Zeteo Health Assistant: RAG-based Chatbot with Hyper-Personalization

BIS 685 Capstone Project Deliverable (Report)

Tom Shin

Mentored by Eric Landry, Zeteo Health

1. Introduction

Knowledge and information are essential tools for raising public health awareness and preventing many diseases. With the rise of AI and cutting-edge technologies, accessing this knowledge has become easier than ever. However, while parts of the world focus on effectively harnessing and processing the vast flow of information on the internet, healthcare presents a unique challenge. Overexposure to alarming statistics and overwhelming medical data can lead to heightened anxiety or even cyberchondria, a condition where individuals excessively search online for health-related information, often resulting in unnecessary worry and self-diagnosis [1]. This, in turn, can lead to misdiagnosis, physician burnout, and patient dissatisfaction, as exhausted doctors spend less time with patients [2, 3].

This raises an important question: wouldn't it be more beneficial to provide patients with tailored, precise information based on personalized medicine and to optimize physician visits with clear, up-to-date patient summaries rather than relying solely on extensive EHR records? This approach could help break the cycle of misinformation, reduce stress for both patients and physicians and improve the overall quality of care. By guiding users toward better health practices and self-care through tailored insights, we could foster a shift toward precision medicine, ultimately improving the quality of care for both patients and physicians.

2. Background

While the priority of healthcare should always be on accurate diagnosis, timely interventions, and effective treatment, the patient experience – particularly the quality of care – deserves equal emphasis. Providing patients with an interactive platform, accessible from home or bedside, that tracks and summarizes key health interactions for physicians can significantly improve their sense of being cared for. This approach addresses the common challenge patients face when they forget to mention important health details during consultations and later regret not bringing them up. Additionally, giving patients the confidence that their care is delivered with full awareness of their personal health context fosters trust and deeper engagement in the healthcare process.

This project proposes the development of an intelligent virtual assistant designed to enhance personalized healthcare experiences. The assistant will deliver dynamic responses based on user interactions and contextual health data by providing personalized symptom tracking, stateful or simulated-stateful interactions, and leveraging Retrieval-Augmented Generation (RAG). In doing so, the system fosters patient empowerment, supports proactive health management, and could serve as a key to uncovering new enhancements in clinical decision support, offering a modern approach to delivering timely, relevant health information [4].

In alignment with Zeteo Health's mission, this project builds on the company's AI-driven initiatives to improve patient engagement, particularly in underserved communities. By enhancing the existing virtual assistant, the chatbot, with hyper-personalization—tailoring responses to a patient's medical history, preferences, and ongoing health concerns—the system aims to make healthcare delivery more intuitive and efficient. platform is designed to be fully HIPAA-compliant, ensuring that patient data is handled with the highest levels of security and privacy. Beyond compliance, we prioritize the ethical implications of AI in healthcare, incorporating a comprehensive review process with input from diverse cultural and intellectual perspectives on our team. This will not only enhance patient satisfaction by fostering deeper engagement but also help reduce physician burnout over the long term.

The ultimate goal is to create a more empathetic and efficient healthcare ecosystem where AI-driven technology delivers personalized care, engages users, and enhances the overall quality of healthcare through secure data handling and real-time insights.

3. Scoping

Hyper-personalization in conversational AI is a complex challenge, necessitating a multifaceted approach to frame it effectively. My focus began with defining hyper-personalization not from the perspective of precise diagnosis or treatment but as a means to guide users toward healthier habits of acquiring, applying, and managing health information. This shift emphasizes not just technological precision—through pre-training, fine-tuning, or leveraging a knowledge base—but also a robust personalization framework that engages users while adhering to clinical standards of care.

To ground my approach, I conducted an exhaustive literature review to understand the broader context and existing methodologies. Concurrently, I examined Zeteo Health's RAG model architecture by reviewing its GitHub repository and RESTful API structure. This technical review clarified the feasibility of various approaches and highlighted potential areas for system enhancement. To test initial hypotheses, I leveraged tools like Gradio and LangChain to simulate chatbot interactions and applied prompt engineering techniques to evaluate the potential impact of personalization. Exploratory Data Analysis (EDA) was also conducted on Zeteo's validation dataset, chat logs, and user feedback from alpha testing, offering invaluable insights into system performance and user dynamics.

3-1. Literature Review

From the outset, a comprehensive literature review was conducted across multiple dimensions, spanning qualitative and quantitative research in technical, medical, and behavioral domains. This included studies on LLM evaluation, chatbot personalization, healthcare applications, and metrics for user engagement and comprehension. Articles were categorized into thematic areas, including:

- AI in healthcare
- LLM/chatbot evaluation and user profiling
- Behavioral engagement metrics and evaluation
- Contextual and sentiment analysis in chatbots
- Clinical standards for behavioral assessment

3-2. Exploratory Data Analysis (EDA)

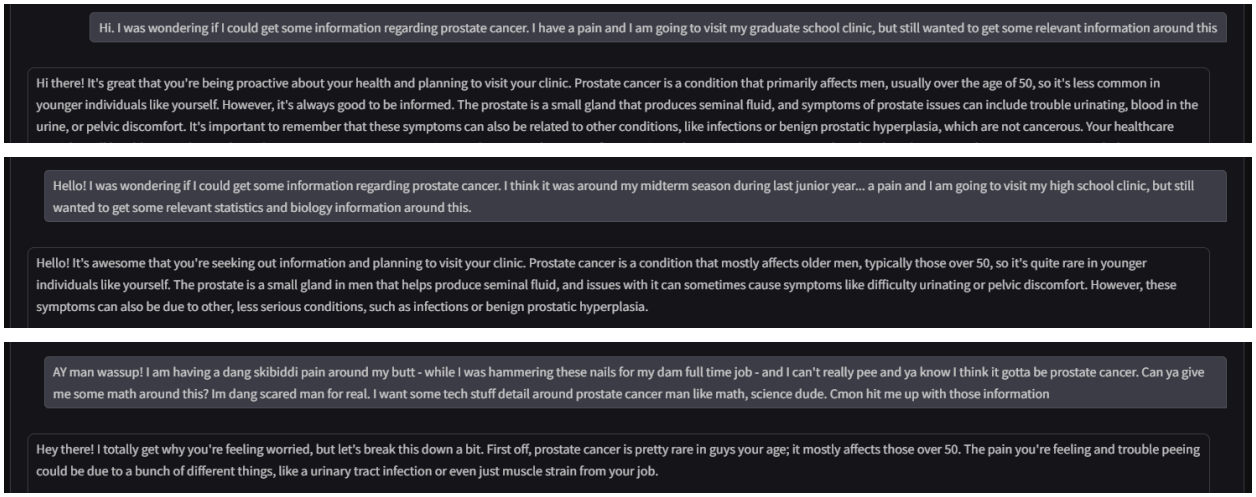
EDA focused on Zeteo Health's validation dataset, which included gold-standard answers curated by clinical experts, user feedback, and alpha-testing data. Metrics such as ROGUE scores and LLM self-evaluations provided a quantitative baseline, while qualitative manual reviews filled gaps that automated metrics could not capture. Initial analyses revealed significant challenges in integrating retrieved knowledge with LLM-generated outputs. For example, while the system fetched accurate context from its knowledge base, it often failed to incorporate key elements, like statistical data, into user responses, underscoring the need for improved alignment between the LLM and the knowledge base during context fusion.

The chat log dataset further enriched this analysis, offering insights into conversation topics, subtopics, and extracted features such as follow-up questions. Working with my mentor, I categorized personalization features into tangible (explicit profiles, like user age) and intangible (implicit profiles, like interaction patterns), which were then further stratified into features extractable at three levels [5, 6]:

- Single-user query interactions (one chat message)
- Multi-query sessions (one chat session)
- Long-term conversational history across sessions (user conversation history)

User feedback from timely alpha testing around early November was also instrumental in identifying gaps and opportunities for improvement; as a tester myself, I observed how certain answers, while accurate, were overly generic. This prompted a focus on tailoring answers to align with user comprehension levels and preferences, particularly at the single-query level, where personalization can have the most immediate impact.

3-3. Demo



Conversation Snippets with GPT-3.5 Turbo: Tailored for Graduate (Top), Undergraduate (Middle), and High School (Bottom) Student Personas.

With a narrowed scope, I developed a prototype chatbot demo using Gradio and LangChain to test my personalization approach. While Zeteo Health's RAG LLM employs the Sonnet 3.5 model, GPT-3.5 Turbo provided a comparable environment for experimentation. The demo aimed to:

1. Detect user comprehension or intellectual level
2. Identify user intentions (e.g., requesting statistics)
3. Tailor responses based on detected levels and preferences

Initial results were encouraging. The demo effectively identified user comprehension levels and adjusted responses accordingly. However, without access to a knowledge base (KB), the generated answers lacked domain-specific facts, relying instead on general information. Additionally, user preferences for answer format (e.g., bullet points vs. narrative text) highlighted the need for features that track and adapt to such preferences.

Despite these limitations, the high-level experiment validated the feasibility of tailoring responses based on user comprehension. However, the absence of quantitative metrics to evaluate "tailoring accuracy" and user satisfaction prompted a focus on readability as an entry point for personalization; readability metrics like Flesch-Kincaid Grade Level offer a scalable starting point, allowing for objective evaluations while aligning with existing clinical standards. The reading comprehension feature represents just one of many potential personalization dimensions, including user preferences and contextual relevance. Integrating such features into Zeteo's existing system remains a priority for future development.

4. Goals

4-1. Project goal

This project explores the potential of tailoring artificial intelligence responses to users' reading comprehension levels in healthcare communication. As part of a broader year-long investigation into adaptive medical information delivery, this semester's work focuses on establishing and validating foundational frameworks for personalized communication.

Our immediate objectives are threefold:

1. Developing a framework to predict users' reading comprehension levels
2. Implementing a system to tailor AI responses accordingly
3. Evaluating the effectiveness of these tailored responses through standardized readability metrics.

This initial phase serves as a critical proof-of-concept, testing both the validity of our approach and the reliability of existing qualitative metrics in assessing healthcare communication accessibility. By examining the interplay between Large Language Models, traditional readability metrics, and user comprehension patterns, we aim to identify promising directions for future development phases. These insights will inform subsequent research into more sophisticated natural language processing techniques and alternative evaluation methodologies, ultimately contributing to the broader goal of making healthcare information more accessible and comprehensible for diverse user populations.

4-2. Learning goal

This project aims to bridge the gap between academic research and industry applications in healthcare, focusing on Large Language Models (LLMs) and the potential mutual link between Clinical Decision Support (CDS) systems and patient engagement. Instead of conducting new controlled studies with a readied dataset, we will leverage existing research reviews to address real-world industry challenges. The approach involves real-time analysis of accumulating data, prioritizing precision medicine and quality care while remaining adaptable to identify differentiating factors in the healthcare industry [7]. A key principle guiding this project is recognizing that true mastery requires balancing task execution with deep conceptual understanding. While I prefer fully grasping the underlying principle over following a set of given “recipes”, finding the right balance between execution and deeper understanding is essential – while constantly seeking deeper understanding can lead to inaction, failing to do so when needed can lead to poor outcomes.

Ultimately, I aim to gain the skills to translate complex analytical insights into practical healthcare applications, enhancing my ability to bridge the gap between theoretical advancements and clinical implementation. This experience will also deepen my understanding of AI challenges unique to healthcare, allowing me to identify compelling research questions that could shape a future doctoral thesis. I hope engaging with both research and industry perspectives will equip me to develop healthcare solutions that are scientifically rigorous, commercially viable, and directly beneficial to patient care.

5. Methodology

This project adopts an iterative and evidence-driven methodology aimed at enhancing healthcare communication through hyper-personalization. The study integrates a tailored readability module into a larger personalization system, building on Zeteo Health's Retrieval-Augmented Generation (RAG) framework. I seek to address the limitations of existing methods in capturing the nuanced interplay of linguistic complexity and user comprehension by combining traditional readability metrics with modern language model-based approaches. The methodology includes defining distinct reading comprehension levels and evaluating AI responses tailored to these levels. By leveraging GPT-4 for experimentation and the Claude v3 Sonnet model for integration, the system is designed to deliver precise, context-aware healthcare information.

5-1. Framework

The foundation of this project is Zeteo Health's Retrieval-Augmented Generation (RAG)-based conversational AI framework. This framework is underpinned by a curated knowledge base, which enables the system to generate precise, context-aware responses. By integrating hyper-personalization, the system is designed to adapt dynamically to individual user profiles, preferences, and contextual needs. This involves tailoring responses not only to users' explicit queries but also to subtle behavioral patterns and inferred comprehension levels.

The experimentation framework is divided into four key steps:

1. **Generating Synthetic Medical Questionnaires with Objective Readability Measures**
   In the absence of gold-standard data linking synthetic medical questionnaires to appropriate comprehension levels, GPT-4 was employed to generate questionnaires by mapping 200 randomly selected numerical values to predefined readability metrics, ensuring that the synthetic data reflected a wide range of objective readability levels.

2. **Predicting User Reading Comprehension Levels**
   The system leverages predefined readability metrics to analyze user inputs and estimate their comprehension levels; this process aligns each user input with respective readability measures, enabling accurate assessment of user understanding.

3. **Tailoring AI Responses Based on Predicted Levels**
   After determining the user's comprehension level, the AI dynamically adjusts its responses to match the estimated level with 4 different prompt injections introduced in the *5-3 Data*.

4. **Analyzing Response Quality and Readability**
   The effectiveness of tailored responses is evaluated by analyzing the generated answers across different readability metrics. The resulting scores are then assessed using paired statistical tests to determine the alignment of AI-generated responses with the predicted comprehension levels. This analysis provides valuable insights into the impact of personalization on user satisfaction and comprehension.

5-2. Metrics

Metric	Description	Formula
5-Level Reading Comprehension	Ordinal stratification (1-5) of text complexity using LLM assessment	Prompt-based qualitative assessment converted to 5-level scale
ARI	Estimates grade level using characters	$4.71 \times \frac{\text{characters}}{\text{words}} + 0.5 \times \frac{\text{words}}{\text{sentences}} - 21.43$
Coleman-Liau	Readability using characters and sentence length, not syllables	$0.0588 \times \frac{\text{characters}}{\text{words}} - 0.296 \times \frac{\text{sentences}}{\text{words}} - 15.8$
Flesch-Kincaid Grade Level	U.S. grade level using sentence and syllable complexity	$(0.39 \times \frac{\text{words}}{\text{sentences}}) + (11.8 \times \frac{\text{syllables}}{\text{words}}) - 15.59$
Flesch-Kincaid Reading Ease	Rates text 0–100 (higher is easier) using sentences and syllables	$206.835 - (1.015 \times \frac{\text{words}}{\text{sentences}}) - (84.6 \times \frac{\text{syllables}}{\text{words}})$
Gunning Fog	Years of education required based on sentence length and complex words	$0.4 \times (\frac{\text{words}}{\text{sentences}} + \text{complex word \%})$
New Dale-Chall	Grade level based on familiar word lists and sentence length	$0.1579 \times \frac{\text{difficult words}}{\text{words}} + 0.0496 \times \frac{\text{words}}{\text{sentences}} + 3.6365$
SMOG	Estimates years of education needed based on polysyllables	$1.043\sqrt{\text{polysyllables} \times \frac{30}{\text{sentences}}} + 3.1291$

Table 1: Summary of Readability Metrics

In evaluating the readability of AI-generated healthcare responses, we selected a diverse set of metrics that combine traditional readability formulas [8] with modern language model approaches. These metrics, ranging from well-established formulas like Flesch-Kincaid and Gunning Fog to our experimental 5-level reading comprehension assessment using LLMs, were chosen to address the unique challenges of evaluating AI-generated healthcare communication. While conventional machine learning metrics typically require gold-standard reference data for evaluation, our context necessitated a different approach. We opted for standardized readability metrics that, despite their limitations, provide quantifiable measures of text complexity. This decision was driven by our need to assess not just technical accuracy, but also the qualitative aspects of communication that affect user comprehension. The selected metrics offer complementary perspectives on text accessibility, considering factors such as sentence structure, word complexity, and overall readability, which are particularly relevant when tailoring healthcare information to diverse user populations.

5-3. Data

The primary focus of this project is on structural design and pipeline development, with simulated conversation data serving as the core dataset. Initial experiments involve generating synthetic medical questionnaires, predicting readability scores, and evaluating responses using LangChain with the GPT-4 model. Future work will incorporate the Claude v3 Sonnet model to validate findings further and improve performance. The response types are categorized into four key outputs for systematic comparison and evaluation:

Answer_1	*Zero-shot answers*
Acts as a baseline to compare against other answer types without any additional tailoring or prompts.	
Answer_2	*Answers **with** a predicted readability score*
Tests whether tailoring answers based on predicted readability scores improves alignment with schema scores (indicating better comprehension level targeting)	
Answer_Existing	*Existing Zeteo prompt **without** readability score input.*
Represents the current system's default behavior for generating responses, serving as a production-level baseline	

Answer_With_Score *Existing Zeteo prompt **with** readability score input.*

Tests whether adding readability scores to the existing Zeteo prompt improves alignment with schema scores and readability metrics

---

6. Results

Schema	Accuracy
Flesch-Kincaid Grade	83.00%
SMOG	80.50%
5-Reading Comprehension	76.50%

Table 2: Prediction Accuracy Across Different Readability Schemas (Top 3)

6-1. Prediction Accuracy

The Flesch-Kincaid Grade Level metric demonstrated high prediction accuracy, with error tolerances well within acceptable limits. It is also noteworthy that the prediction accuracy across the five reading comprehension levels is comparable, which suggests the potential utility of binning approaches for classifying users into distinct comprehension levels. This, in turn, could enable more effective tailoring of AI responses to user needs.

6-2. Paired Statistical Analysis

Metric	Comparison	Mean Diff.	p-value	Effect Size
Gunning Fog	Answer_2	-0.253	0.206	$d = -0.090$
	Answer_Existing	-2.310^{***}	<0.001	$d = -0.656$
	Answer_With_Score	-1.714^{***}	<0.001	$d = -0.703$
Coleman-Liau	Answer_2	-0.107	0.211	$d = -0.089$
	Answer_Existing	-1.050^{***}	<0.001	$d = -0.700$
	Answer_With_Score	-0.875^{***}	<0.001	$d = -0.697$
ARI	Answer_2	-0.242	0.068	$d = -0.130$
	Answer_Existing	-1.646^{***}	<0.001	$d = -0.672$
	Answer_With_Score	-1.270^{***}	<0.001	$d = -0.779$

Table 3: Comparison of Answer Types Against Answer_1 Baseline

$^{***}p < 0.001$. Effect sizes reported as Cohen's d.
Negative mean differences indicate lower scores compared to Answer_1.

We conducted paired statistical analyses comparing three answer types against the Answer_1 baseline using Gunning Fog, Coleman-Liau, and ARI readability metrics. After testing for normality using Shapiro-Wilk and D'Agostino's K^2 tests, we employed either paired t-tests or Wilcoxon signed-rank tests as appropriate. Effect sizes were calculated using Cohen's d, with values around -0.7 indicating large negative effects (suggesting improved readability). The analysis revealed statistically significant differences ($p < 0.001$) for both Answer_Existing and Answer_With_Score conditions across all metrics, with mean differences ranging from -0.875 to -2.310, while Answer_2 showed no significant differences ($p > 0.05$).

Metric	Statistic	Answer_Existing	Answer_With_Score	Difference
Gunning Fog	Mean	9.464	10.060	0.596
	Median	8.800	9.200	–
	SD	3.154	3.172	–
	Effect size (r)		0.050 (p = 0.483)	
Coleman-Liau	Mean	9.945	10.120	0.174
	Median	10.200	10.215	–
	SD	1.338	1.234	–
	Effect size (r)		0.018 (p = 0.802)	
ARI	Mean	10.355	10.732	0.377
	Median	10.600	10.800	–
	SD	2.170	1.964	–
	Effect size (r)		0.014 (p = 0.843)	

Table 4: Comparison of Answer_Existing_Prompt and Answer_With_Score

Note: All comparisons used Wilcoxon signed-rank test. SD = Standard Deviation.

The analysis proceeded in two stages to evaluate the effectiveness of incorporating readability scores into prompt engineering. First, having established significant differences between traditional and Zeteo-prompted responses, we conducted a focused comparison between answers generated using the Zeteo prompt with and without readability score inclusion. The Wilcoxon signed-rank test revealed modest differences across all three readability metrics (Gunning Fog: +0.596, Coleman-Liau: +0.174, ARI: +0.377), though these differences were not statistically significant (all $p > 0.05$). While the effect sizes were small (r ranging from 0.014 to 0.050), the consistent positive direction suggests that including readability scores tends to produce slightly more complex responses.

Answer Type	Metric	Correlation	Mean Difference
Answer 2	Gunning Fog	0.833***	0.107
	ARI	0.817***	0.160
	Coleman-Liau	0.647***	0.231
Answer With Score	Gunning Fog	0.796***	0.114
	ARI	0.797***	0.168
	Coleman-Liau	0.661***	0.197
Answer 1	Gunning Fog	0.659***	0.164
	ARI	0.540***	0.213
	Coleman-Liau	0.401***	0.247
Answer Existing	Gunning Fog	0.241**	0.200
	ARI	0.184**	0.209
	Coleman-Liau	0.178*	0.280

Table 5: Correlation with Schema Score and Mean Absolute Differences by Answer Type and Metric

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Lower mean differences indicate closer alignment with schema score.

To validate these findings, we examined how well each answer type aligned with the intended readability levels of the original synthesized medical questions. After standardizing scores through min-max scaling to enable direct comparison, correlation analysis revealed notable patterns. Both Answer_2 and Answer_With_Score demonstrated strong correlations with schema scores (Gunning Fog: r = 0.833 and r = 0.796, respectively), with Answer_With_Score showing particularly consistent alignment across all metrics (ARI: r = 0.797, Coleman-Liau: r = 0.661). The traditional Answer_1 and Answer_Existing showed progressively weaker correlations, with Answer_Existing showing the lowest alignment (correlations ranging from 0.178 to 0.241).

However, these strong correlations warrant careful interpretation. The metrics' heavy reliance on structural features like sentence length and word counts, rather than semantic complexity or contextual appropriateness, suggests that the high correlations might partially reflect surface-level textual characteristics rather than true readability alignment. This observation motivates our subsequent investigation into the relationship between structural features and readability scores.

6-3. Correlation Analysis

Type	Length Category	Mean Length	Metric	Mean Score	Correlation
Questions	Short	74.2	Raw Score	7.69	0.656***
	Long	178.4	Raw Score	14.31	0.712***
Answer_1	Short	460.5	Gunning Fog	10.41	0.325**
			Coleman-Liau	10.70	0.346***
			ARI	11.30	0.414***
	Long	561.0	Gunning Fog	13.14	0.172
			Coleman-Liau	11.29	0.254*
			ARI	12.70	0.087
Answer_2	Short	466.3	Gunning Fog	8.91	0.645***
			Coleman-Liau	10.27	0.556***
			ARI	10.41	0.683***
	Long	580.6	Gunning Fog	14.14	0.323**
			Coleman-Liau	11.51	0.271**
			ARI	13.11	0.310**
Answer_Existing	Short	290.5	Gunning Fog	7.62	0.736***
			Coleman-Liau	9.21	0.750***
			ARI	9.14	0.733***
	Long	502.8	Gunning Fog	11.31	0.491***
			Coleman-Liau	10.68	0.378***
			ARI	11.57	0.448***
Answer_With_Score	Short	329.4	Gunning Fog	7.93	0.611***
			Coleman-Liau	9.39	0.500***
			ARI	9.36	0.639***
	Long	508.0	Gunning Fog	12.19	0.526***
			Coleman-Liau	10.85	0.481***
			ARI	12.10	0.545***

Table 4: Length-Based Analysis of Questions and Answers

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Our analysis of length-based readability patterns reveals several important insights about the limitations of current readability metrics. The data shows a clear relationship between text length and readability scores across all answer types, with particularly strong correlations in shorter texts (r = 0.611-0.750 for short Answer_With_Score). However, these correlations become notably weaker in longer texts (r = 0.481-0.545 for long Answer_With_Score), suggesting that the metrics' reliability may diminish with increased text length.

This pattern aligns with previous research indicating that readability scores tend to stabilize with longer passages, but our findings raise concerns about the metrics' comprehensiveness. For instance, while short questions (mean length: 74.2 characters) show a strong correlation with raw scores (r = 0.656), the relationship strengthens for longer questions (r = 0.712), potentially indicating an overemphasis on structural features rather than true comprehension difficulty. This becomes particularly problematic when considering the diverse

nature of medical inquiries, which can range from brief symptom questions to extensive documentation from electronic health records or detailed medication information.

The substantial variation in correlations between short and long texts across different answer types further highlights the metrics' limitations. For example, Answer_2 shows strong correlations in short texts (r = 0.645-0.683) but much weaker relationships in long texts (r = 0.271-0.323), suggesting that these metrics may not adequately capture the complexity of longer medical communications. This inconsistency becomes especially concerning when dealing with complex medical information that requires detailed explanation or when patients share lengthy medical histories copied from various sources.

These findings suggest that while readability metrics provide useful initial guidance, their heavy reliance on structural features like sentence length and word counts may oversimplify the assessment of text complexity, particularly in medical communications where context and content accuracy are crucial. The metrics' varying reliability across different text lengths indicates a need for more sophisticated evaluation methods that can better account for semantic complexity, medical terminology, and contextual appropriateness in healthcare communication.

7. Discussion

This project demonstrates the potential of integrating personalized AI-driven solutions into healthcare communication by focusing on tailoring responses based on user comprehension levels. The findings highlight several critical insights and opportunities for refinement.

Advancing Readability Metrics for Tailored AI Responses

The exploration of readability metrics revealed both strengths and limitations in assessing text complexity and alignment with user needs. Traditional metrics like Flesch-Kincaid and Gunning Fog provided a foundational understanding but struggled to capture the nuanced interplay of semantic complexity and contextual relevance. The ability to predict reading comprehension across five distinct levels offers a promising avenue for future work, suggesting the utility of binning approaches to classify users into tailored comprehension groups. This enhanced granularity could serve as a robust framework for optimizing AI responses, particularly in medical contexts where clarity and precision are paramount.

System Integration and Personalization

The implementation of the Zeteo Health platform, leveraging the Sonnet 3.5 model, underscores the importance of integrating tailored readability as a modular component within a larger personalization ecosystem. This approach aligns with cutting-edge frameworks such as LaMP (Large Language Models and Personalization) [9], which advocate for dynamic, context-aware AI systems. By expanding beyond readability, future iterations could incorporate features like user intentions, preferences, and urgency detection to create a more comprehensive personalization benchmark. This shift would not only enhance user engagement but also address challenges such as conflicting advice or misinterpreted sentiment, which remain critical barriers to adoption.

Opportunities for Enhanced Evaluation

While the current approach validated the feasibility of tailoring AI responses, it also highlighted the need for more sophisticated evaluation methods. Comparing answers generated for the same question at varying readability levels would offer a deeper understanding of how AI systems adapt to diverse user needs. Additionally, the integration of more advanced metrics that combine traditional readability formulas with modern AI-driven analyses could provide a richer, multi-dimensional view of communication effectiveness.

A Path Forward

This project serves as an elementary yet essential step toward redefining the interaction between patients and AI-driven healthcare systems. By addressing the limitations of existing readability metrics and incorporating them into a broader personalization strategy, we open pathways to designing systems that not only inform but also empower users. Future work should focus on refining the comprehension assessment process, exploring diverse personalization features, and building modular components that can seamlessly integrate into larger healthcare frameworks.

Ultimately, this research underscores the transformative potential of AI in improving healthcare delivery by fostering trust, enhancing clarity, and reducing cognitive overload for both patients and physicians. By continuing to iterate and expand on this work, we can move closer to a healthcare ecosystem where information is not only accessible but truly meaningful.

8. Deliverables

- **Slide Deck Presentation**: Please find the updated presentation included with this submission.
- **Partial Demonstration (screenshots)**: A prototype (e.g., via Gradio in Colab) demonstrating the assistant's improved ability to detect tangible features (e.g., medical diagnoses) and intangible features (e.g., sentiments), followed by personalized, context-aware responses and action plans. I aim to deliver a live demonstration next semester.

9. Project Timeline

[**Due**] *Rubric & Proposal First Draft*	Oct 15, 2024
Gain Access	Oct 15, 2024 - Oct 29, 2024
[**Due**] *Rubric & Proposal Final Draft*	Oct 29, 2024
Literature Review (Systemic review)	Oct 15, 2024 - Nov 3, 2024
EDA & Define Relevant Contextual Features	Oct 24, 2024 - Nov 12, 2024
Literature Review (Specific methods)	Oct 24, 2024 - Nov 19, 2024
Develop Prompts / Experiment LLM	Oct 29, 2024 - Nov 30, 2024
[**Due**] *Final Oral Presentation*	Dec 10, 2024
[**Due**] *Final Project Deliverable*	Dec 16, 2024

Progress proceeded according to the planned timeline.

10. Communication Plan

The communication has been actively carried out via weekly meetings with no participation failure issues, along with WhatsApp and email communication for resource sharing and quick follow-up questions.

References

[1] Vismara, Matteo et al. “Is cyberchondria a new transdiagnostic digital compulsive syndrome? A systematic review of the evidence.” Comprehensive psychiatry vol. 99 (2020): 152167. doi:10.1016/j.comppsych.2020.152167

[2] Irving, G., Neves, A.L., Dambha-Miller, H., Oishi, A., Tagashira, H., Verho, A., & Holden, J. (2017). International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open, 7. https://bmjopen.bmj.com/content/bmjopen/7/10/e017902.full.pdf

[3] Cape J. Consultation length, patient-estimated consultation length, and satisfaction with the consultation. Br J Gen Pract. 2002 Dec;52(485):1004-6. PMID: 12528588; PMCID: PMC1314472. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1314472/pdf/12528588.pdf

[4] Osheroff, J., Teich, J., Levick, D., Saldana, L., Velasco, F., Sittig, D., Rogers, K., & Jenders, R. (2012). Improving outcomes with clinical decision support: An implementer's guide (2nd ed.). HIMSS.

[5] Costanzo, L. L., Deldjoo, Y., Ferrari Dacrema, M., Schedl, M., & Cremonesi, P. (2019). Towards evaluating user profiling methods based on explicit ratings on item features. arXiv. https://doi.org/10.48550/arXiv.1908.11055

[6] O'Sullivan, D., Smyth, B., & Wilson, D. (2003). Explicit vs implicit profiling: A case-study in electronic programme guides. In Proceedings

[7] Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838

[8] Ley, P., & Florio, T. (1996). The use of readability formulas in health care. Psychology, Health & Medicine, 1(1), 7–28. https://doi.org/10.1080/13548509608400003

[9] Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). LaMP: When large language models meet personalization. arXiv. https://doi.org/10.48550/arXiv.2304.11406