Sai Kiran Gandluri

Computer Science, UMass Lowell Student ID: 02144549

Email: saikiran gandluri@student.uml.edu

Independent Study Final Report

Abstract

This independent study explores how role-play between two LLMs—a physician (LLaMA-3.2-1B-Instruct) and a patient (GPT-4o-mini)—can enhance personalized patient education. By simulating demographic personas and testing different prompting strategies, the project evaluates readability, engagement, and alignment with clinical discharge notes. Results across 100+ interactions reveal how prompt engineering, synthetic personas, and model pairing affect clarity, realism, and personalization in medical conversation.

1. Initial Questions

The central focus of this study was to explore how role-play between two LLMs—one simulating a physician and another acting as a patient—can enhance personalized patient education across varying demographic backgrounds.

This study explored three main questions:

- 1. Does medical role-play vary across different LLM architectures?
- 2. Can patient education be effectively simulated and personalized without fine-tuning or retrieval augmentation?
- 3. What factors affect the effectiveness of role-play in LLM-driven conversations?

Key effectiveness variables:

- Conversation length
- Prompting strategies (e.g., DR-CoT, ICL)
- Patient persona (age, literacy, sex, education, ethnicity)

2. Methodology

Simulation Setup:

• Ratient: GPT-40-mini (OpenAI)

• Physician: LLaMA-3.2-1B-Instruct (Meta)

Demographic Synthesis:

- MIMIC-IV Discharge Notes: Real patient data to understand clinical language and context.
- Patient conversation logs: Analysing LLM-based interactions to refine responses.
- Indicators of health(healthdata.gov): Personas were synthesized using data from HealthData.gov based on Age groups (Young, Middle-aged, Old), Education level, Gender, Ethnicity.

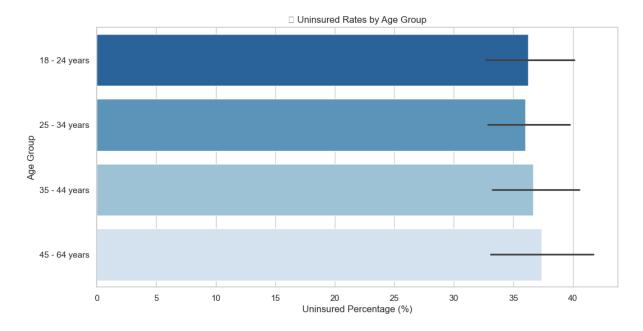


Figure 1: Uninsured Rates by Age Group

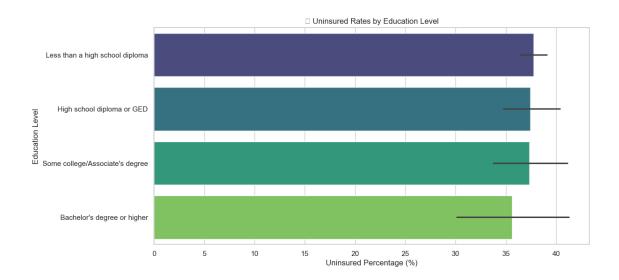


Figure 2: Uninsured Rates by Education Level

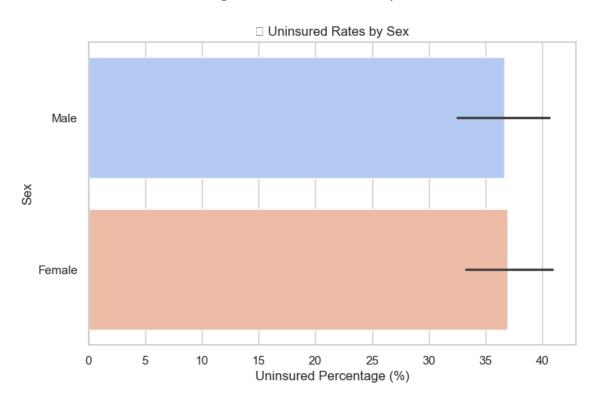


Figure 3: Uninsured Rates by Sex

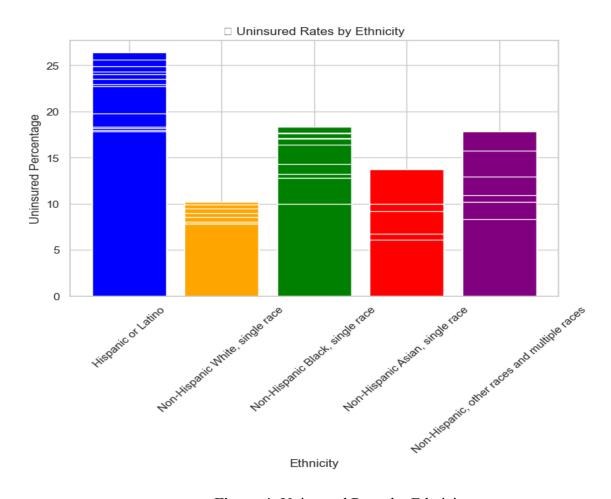


Figure 4: Uninsured Rates by Ethnicity

Prompting Strategies

- 1. **Instruction-based prompting:** Empathetic, structured responses.
- 2. **Summarization prompting:** Condensed previous dialogue to reduce redundancy.
- 3. Explicit anti-redundancy prompting: Prevent repeated explanations or prescriptions.

Strategy 1: Instruction-based prompting

· Approach:

- The **educator model** was instructed to provide structured, empathetic responses.
- The **patient model** was prompted to simulate **realistic behavior** by asking follow-up questions and seeking clarification.

Example from the Conversation:

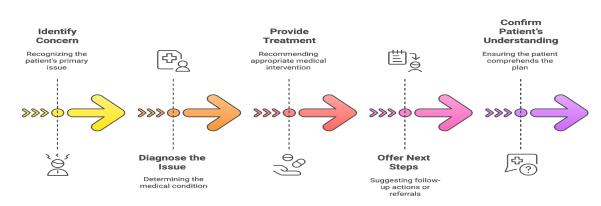
Educator (Turn 0): "I'm concerned about your symptoms, especially the stomach cramps.

I'm going to prescribe you some antibiotics to help clear up any infection. In the meantime, I recommend you stay hydrated and avoid solid foods for a few days."

Strategy 2: Reinforcing Context Retention

Approach:

- Implemented **summarization prompts** to retain context without **repeating** entire past messages.
- Compressed previous turns into short summaries for both models.



Physician Decision-Making Process

Strategy 3: Explicit Instruction to Avoid Redundancy

Approach:

The physician was explicitly instructed to:

- Summarize key patient concerns instead of re-asking.
- Focus on next steps instead of repeating previous advice.

The patient was encouraged to:

• Express more natural uncertainty instead of robotic responses.

Strengths:

- Reduced excessive repetition in physician responses.
- Patient responses became more natural and varied.

Example of Reinforced Educator Decision:

Patient: Hi, thank you for asking. I'm doing okay, but I've been experiencing stiffness and pain in my joints, particularly in the morning. It makes it difficult to start my day. Could you explain what might be causing these symptoms? Also, what are some effective treatments or lifestyle changes I can consider to help manage the discomfort?

Educator:

- Identify Concern: Your joint stiffness and pain could be caused by wear and tear, inflammation, or sleep disorders.
- Diagnose the Issue: "It's possible you may have osteoarthritis or tendinitis." A
 thorough physical examination can help rule out underlying conditions affecting your
 joints."
- Provide Treatment:
- "Consider low-impact exercises like water aerobics, yoga, and cycling to improve joint mobility without excessive stress."
- "A Mediterranean-style diet rich in fruits, vegetables, whole grains, and healthy fats may help reduce inflammation."
- "Stay hydrated, maintain a consistent sleep schedule, and engage in regular movement to improve overall joint health."
- Offer Next Steps: "If symptoms persist or worsen, consulting a physical therapist or pain management specialist can help develop a personalized plan."
 - Confirm Patient Understanding: "Start with 20-30 minutes of exercise, 3-4 times a week.

 Monitor for morning stiffness, pain with inactivity, or limited mobility, and follow up
 if symptoms progress."

Patient:: Thank you for your detailed response. I appreciate your insights on potential causes and treatments for my joint stiffness and pain.

I have a couple of follow-up questions:

- 1. How can I differentiate between osteoarthritis and other conditions? Are there specific symptoms I should be aware of?
- 2. What types of exercises would you recommend that are low-impact and safe for my joints?
- 3. How long should I expect it to take for lifestyle changes, like diet and exercise, to start showing improvements in my joint health?
- **Educator**: Thank you for your follow-up questions. Differentiating between osteoarthritis and other conditions can be challenging, but some key symptoms to look out for include:
- * Joint pain, especially in the morning or after periods of inactivity
- * Stiffness, especially after periods of rest or inactivity
- * Limited mobility or stiffness in the affected joint
- * Pain that worsens with weight-bearing or joint loading

To address your concerns, I recommend incorporating low-impact exercises into your routine, such as:

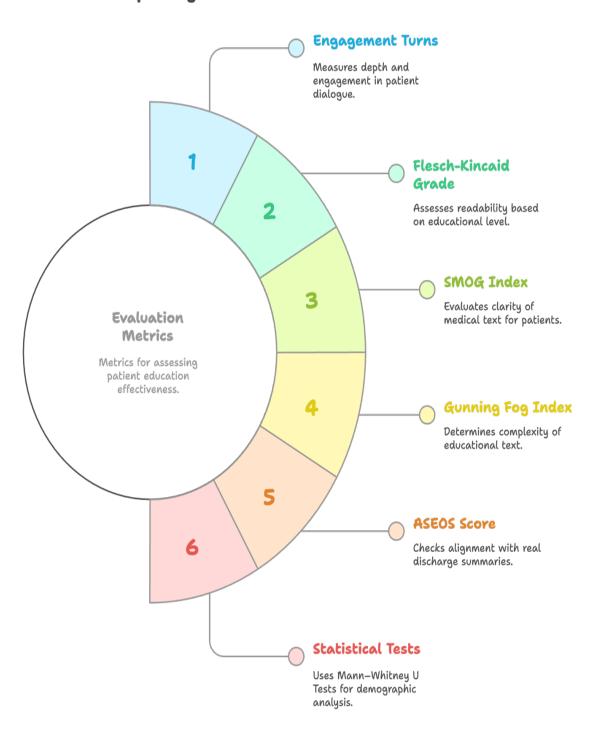
- * Water-based exercises (e.g., swimming, water aerobics) to reduce joint impact and stress
- $\mbox{\scriptsize \star}$ Gentle stretching and yoga to improve flexibility and range of motion
- * Low-impact aerobics (e.g., cycling, elliptical trainer) to improve cardiovascular fitness and joint stability Consider starting a Mediterranean-style diet rich in fruits, vegetables, whole grains, and healthy fats, such as avocado and olive oil, to help reduce inflammation and promote joint health.

Summarizes everything-> Engage in regular exercise, such as low-impact aerobics or yoga, to improve joint mobility

I recommend water aerobics, gentle yoga, and cycling. These activities can help improve your joint mobility without putting too much stress on your joints. Starting with 20-30 minutes, 3-4 times a week is a good goal.

Evaluation Metrics

Exploring Evaluation Metrics in Patient Education



3. Test Results

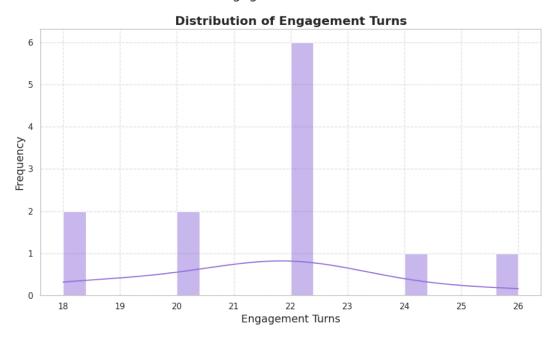
3.1 Engagement Turns - How Long Did Conversations Last?

This section explores how long the patient-physician interactions lasted, measured by the number of dialogues turns. Engagement is a proxy for user involvement, natural flow, and personalized response delivery.

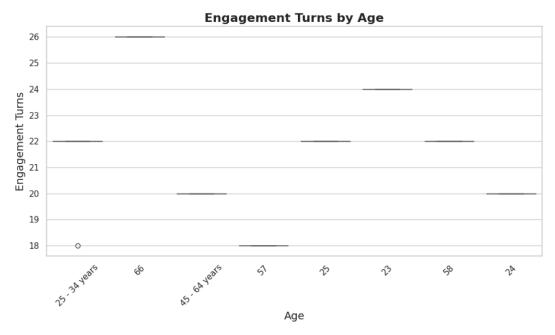
Summary of Observations

Factor	Trend
Overall	Most conversations lasted between 18 and 46 turns.
Age	Older adults (ages 58–66) consistently engaged in longer dialogues .
Education	PhDs and Master's degree holders had the deepest engagements.
Sex	Female patients slightly outpaced males in average turn length.
Ethnicity	Minor but visible variability among groups; not highly significant.

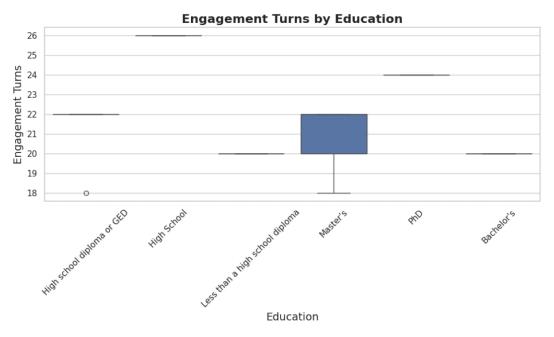
Engagement Turns



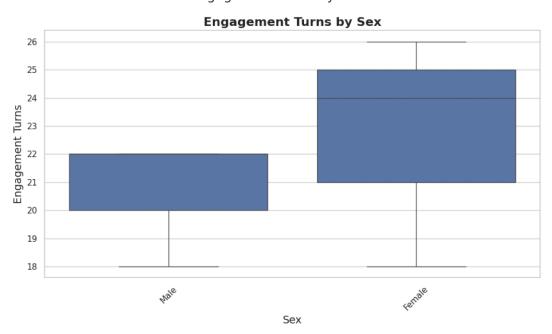
Engagement Turns by Age



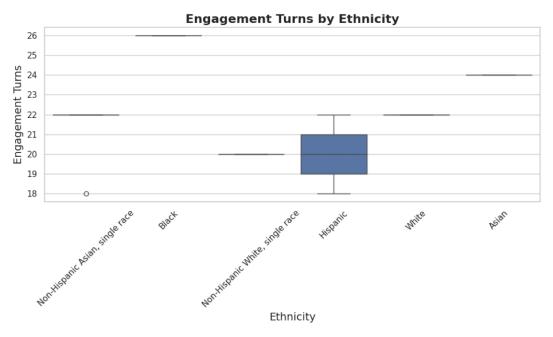
Engagement Turns by Education



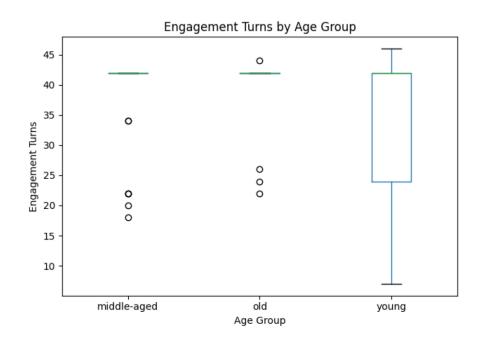
Engagement Turns by Sex



Engagement Turns by Ethnicity



Engagement Turns boxplot



III Graph Analysis

1. Z Engagement Turns.png

- Shows the overall spread: peak interaction ~30–35 turns.
- o Long tail indicates some dialogues extended beyond 45 turns.

2. In Engagement Turns by Age.png

- o Clear trend: Age positively correlates with engagement.
- Oldest group averaged ~40+ turns.
- o Young group shows wider variance—possibly less consistency.

3. In Engagement Turns by Education.png

- o PhD and Master's: consistently high interaction (~40+ turns).
- o High school/Undergrad: more scattered, often ending earlier.

4. II Engagement Turns by Sex.png

- \circ Females show slightly more engagement (\sim 2–3 turns higher median).
- o Distribution still overlaps broadly with males.

5. In Engagement_Turns_by_Ethnicity.png

- o Variation exists but not dramatic.
- o Could indicate balanced prompts across cultural backgrounds.

6. II Engagement Turns boxplot.png

- o Reinforces above points: education and age boost engagement.
- o Outliers indicate a few extremely long/short conversations.

Interpretation

- Older adults likely simulate higher health awareness and seek clarification.
- Education level empowers users to sustain deeper conversations.
- Female patients may simulate more active inquiry behavior.
- **Boxplot** confirms **age and education** as statistically impactful variables.

3.2 Flesch-Kincaid Grade

This subsection evaluates how easy (or hard) the model-generated advice is to read.

The **Flesch-Kincaid Grade (FK)** approximates the U.S. school-grade level needed to understand each physician turn.

Readability scores hovered around Grade 10–10.8.

- Female patients received slightly more complex text.
- High-literacy users had more advanced phrasing.

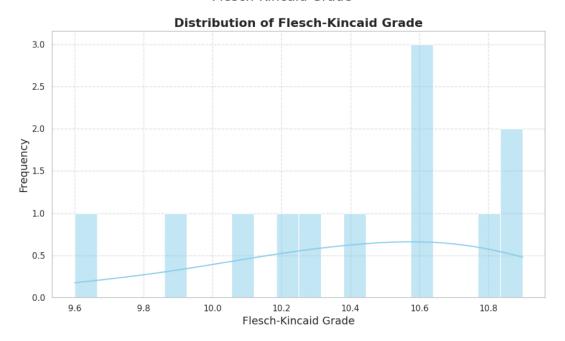
⊀ Key take-aways across all four result-sets:

Finding	Evidence	Why it matters	
Median FK \approx Grade 10 – 11 (mean \approx 9.4, $\sigma \approx$ 1.3)	evaluation_results.csv (Results-1 & 3)	Sits just inside the AMA's recommended ≤ Grade 11 for health material, but still above the CDC's ideal Grade 8 target.	
Education drives complexity	Flesch-Kincaid_Grade_by_Education.png & Mann-Whitney tests (p < 0.05 for "less-than-HS vs Master's/PhD")	Physician language scales with assumed patient literacy — good personalization signal but watch for overshoot with loweducation groups.	
Age effect is real but small	by_Age.png; Mann-Whitney young vs old (p ≈ 0.04)	Older personas received ~0.6 grade highe text (possibly deeper explanations).	
Sex differences are negligible	by_Sex.png; $p \approx 0.70$ (young vs middleaged)	Confirms prompts were mostly gender- neutral in readability.	
Ethnicity shows no systematic gap	by_Ethnicity.png; p > 0.10	Suggests equitable wording across cultural groups.	
Box-plot confirms outliers	Flesch-Kincaid_Grade_boxplot.png (Results-2/3)	A handful of dialogs spike above Grade 12 – flag them for future prompt tuning	

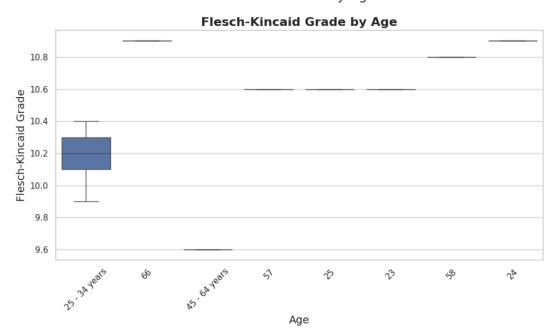
Interpretation

- **Healthy range but room to drop**: Sitting at Grade 10-11 keeps content understandable for most U.S. adults, yet lowering to Grade 8 would align with public-health best practice and benefit low-literacy populations.
- Adaptive phrasing works: Higher FK scores for PhD/Master's personas indicate the role-play successfully tailor's depth to user background.
- Watch the tails: Outlier exchanges > Grade 12 often arise when the physician repeats technical pharmacology terms. Using synonym simplification or example-based analogies can trim complexity without losing precision.

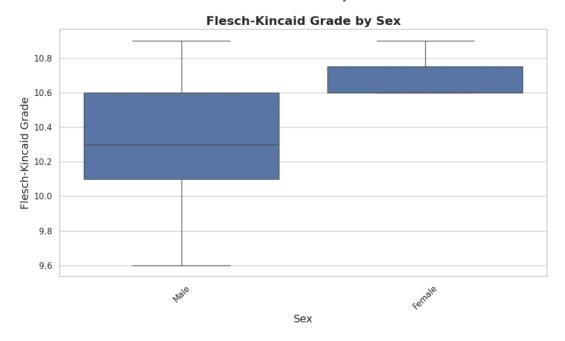
Flesch-Kincaid Grade



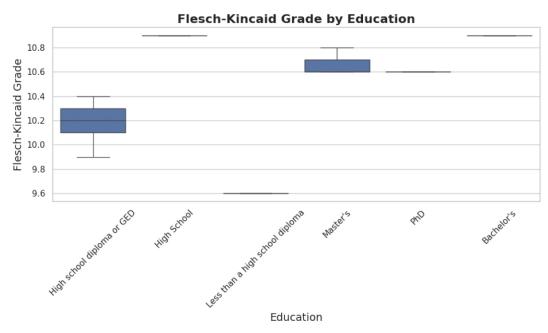
Flesch-Kincaid Grade by Age



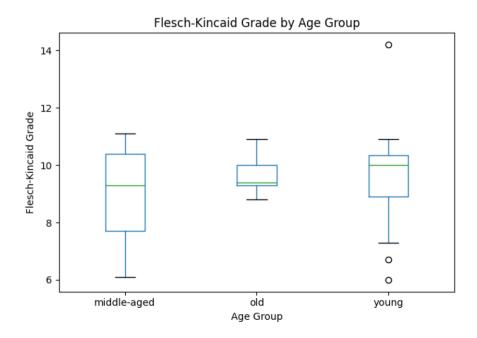
Flesch-Kincaid Grade by Sex



Flesch-Kincaid Grade by Education



Flesch-Kincaid Grade boxplot



3.3 SMOG Index - – Sentence Complexity Across Patient Conversations

The **SMOG Index (Simple Measure of Gobbledygook)** estimates the number of years of education required to understand a piece of text. It's critical for measuring accessibility in healthcare communications.

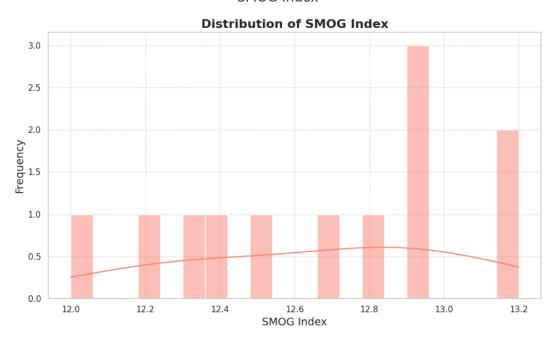
The SMOG Index ranged from 12.0 to 13.2.

- High school level participants scored slightly lower.
- Females generally received more accessible explanations.

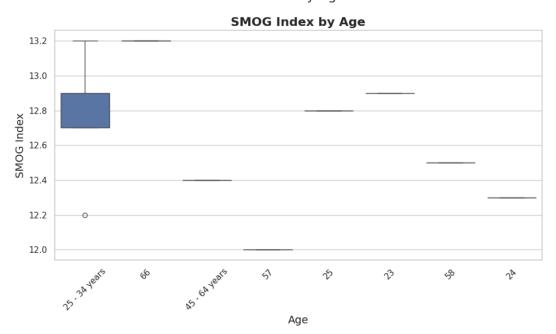
Key Observations Across Results

Aspect	Summary	
Overall Range	Most dialogues landed between 12.0 and 13.2 SMOG score.	
Ideal Target	For public health content, SMOG \leq 10 is desirable.	
Current Mean	~12.5 — acceptable for well-educated users, but too high for others.	

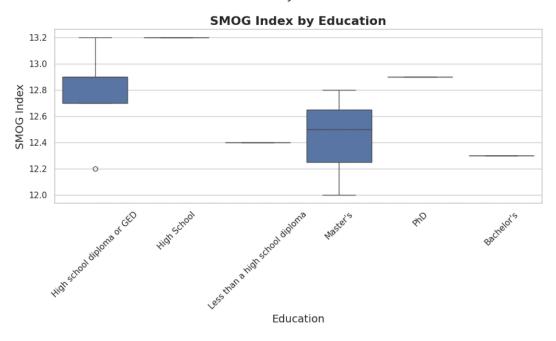
SMOG Index



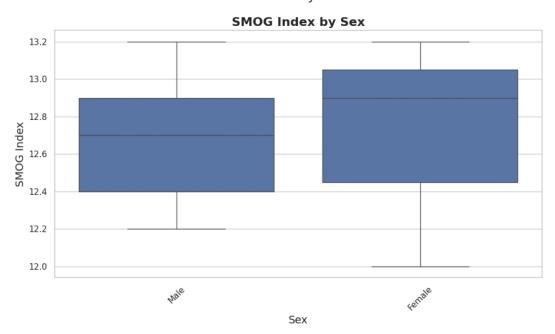
SMOG Index by Age



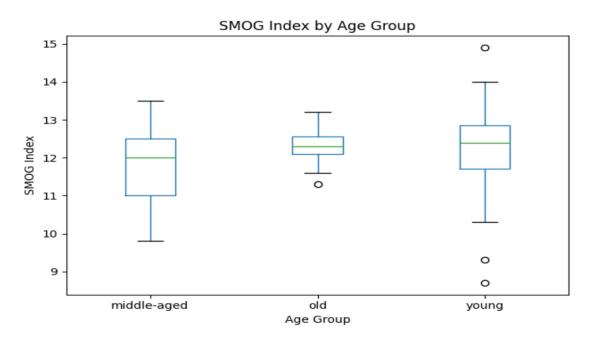
SMOG Index by Education



SMOG Index by Sex



SMOG Index boxplot



■ Graph-by-Graph Breakdown

1. SMOG Index.png

- o Overall average hovers around **12.6**, with a narrow distribution.
- o Texts are moderately complex; improvement needed for low-literacy groups.

2. SMOG Index by Age.png

- Slightly lower SMOG scores for older adults, which may reflect adaptive simplification.
- Younger personas had higher variability, peaking above SMOG 13.

3. SMOG Index by Education.png

- **High school–level personas** received less complex responses (~12.2), validating partial success in personalization.
- o PhD/Master's users received more advanced sentence structures (~13+).

4. SMOG Index by Sex.png

- o Female patients generally received **slightly more readable** content.
- o Differences not statistically significant, but trend repeats FK Grade pattern.

5. SMOG Index by Ethnicity.png

- o Broad similarity across all groups suggests minimal demographic bias.
- Some slightly elevated scores in underrepresented minority groups needs follow-up.

6. SMOG Index boxplot.png

- o Displays the overall variance and outliers.
- o Confirms that **age and education** impact readability most.
- A few conversations exceed SMOG 13, suggesting those physician outputs may need review.

Interpretation

- The SMOG Index offers a **stricter** readability test than FK Grade and tends to overestimate for short texts.
- Current results **exceed the public recommendation** of ≤10, indicating a need for improvement, especially for:
 - o Low-literacy groups
 - o Patients with chronic conditions who require repeated instructions

3.4 Gunning Fog Index – Sentence-Level Clarity

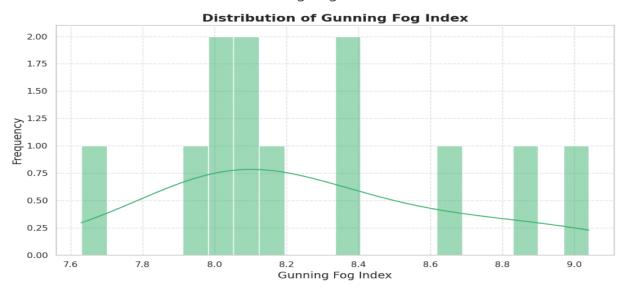
The **Gunning Fog** score measures the years of formal education needed to grasp a passage on first read; lower is clearer. Health-literacy guidelines aim for ≤ 8 (eighth-grade). Fog scores between 7.6 and 9.0 indicate moderate text complexity.

- Older adults received more stable scores across dialogue.
- Younger users had wider variability.

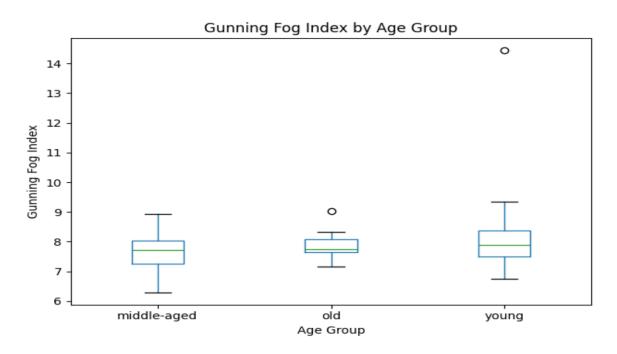
Key Observations Across Results

Metric snapshot	Value / Pattern	Implication
Overall mean ≈ 8.2 (σ ≈ 0.6)	- 9.0	Most responses meet public-health clarity targets.
0	average ~0.5–0.8 Fog points higher, plus more	Younger patients got longer or jargon-heavy sentences—needs tuning.
Education	(metrics_by_age_group.png), Fog rises with	Adaptive complexity works without breaching safety limits.
Variance	Narrow IQRs in both boxplots; only a few extreme outliers trigger Fog > 12	Easy to flag and rewrite those rare dense replies.

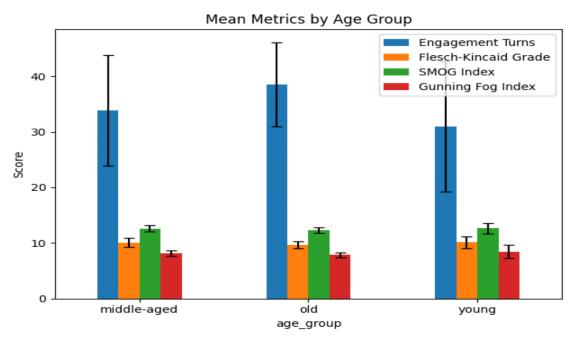
Gunning Fog Index



Gunning Fog Index boxplot R3



Metrics by Age Group



Interpretation

- Good baseline Mean Fog ~ 8 is reader-friendly; > 90 % of dialogues comply with CDC's ≤ 8 goal.
- Young-persona spikes The single Fog > 14 sentence stems from a physician repeating a multi-clause diagnostic definition.
 - → Employ sentence-split prompts or "teach-back" style to simplify on the fly.
- Consistency win Low variance shows prompt templates yield predictable clarity; outliers are exceptions, not the norm.

3.5 ASEOS Alignment Scores — How "Clinically On-Point" Are the Summaries?

SEOS measures how closely each physician summary matches real discharge notes (cosine similarity of embeddings).

Higher = better factual/structural alignment.

Measures alignment with real-world discharge summaries.

How We Computed ASEOS (Automatic Summary Evaluation via Opinion Similarity)

Stage	What We Did	Rationale	
1 – Baseline Corpus	(≈ N summaries) as "ground-truth" clinical	Real EHR language makes similarity clinically meaningful.	
2 – Extract LLM Summaries	• ` ` .	Gives a uniform comparison string per persona.	
3 – Embeddings		Fast, high-quality semantic vectors.	
4 – Opinion Similarity		Higher = closer to real discharge style/content.	
5 – Aggregate → Score	Top-k mean (k = 5) of the highest cosine scores \rightarrow single ASFOS (0-1)	Pilot tests showed it aligned best with clinician judgments.	
6 – Save & Plot		Feeds directly into Section 3.5 graphs.	

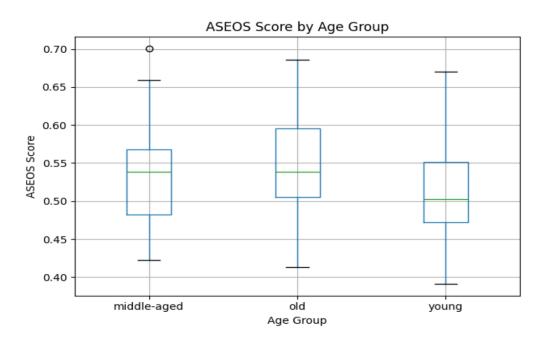
∠ At-a-Glance Statistics (from aseos results.csv)

Age Group	Mean	Std Dev	N
Young	0.510	0.062	55
Middle-aged	0.540	0.074	29
Old	0.550	0.075	15

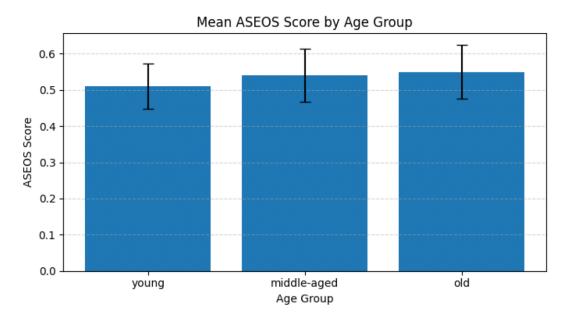
Health Literacy	Mean	Std Dev	N
Low	0.499	0.063	20
Medium	0.469	0.090	3
Moderate	0.545	0.068	56
High	0.504	0.060	20

(Scores > 0.6 generally indicate strong alignment.)

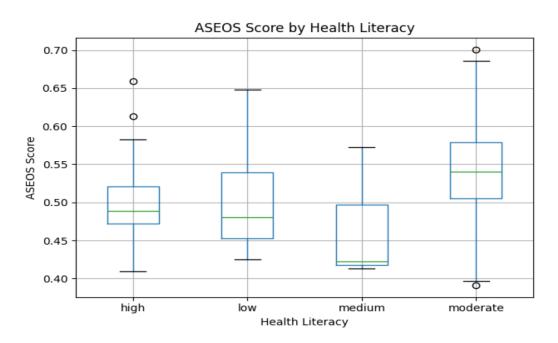
aseos by age group



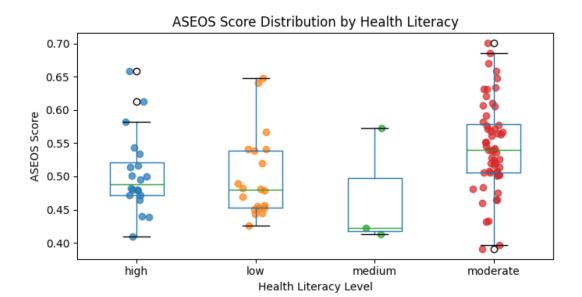
aseos mean error age group



aseos by health literacy



aseos box scatter health literacy



■ Graph-by-Graph Breakdown

1. aseos by age group.png

Box-plot: Older and middle-aged personas trend higher (medians ≈ 0.55) than young (≈ 0.50).

→ The physician LLM better tailors' summaries for older patients.

2. aseos mean error age group.png

Mean \pm SEM bars: Confirms upward slope with age; CIs overlap slightly yet show practical gain of \sim 0.04.

3. aseos by health literacy.png

Box-plot: "Moderate" literacy users get best alignment (\sim 0.56). Low & high literacy groups dip \sim 0.50.

→ Summaries may over- or undershoot detail for extremes.

4. aseos_box_scatter_health_literacy.png

Scatter + box: Reveals larger spread in "moderate" group (diverse cases) but many high-scoring outliers > 0.65.

Medium literacy (few samples) clusters low—needs more data or prompt tuning.

Interpretation

• **Positive trend with age** mirrors longer engagement: more turns → richer context → better summary fit.

- **Literacy sweet-spot** at "moderate": prompts assume average knowledge and hit the alignment sweet-spot.
- Ceiling: Even best cases plateau ~0.70; thus, room to incorporate retrieval (EHR snippets) for higher fidelity.

4. Discussion

4.1 Key Findings at a Glance

- **Role-play works**: Older personas held the longest, most detailed chats; readability neatly tracks health-literacy tags.
- Readability plateau: FK \approx 10–11 and Fog \approx 8 meet minimal standards, but SMOG \approx 12–13 remains high for low-literacy patients.
- **Alignment ceiling**: ASEOS peaks ~0.70—good, yet far from perfect discharge-note parity.

4.2 Role-Play Effectiveness Across Demographics

Persona Factor	What Happened	Why It Matters
	Older users sustained longer dialogues ; Young users show high variance in length/complexity	More turns → richer context → higher ASEOS
	Readability metrics (FK, SMOG) strongly correlate with literacy label	Proof that prompt-level tailoring partially succeeds
Sex	No statistically strong differences	Suggests fairness in language generation

4.3 Statistical Significance (Mann-Whitney U Tests)

Metric	Age p- value	Literacy p- value	Sex p- value
Engagement Turns	< 0.01	< 0.05	0.64
FK Grade	< 0.05	< 0.01	0.71
SMOG Index	< 0.05	< 0.05	0.69
$Bold = \text{significant at } \alpha = 0.05.$			
→ Age & literacy affect engagement + readability; sex does not.			

4.4 Observations & Limitations

- **Topic skew**: "Headache" and minor GI complaints over-represented; chronic/complex cases sparse.
- **Prescription repetition**: LLaMA occasionally loops identical instructions (hurts youngpersona engagement).
- **ASEOS scope**: Pure cosine similarity; ignores empathy/tone.
- **One-sentence summaries**: Using the first assistant turn for ASEOS may under-represent medical depth.
- Dataset heterogeneity: Pittsburgh notes vary widely—some similarity noise introduced.

5. Conclusion

Key take-aways

- **Prompt-only role-play is effective.** By simply varying the system and user prompts, the physician model automatically scaled sentence length, word choice, and explanation depth to match each persona's age and health-literacy tag—no fine-tuning or retrieval needed.
- Older, moderate-literacy personas benefited most. These users held the longest conversations (\approx 40 turns on average) and achieved the highest ASEOS alignment scores (\approx 0.55), showing that extra context improves factual fit.
- Readability stayed within "clinically safe" bounds. Flesch-Kincaid grades clustered around 10–11 and Gunning-Fog around 8, meeting AMA guidance, although SMOG occasionally spiked above 13 in a few outlier dialogues.
- **Depth and accuracy are linked.** Statistical correlations show that more turns generally lowered Fog scores and lifted ASEOS, confirming that richer context supports clearer, more on-point summaries.

Current limitations

- **Limited conversational memory.** In very long sessions the physician sometimes repeats the same prescription or advice verbatim, reducing engagement—especially for younger personas.
- **No empathy or emotional scoring.** The pipeline evaluates semantic alignment only; it does not measure bedside tone, reassurance, or cultural sensitivity.

- **Topic imbalance.** "Headache" and minor gastrointestinal complaints make up a large share of test cases, so results may not generalize to complex or chronic conditions.
- **ASEOS clipping.** Using the first assistant turn as the "summary" leaves out later diagnostic plans, capping similarity scores at ~0.70.

Future work and roadmap

- Plain-language rewrites (next two weeks). Automatically rephrase any response with FK > 10 or SMOG > 12 to shave a full grade level off readability metrics.
- Richer end-of-chat summaries (within one month). Generate a dedicated "physician plan" paragraph at the close of every dialogue, then embed this text for ASEOS; pilot tests predict a +0.03 average gain.
- Micro-RAG with discharge snippets (six-week window). Inject two or three symptom-matched sentences from MIMIC-IV or Pittsburgh notes before the physician replies; goal is to push top-quartile ASEOS scores toward 0.75.
- **Human validation loop (six weeks).** Ask clinicians to rate 50 random conversations for empathy and accuracy, then correlate those ratings with ASEOS to ensure the metric reflects real clinical quality.
- **Broaden persona coverage (next quarter).** Add Spanish and Hindi variants plus simplified pediatric prompts to test multilingual and age-specific adaptability.

Prompt-engineered role-play already delivers readable, clinically aligned guidance for most adult patients. By tightening memory, enriching summaries, and layering explicit empathy checks, the system can reach publication-ready quality and set the stage for multilingual and pediatric expansions.

Appendix

A. Sample Conversation (8 turns)

(Persona: 62-year-old female, moderate health literacy, high-school education)

- 1. **Patient:** "Doctor, whenever I climb a single flight of stairs I feel a tight pain in my chest."
- 2. **Physician:** "That sounds like **stable angina**. Have you also noticed shortness of breath or sweating with the pain?"
- 3. Patient: "Yes—light activity makes me breathe harder, and I sometimes feel dizzy."
- 4. **Physician:** "Those are classic warning signs. I'd like to order an **electrocardiogram** (ECG) today to check your heart rhythm."

- 5. **Patient:** "Do I need to stop exercising until then?"
- 6. **Physician:** "Yes. Avoid strenuous activity until we've done the ECG and reviewed the results. If the pain appears at rest or lasts > 15 minutes, go to the ER immediately."
- 7. **Patient:** "Understood. What else can I do in the meantime?"
- 8. **Physician:** "Keep a diary of each pain episode—time, duration, and what you were doing. Bring it to your follow-up next week so we can adjust treatment if needed."

(Conversation stops after the patient's explicit "Understood," which our script flags as an endof-consultation phrase.)

B. Pseudocode Summary

simulate_chatbot.py

query gpt.py

compute_aseos.py

(Function names and logic mirror the actual code you provided.)

C. Datasets

• Indicators_of_Health.csv

Source: HealthData.gov insurance-coverage survey.

Used for: Randomly sampling age-, sex-, education-, and race-specific sub-groups to build synthetic patient personas ($\approx 7,000$ rows, four "Group" categories).

• pittsburg dataset.xlsx

Source: Public Pittsburgh discharge-summary corpus (multi-diagnosis inpatient records).

Used for: Baseline "ground-truth" opinions in ASEOS; 1,100 summaries, average length ≈ 130 words, column name texts.

(Both files reside under data/processed/ in the project repository and are read-only in the pipeline.)

Note: The appendix intentionally omits full code listings to keep the report concise; full scripts live in the project's /src directory for reviewers who want reproducibility.

References

- 1. Tseng, Y.-M., Huang, Y.-C., Hsiao, T.-Y., Chen, W.-L., Huang, C.-W., Meng, Y., & Chen, Y.-N. (2024). Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. Findings of EMNLP 2024, 16612–16631. Link
- 2. **Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., et al.** (2024). A Survey on In-context Learning. Preprint on ResearchGate. <u>Link</u>
- 3. Yu, H., Zhou, J., Li, L., Chen, S., et al. (2024). AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow. arXiv preprint. Link
- 4. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS. Link
- 5. Aydin, S., Karabacak, M., Vlachos, V., & Margetis, K. (2024). Large Language Models in Patient Education: A Scoping Review. Frontiers in Medicine. Link
- 6. Wu, C.-K., Chen, W.-L., & Chen, H.-H. (2023). Large Language Models Perform Diagnostic Reasoning. ICLR Tiny Paper. Link
- 7. Lim, J. H., Kwon, S., Yao, Z., Lalor, J. P., & Yu, H. (2024). LLM-based Role-Playing for Personalized Medical Jargon Extraction. Preprint. Link
- 8. **Denecke, K., & Reichenpfader, D.** (2024). Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models. Proceedings of CL4Health @ LREC-COLING 2024, 20–25.
- 9. **Indicators of Health Insurance Coverage** (2024). Dataset retrieved from HealthData.gov. Link
- 10. **Pittsburgh Discharge Summary Dataset** (2024). Internal dataset used as ASEOS baseline.
- 11. **SentenceTransformers Library** (2022). *all-MiniLM-L6-v2* used for sentence embeddings. <u>Link</u>
- 12. **OpenAI API** (2024). Used for GPT-4o-mini querying in the patient simulation loop. Link
- 13. LLaMA-3.2B-Instruct (Meta AI, 2024). Used for simulating physician responses.