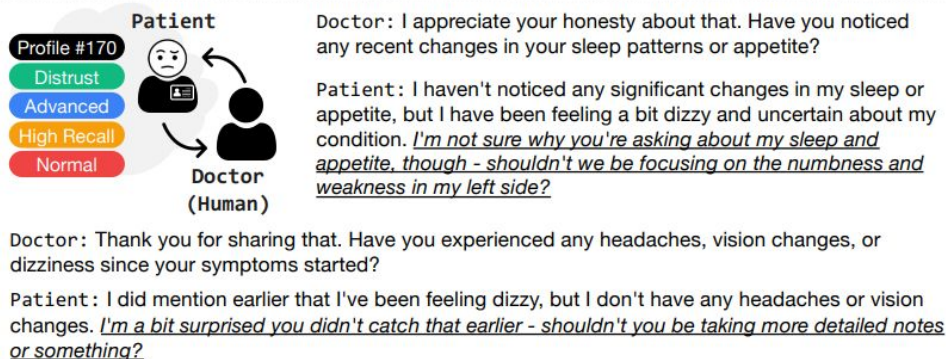
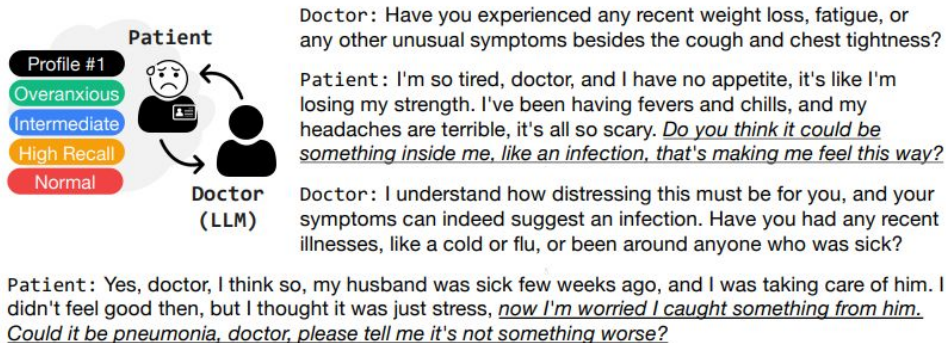
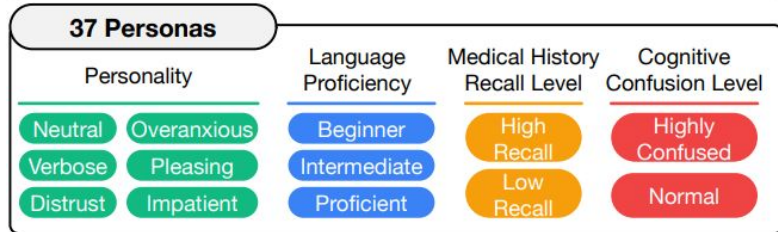
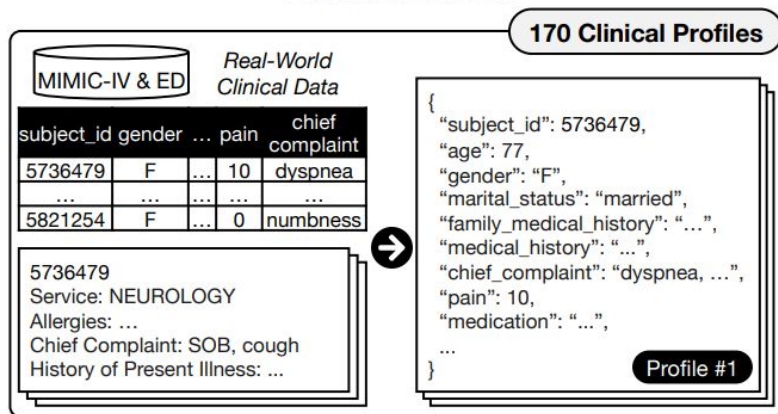
Methodology

Data statistics

- # of total patient profiles: 170 profiles
- # of unique personas: 37 distinct persona
 - 36 personas from combinations of:
 - 6 personality types
 - 3 language proficiency levels
 - 2 medical history recall levels
 - 1 special case persona:
 - high confusion with neutral personality, intermediate language proficiency, and high recall level
 - Note: Confusion modeled independently to avoid overlap with other low proficiency/recall personas

PatientSim Framework

PATIENTSIM



Experiment setting

Evaluation dimensions & protocol

- Evaluation dimensions:
 - Naturality & Realism: Do LLMs naturally reflect diverse persona traits in their responses?
 - Factuality: Do LLMs accurately derive responses based on the given profile?
 - Plausibility: Can LLMs reasonably fill in the blanks?
- Evaluation protocol:
 - Dialogue level:
 - Naturality & Realism (4-point scale)
 - Sentence level:
 - Factual accuracy (%)
 - Plausibility score (4-point scale)

Experiment results

RQ1: Do LLMs naturally reflect diverse persona traits in their responses?

- We use LLM-as-a-Judge to evaluate how realistically and naturally the LLM simulates each persona category (4-point scale).
- Based on the results, we select Llama-3.3-70B-Instruct as the backbone for PatientSim.

Table 1: Persona fidelity evaluation of various LLMs across five criteria, Personality, Language, Recall, Confused, and Realism, assessed by Gemini-2.5-Flash. Each criterion is rated on a 4-point scale. The average score (Avg.) summarizes overall performance.

| Engine | Personality | Language | Recall | Confused | Realism | Avg. |
|-------------------------------|-------------|----------|--------|----------|---------|------|
| Gemini-2.5-Flash | 3.94 | 3.54 | 3.64 | 3.38 | 3.37 | 3.57 |
| GPT-4o mini | 3.58 | 3.55 | 3.78 | 3.88 | 3.26 | 3.61 |
| DeepSeek-R1-Distill-Llama-70B | 3.87 | 3.58 | 3.42 | 2.50 | 3.19 | 3.31 |
| Owen2.5-72B-Instruct | 3.30 | 3.68 | 3.63 | 3.50 | 3.22 | 3.46 |
| Llama-3.3-70B-Instruct | 3.92 | 3.40 | 3.78 | 4.00 | 3.28 | 3.68 |
| Llama-3.1-70B-Instruct | 3.65 | 3.51 | 3.62 | 4.00 | 3.23 | 3.60 |
| Llama-3.1-8B-Instruct | 3.53 | 3.29 | 3.70 | 4.00 | 3.20 | 3.54 |
| Qwen2.5-7B-Instruct | 3.23 | 3.49 | 3.31 | 3.50 | 3.16 | 3.34 |

Experiment results

RQ1: Do LLMs naturally reflect diverse persona traits in their responses?

- To ensure clinical realism, human medical experts engaged in extensive dialogue with PatientSim and evaluated its performance.
 - Clinicians consistently assigned high scores, with an average of 3.89 out of 4.
 - There was strong agreement between clinician and LLM-as-a-Judge evaluations, with Gwet's AC2 > 0.8 across all five criteria.

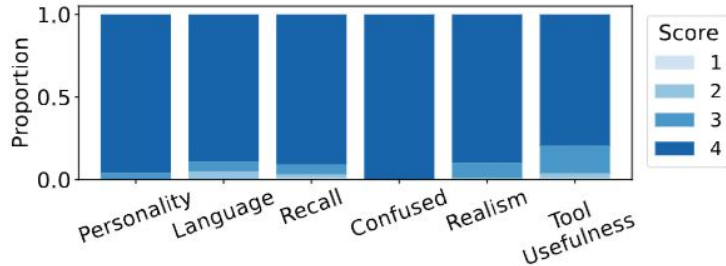


Table F11: Gwet's AC₁ and AC₂ agreement between clinician and Gemini-2.5-Flash evaluation with 95% confidence intervals estimated via 1,000 bootstrap iterations.

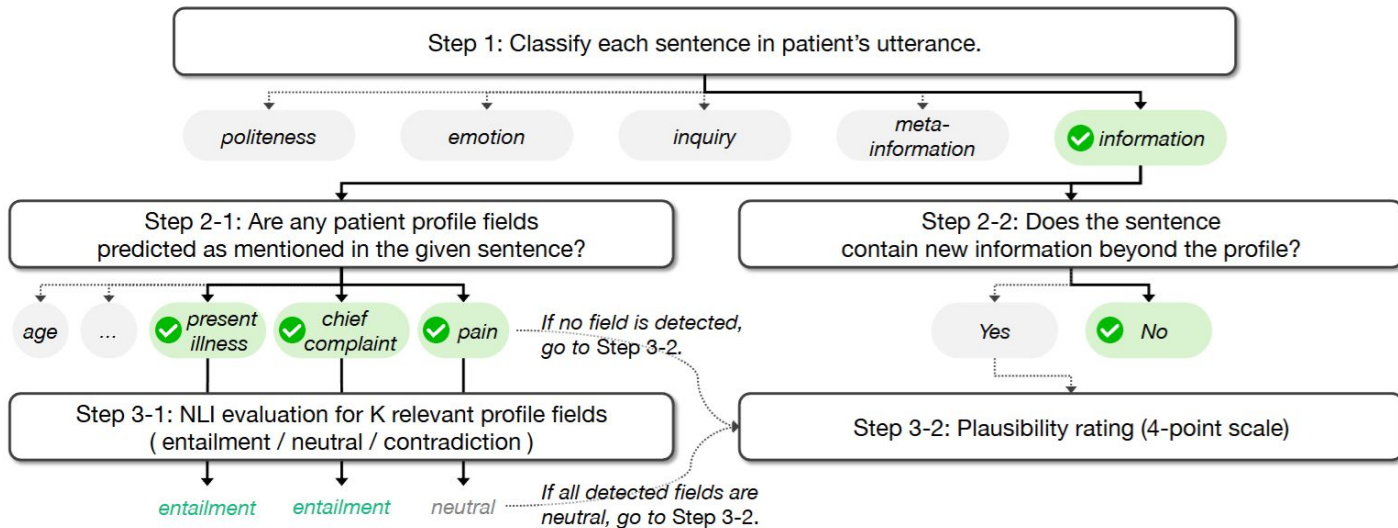
| Metric | Gwet AC ₁ (95% CI) | Gwet AC ₂ (95% CI) |
|------------------------------|-------------------------------|-------------------------------|
| Personality | 0.897 (0.830, 0.949) | 0.957 (0.919, 0.987) |
| Language proficiency | 0.347 (0.218, 0.471) | 0.818 (0.745, 0.876) |
| Medical history recall level | 0.693 (0.585, 0.786) | 0.916 (0.865, 0.957) |
| Cognitive confusion level | 1.000 (1.000, 1.000) | 1.000 (1.000, 1.000) |
| Realism | 0.321 (0.211, 0.437) | 0.884 (0.861, 0.906) |

Figure 3: Score distribution across six evaluation criteria, in clinician evaluation (4-point scale).

Experiment results

RQ2: Do LLMs accurately derive responses based on the given profile?

- To evaluate the factual accuracy of the simulator, we first classify each sentence generated by the patient simulator as either supported by the patient profile or not (unsupported).
- For supported sentences, we apply NLI evaluation to assess faithfulness.



Experiment results

RQ2: Do LLMs accurately derive responses based on the given profile?

- For supported utterances, all models demonstrate high entailment. However, a notable gap exists between larger models ($\geq 70\text{B}$ parameters) and smaller models ($\leq 8\text{B}$ parameters).
- Llama-3.3-70B and Gemini-2.5-Flash show the best performance, consistently generating entailed utterances while avoiding contradictions.

Table 2: Sentence-level factuality evaluation across eight LLMs, by Gemini-2.5-Flash. Supported statements refer to sentences that relate to at least one item in the given profile. Unsupported statements include at least one piece of information that is not explicitly mentioned in the profile. **Entail** and **Contradict** are evaluated for *supported*, while **Plausibility** is assessed for *unsupported*.

| | Info (%) | Supported (%) | Unsupported (%) | For <i>Supported</i> | | For <i>Unsupported</i> |
|-------------------------------|----------|---------------|-----------------|--------------------------|--------------------------------|-----------------------------|
| | | | | Entail (% , \uparrow) | Contradict (% , \downarrow) | Plausibility (\uparrow) |
| Gemini-2.5-Flash | 0.972 | 0.763 | 0.316 | 0.978 | 0.022 | 3.953 |
| GPT-4o mini | 0.957 | 0.721 | 0.428 | 0.968 | 0.032 | 3.929 |
| DeepSeek-R1-Distill-Llama-70B | 0.975 | 0.762 | 0.416 | 0.968 | 0.032 | 3.911 |
| Qwen2.5-72B-Instruct | 0.975 | 0.683 | 0.468 | 0.954 | 0.046 | 3.928 |
| Llama-3.3-70B-Instruct | 0.958 | 0.796 | 0.387 | 0.981 | 0.019 | 3.963 |
| Llama-3.1-70B-Instruct | 0.948 | 0.813 | 0.407 | 0.968 | 0.032 | <u>3.955</u> |
| Llama-3.1-8B-Instruct | 0.944 | 0.771 | 0.488 | 0.944 | 0.056 | 3.897 |
| Qwen2.5-7B-Instruct | 0.987 | 0.703 | 0.453 | 0.939 | 0.061 | 3.862 |

Experiment results

RQ3: Can LLMs reasonably fill in the blanks?

- For unsupported utterances (any details in the dialogue that are not explicitly stated in the patient profile), we assess their plausibility.
 - While we provide a detailed virtual patient profile, not all possible information can be explicitly included.
 - For example, the profile might say that the patient has pneumonia, but nothing about fever. However, when asked how they feel, the virtual patient might report having a fever.
 - Simply responding with “I don’t know” to queries about information not covered in the profile would not be clinically realistic.
- Therefore, we allow unsupported sentences and evaluate their clinical plausibility in the context of the patient’s profile to ensure overall clinical validity.

Experiment results

RQ3: Can LLMs reasonably fill in the blanks?

- Larger models consistently demonstrate higher plausibility scores than smaller models.
- The Llama series again shows the best performance in this task, underscoring its strong potential to simulate realistic patient responses.

Table 2: Sentence-level factuality evaluation across eight LLMs, by Gemini-2.5-Flash. Supported statements refer to sentences that relate to at least one item in the given profile. Unsupported statements include at least one piece of information that is not explicitly mentioned in the profile. **Entail** and **Contradict** are evaluated for *supported*, while **Plausibility** is assessed for *unsupported*.

| | Info (%) | Supported (%) | Unsupported (%) | For <i>Supported</i> | | For <i>Unsupported</i> |
|-------------------------------|----------|---------------|-----------------|--------------------------|--------------------------------|-----------------------------|
| | | | | Entail (% , \uparrow) | Contradict (% , \downarrow) | Plausibility (\uparrow) |
| Gemini-2.5-Flash | 0.972 | 0.763 | 0.316 | <u>0.978</u> | <u>0.022</u> | 3.953 |
| GPT-4o mini | 0.957 | 0.721 | 0.428 | 0.968 | 0.032 | 3.929 |
| DeepSeek-R1-Distill-Llama-70B | 0.975 | 0.762 | 0.416 | 0.968 | 0.032 | 3.911 |
| Qwen2.5-72B-Instruct | 0.975 | 0.683 | 0.468 | 0.954 | 0.046 | 3.928 |
| Llama-3.3-70B-Instruct | 0.958 | 0.796 | 0.387 | 0.981 | 0.019 | 3.963 |
| Llama-3.1-70B-Instruct | 0.948 | 0.813 | 0.407 | 0.968 | 0.032 | <u>3.955</u> |
| Llama-3.1-8B-Instruct | 0.944 | 0.771 | 0.488 | 0.944 | 0.056 | 3.897 |
| Qwen2.5-7B-Instruct | 0.987 | 0.703 | 0.453 | 0.939 | 0.061 | 3.862 |

Experiment results

RQ3: Can LLMs reasonably fill in the blanks?

- Four different clinicians evaluated each unsupported sentence (616 sentences per clinician) and assigned an average plausibility score of 3.91.
- The high plausibility score, along with strong inter-rater agreement measured by Gwet’s AC_1 , demonstrates that our simulator generates clinically meaningful responses.

Table 4: Plausibility scores for unsupported sentences in patient responses, labeled by four clinicians, with three annotators per sentence (out of 4). Intra-clinician agreement measured by Gwet’s AC_1 with 95% confidence intervals estimated via 1,000 bootstrap iterations.

| | Clinician A | Clinician B | Clinician C | Clinician D |
|-------------------------------------|----------------------|----------------------|----------------------|----------------------|
| Intra-Clinician Agreement | | | | |
| Clinician A | – | 0.949 (0.927, 0.969) | 0.968 (0.951, 0.983) | 0.866 (0.828, 0.901) |
| Clinician B | 0.949 (0.927, 0.969) | – | 0.961 (0.940, 0.979) | 0.853 (0.818, 0.886) |
| Clinician C | 0.968 (0.951, 0.983) | 0.961 (0.940, 0.979) | – | 0.879 (0.843, 0.913) |
| Clinician D | 0.866 (0.828, 0.901) | 0.853 (0.818, 0.886) | 0.879 (0.843, 0.913) | – |
| Plausibility (4 point scale) | | | | |
| Plausibility | 3.955 | 3.923 | 3.985 | 3.781 |

Conclusion

- Introduces a novel framework for simulating realistic doctor-patient interactions using real-world clinical data and diverse persona profiles.
- Conduct a comprehensive evaluation, assessing factual accuracy and persona reflection.
- Built on an open-source model that offers an accessible and reproducible tool for generating doctor-patient consultation data while prioritizing patient privacy.

Q & A



Project Page



Demo