Enhancing Automated Health Literacy Feedback: A Merged-Concept Fine-Tuning Approach with Large Language Models

Donger Chen[1], Hang Tran[1], Sameep Shah[2], Ying He[1], Yunhe Feng[1], Carol Howe[3], Lindsey Patton[4], Liran Ma[5], Song Fu[1]

[1] Department of Computer Science and Engineering, University of North Texas, Denton, Texas, USA
[2] Department of Computer Science, Texas Christian University, Fort Worth, Texas, USA
[3] Harris College of Nursing & Health Sciences, Texas Christian University, Fort Worth, Texas, USA
[4] Children's Health, Dallas, Texas, USA
[5] Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA

***Abstract*—Effective clinical communication, underpinned by strong health literacy (HL) principles, is crucial for optimal patient outcomes. Providing scalable and actionable feedback to clinicians on HL practices such as Teach-Back, Plain Language, and Open-Ended Questioning remains a challenge. This paper extends our prior work on the HealthLit system, which utilized Retrieval Augmented Generation (RAG) with Large Language Models (LLMs) for HL feedback. We introduce a novel "Merged Finetuning" approach, where LLMs (Llama 8B and Mistral 7B) are fine-tuned on a dataset combining annotations across all three HL practices in a unified format to facilitate holistic feedback generation. This contrasts with baseline approaches that often handle concepts in an isolated manner. Our evaluation on a test set of 9 clinical dialogue scripts demonstrates that the "Merged Finetuning" Llama 8B model significantly improves agreement with human expert ratings, achieving 62.96% agreement compared to 25.93% in previous work. The "Merged Finetuning" Mistral 7B model also showed improvement, though more modest with 37.04% agreement vs. 29.63% for baseline. Qualitatively, the merged fine-tuning approach yields more coherent, efficiently generated, and, particularly for Llama 8B, more actionable feedback with fewer inconsistencies relative to the baseline RAG models. These findings highlight the potential of data-centric fine-tuning strategies to create more reliable and effective AI tools for enhancing health communication.**

***Index Terms*—AI in Healthcare, Language Models, Natural Language Processing, Human-Computer Interaction, Intelligent Multimedia Systems.**

I. Introduction

The rise of intelligent multimedia systems has revolutionized domains such as healthcare, where textual, auditory, and visual data drive decision-making. Effective communication between clinicians and patients is a high-dimensional challenge, shaped by linguistic nuances (e.g., semantics, context), non-verbal cues (e.g., tone, gestures), and patient-specific factors. In pediatric care, where interactions frequently involve caregivers and sensitive health information, these challenges become even more intricate due to strict privacy regulations and the necessity for lightweight, localized systems to ensure data protection. To tackle these challenges, we analyze transcribed nurse-patient audio interactions in the clinical setting, removing non-verbal elements while preserving textual meaning. However, even in text form, these conversations remain unstructured and lengthy, making scalable evaluation difficult. Our prior work, the HealthLit system [1], pioneered an approach to address this by employing pre-trained Large Language Models (LLMs) such as Mistral 7B and Llama 8B models, combined with Retrieval Augmented Generation (RAG), to automate the auditing of these crucial HL practices in transcribed pediatric clinical dialogues. HealthLit demonstrated the feasibility of a scalable, privacy-aware system for delivering valuable communication feedback.

Building upon insights from previous work, we recognized that while analyzing individual health literacy (HL) concepts provides useful feedback, clinical communication is inherently holistic, with multiple principles often interacting within a single conversational exchange. An approach that accesses HL concepts in isolation, where each concept is treated as a distinct task, might not fully capture these interdependencies or provide the most comprehensive and integrated feedback. For example, with such a strategy, analyzing a clinician's utterance like "This new pill, the metoprolol, is for your blood pressure, okay?" might involve separate, sequential queries to the LLM: first, to evaluate its use of plain language (e.g., assessing "metoprolol"); second, to determine if the question "okay?" constitutes an effective open-ended question or teach-back attempt. Each step would yield a distinct piece of feedback.

We hypothesized that fine-tuning an LLM on a dataset specifically structured to present conversational snippets alongside a comprehensive evaluation of all relevant HL concepts simultaneously could enable a model to achieve a more nuanced understanding and the ability to generate more cohesive, actionable feedback. This led to the development of a "merged dataset design" (detailed in Section III-B1), where each training instance pairs a conversational input with a target output that encompasses assessments for Teach-Back, Plain Language, and Open-Ended Questions. In contrast to the previous method, our new Merged Finetuning approach trains a single model to process the same utterance once and generate a unified report. This report would simultaneously assess the use of medical terminology ("metoprolol"), the closed nature of the question ("okay?"), and the absence or presence of a teach-back, all within a single integrated analysis. This holistic analytical capability is a central advancement of the work presented in this paper.

This paper builds upon our initial system by introducing the "Merged Finetuning" strategy, an advancement developed by fine-tuning LLMs on a novel dataset that integrates multiple HL concepts into a unified structure. The key contributions of this extended work are:

- We create a specialized fine-tuning dataset designed for holistic HL assessment. This includes data aggregation and cleaning procedures, the unique structure consolidating annotations of three HL practices for each conversational input, and the comprehensive instructional prompt engineered to guide the fine-tuning of a single model for multifaceted HL evaluation.
- We provide an empirical evaluation of the Merged Finetuning model, comparing its HL assessments against the original HealthLit baseline system and human health literacy expert ratings. The results demonstrate improved model accuracy and higher alignment with expert judgment.
- The model's feedback is designed for seamless integration into nurse training programs and continuous professional development. This Human-Computer Interaction (HCI)-focused output aims to provide tangible guidance that helps healthcare professionals refine their communication strategies, ultimately fostering improved patient understanding and engagement.

II. Background and Related Work

A. LLMs in Healthcare

Modern healthcare systems increasingly rely on AI and multimedia technologies to process diverse data modalities, including text, audio, and video, to improve clinical decision-making. Recent advances in large language models (LLMs) have demonstrated their versatility as "health system-scale prediction engines" capable of tasks ranging from diagnosis coding to patient outcome forecasting [2]. Yang et al. [3] outlines the development of LLMs and their applications in health care, including clinical decision support, medical documentation, and patient engagement, while also highlighting challenges such as data privacy, reliability, and ethical considerations. Similarly, the survey by Luo et al. [4] details the extensive use of pre-trained language models in medicine, covering various architectures, datasets for pre-training and fine-tuning, and current research trends. However, fine-tuning with sensitive medical data introduces significant safety and ethical considerations. Critical issues such as the risk of models learning and perpetuating biases present in training data, generating medically incorrect information, and concerns related to patient privacy and data security still need careful consideration [5]. These challenges underscore the need for careful dataset curation, robust evaluation protocols, and ongoing safety monitoring when fine-tuning LLMs for healthcare applications.

These models excel at parsing unstructured clinical narratives to extract actionable insights. For instance, LLMs with RAG have proven effective in summarizing lengthy electronic health records (EHRs) while preserving critical clinical context [6]. Such systems highlight the potential of LLMs to transform healthcare workflows by automating labor-intensive tasks like documentation and audit processes. Another study [7] proved that LLMs such as ChatGPT can achieve similar or better outcomes compared to human experts in clinical dialogue summarization tasks. However, they also found that sometimes, especially for non-English dialogues, the output may include hallucinations or omit crucial information compared to human experts, indicating areas for further development.

B. AI for Health Literacy and A&F

While LLMs are widely applied within the medical field, their use in evaluating HL practices in clinician-patient interactions remains underexplored. Health literacy, defined as an individual's capacity to understand and act on health information [8], is fundamental to effective care. Deficiencies in health literacy are linked to adverse health outcomes, including increased hospitalization rates and reduced treatment adherence [9]. Consequently, health organizations like *Healthy People 2030* [10] advocate for organizational health literacy, encouraging healthcare systems to implement clear communication strategies such as teach-back, the use of plain language, and open-ended questioning [11], [12].

Traditionally, ensuring adherence to these communication practices has relied on Audit & Feedback (A&F) methods involving manual expert reviews, which are resource-intensive and potentially subjective [13]. AI, particularly LLMs, offers a pathway to automate and scale A&F. Early research discussed the broader potential of AI to analyze and teach communication in healthcare, laying groundwork for more advanced applications [14]. More recently, research has explored adapting LLMs for structured information extraction from clinical dialogues [15]. Bodonhelyi et al. [16] investigated LLMs for simulating challenging patient interactions, providing a scalable tool for medical communication training. These endeavors highlight the potential for AI to not only analyze but also actively improve clinical communication skills. Our work builds on these advancements by focusing on automated feedback for specific HL practices.

C. Advancements in Prompt Engineering

Prompt engineering has emerged as a critical discipline for effectively utilizing pre-trained LLMs. A prompt can be defined as the textual guidance that users provide to shape LLMs' outputs [17]. Prompts usually are combinations of explicit instructions, questions, input data, and illustrative examples [17]. The goal is to design the most effective input to achieve the desired output from a generative model. This process is complex because it involves more than giving simple direction, but requires a deep understanding of what a model can and cannot do, along with the special circumstances of its application. A significant benefit of prompt engineering is its capacity to improve model performance on specific tasks using tailored instructions, without requiring changes to the underlying model's weights [18]. This methodology is particularly relevant to our work, where the goal is to adapt general LLMs for specialized applications in HL.

To enhance the reasoning capabilities and factual accuracy of LLMs, various prompting techniques have been developed. Chain-of-Thought (CoT) prompting [19] is a prominent technique where the input illustrates a step-by-step thought process to reach a conclusion. This encourages the model to emulate human-like problem decomposition into manageable, logical stages [18]. A simple sentence like "Let's think step-by-step" can trigger this reasoning pattern. By leading the model through a more organized and traceable reasoning path, CoT can enhance the reliability of its outputs [17]. Building on this, Step-Back Prompting offers another sophisticated method, enabling models to perform abstraction by identifying overarching concepts and principles from detailed examples, typically through a two-phase process involving Abstraction and then Reasoning [20].

Another technique is to include few-shot examples within the prompts for steering model's behavior. While traditional CoT often uses static, human-curated examples, newer approaches like Active-Prompt [21] focus on strategically selecting the most beneficial questions for human annotation. This technique, drawing from uncertainty-based active learning, uses metrics to identify and prioritize the most ambiguous questions for annotation to help the model have a better performance [18], [21].

The system prompt, which provides high-level instructions, is also a key area for optimization. Previous works have shown that refining system prompts can yield performance improvements as significant as those from fine-tuning task-specific prompts. Furthermore, an effectively tuned system prompt can perform as well as prompts tailored for individual tasks, and integrating both system and task-level prompt optimizations can offer synergistic benefits [22].

Finally, research also explored the effect of assigning a role to the LLM within prompts. This technique has been applied to various tasks, including text summarization [23] and to augment reasoning [24]. Kong et al. [24] found that role-play can implicitly activate CoT-like reasoning and lead to better results. While other researchers suggest that role-playing might sometimes detract from an LLM's reasoning capacity by causing distraction [25]. Therefore, a careful application with evaluation of this method is crucial for tasks in different fields.

III. Methodology

Our methodology extends prior work on the HealthLit system by developing and evaluating a novel fine-tuning approach using a merged-concept dataset. This section details the baseline HealthLit system, the construction and rationale for our new merged dataset and fine-tuning process, and the framework for evaluating its performance.

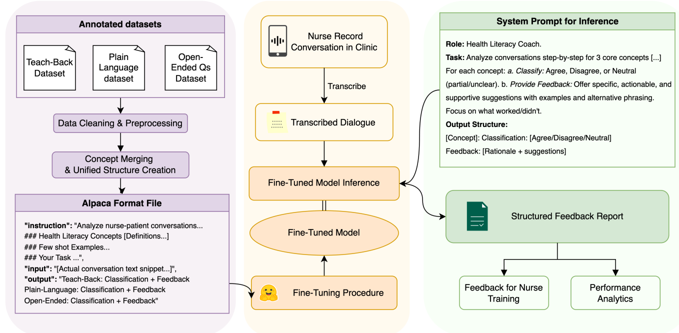

Fig. 1: The end-to-end pipeline of the Merged Finetuning model, illustrating data preparation, model fine-tuning, and the inference process for generating health literacy feedback.

A. Baseline System

The original HealthLit system served as the baseline for this work. It was designed to provide automated feedback on HL practices in clinical dialogues using Large Language Models (LLMs), specifically Mistral-7B-Instruct and Llama 3.1-8B-Instruct, augmented with an RAG framework. The primary goal was to assess key health literacy components such as Teach-Back, Plain Language, and Open-Ended Questioning. For a comprehensive description of the baseline HealthLit system architecture and its initial evaluation, please refer to our previous work [1].

The training or prompting data for the baseline system involved assessing each HL concept in a somewhat isolated manner. If we denote an input conversation snippet as x_i, and the set of health literacy (HL) concepts as $\mathcal{C} = \{c_1, c_2, \dots, c_M\}$ (e.g., c_1 = Teach-Back, c_2 = Plain Language, c_3 = Open-Ended Questions), the baseline task can be modeled as evaluating snippet x_i for a single HL concept $c_j \in \mathcal{C}$. The dataset for this baseline approach, denoted as $\mathcal{D}_{\text{split}}$, is represented as a collection of tuples: $\mathcal{D}_{\text{split}} = \{(x_i, c_j, y_{i,j})\}$, where $y_{i,j}$ is the assessment output (e.g., classification result and corresponding feedback) for snippet x_i regarding the j-th concept c_j.

This $\mathcal{D}_{\text{split}}$ structure necessitates that, to comprehensively evaluate a single conversation snippet x_i across all M HL practices, the baseline language model must execute M separate evaluation passes. In each pass, the model processes x_i using a distinct prompt specifically tailored to the particular HL concept c_j, resulting in isolated assessments of each HL practice for the same snippet. To mitigate the limitations arising from segment-level annotation sparsity and context fragmentation inherent in this approach, we constructed a merged dataset ($\mathcal{D}_{\text{merged}}$) as described in this paper.

B. The Merged-Concept Fine-tuning Approach

To foster a more holistic and nuanced assessment of clinical communication, we hypothesized that a model fine-tuned to concurrently evaluate multiple HL concepts would yield more contextually aware and practically useful feedback. This led to the development of our Merged Finetuning approach, centered around a novel dataset structure and comprehensive prompts.

The end-to-end pipeline for developing and utilizing our Merged Finetuning model is illustrated in Figure 1. This diagram outlines the key stages, commencing with the collection and preprocessing of data. Then, the merged-concept dataset is constructed, where each conversational input is structured for holistic HL assessment. Following this, the diagram shows the fine-tuning stage, where base LLMs (Llama 8B and Mistral 7B) are trained using this merged dataset along with our specifically engineered instruction and system prompts. Finally, there is the inference phase where the fine-tuned model takes a new transcribed clinical dialogue and a system prompt as input to generate a single, structured feedback report covering all three HL concepts.

1) Dataset Construction and Preprocessing: The foundation of our Merged Finetuning model is a specialized dataset constructed from three original human-annotated spreadsheets, each focusing on a specific HL concept: Teach-Back (TB), Plain Language (PL), and Open-Ended Questions (OE).

We began with three independently annotated datasets. Each dataset contained: (1) the transcribed nurse-caregiver dialogue, (2) an HL practice label (*Agree*, *Neutral*, or *Disagree*), and (3) written feedback from human experts. Preprocessing involved several steps:

a) Data Cleaning: Missing 'Feedback' values were imputed (e.g., with a default statement like "Your health literacy practice is good" for agree cases). labels (Agree, Disagree, Neutral) were standardized to a capitalized format. HTML tags and extraneous whitespace were removed from the input text, output classifications, and feedback narratives.

b) Merged Data Structure:: Unlike the previous split approach ($\mathcal{D}_{\text{split}}$), the newly constructed merged dataset, $\mathcal{D}_{\text{merged}}$, associates each input conversation snippet x_i with a composite output Y_i. This composite output explicitly encapsulates assessments for all three HL concepts (Teach-Back, Plain Language, Open-Ended Questions) as pairs of classification labels ($l_{i,c}$) and feedback texts ($f_{i,c}$):

$$Y_i = \left((l_{i,\text{TB}}, f_{i,\text{TB}}), (l_{i,\text{PL}}, f_{i,\text{PL}}), (l_{i,\text{OE}}, f_{i,\text{OE}}) \right).$$

If a concept $c \in \{\text{TB, PL, OE}\}$ was not originally assessed for snippet x_i, the corresponding label $l_{i,c}$ was set to *"Not Assessed"*, and the feedback $f_{i,c}$ was left empty. This ensures that each training instance provides a comprehensive, albeit potentially sparse, target for all three concepts.

This merged design enables the model to learn a joint mapping: $M_{\text{merged}} : X \to Y_{\text{TB}} \times Y_{\text{PL}} \times Y_{\text{OE}}$, thus promoting a holistic understanding of snippet x_i across multiple communication dimensions simultaneously. Moreover, this structure efficiently leverages all available annotated data, even when annotations are not uniformly present for each concept across all snippets. After the merging operation, the final dataset comprised 105 data points.

2) Instruction and System Prompt Engineering: Effective fine-tuning heavily relies on carefully engineered prompts that clearly define the task and guide the model towards the desired output format and reasoning process. Our approach employs a two-stage prompt structure in both the instruction and system prompts, where HL concepts are first defined for the model (**abstraction**), followed by the specification of the analytical task based on these concepts (**reasoning**). This mirrors the abstraction-then-reasoning framework suggested by Zheng et al [20], which encourages LLMs to establish high-level understanding before proceeding to detailed reasoning. For our Merged Finetuning model, we utilized a detailed instruction prompt, consistent across all training instances in $\mathcal{D}_{\text{merged}}$, complemented by a specific system prompt.

a) Instruction Prompt Design: The instruction prompt served as the primary guide for the LLM during fine-tuning. Its structure was designed in three parts to provide comprehensive context and clear task definition:

Firstly, we explicitly define the three HL concepts:

- *Teach-Back:* Characterized by clinician verification of patient understanding through patient demonstration or explanation in their own words, while avoiding simple yes/no checks like "Do you understand?".
- *Plain Language:* Defined by the replacement of medical terms with common words and the avoidance of unexplained jargon.
- *Open-Ended Questions:* Identified by questions typically starting with "How", "What", or "Why", which encourage elaboration beyond simple yes/no answers and avoid leading questions.

This initial definition phase provides the necessary conceptual grounding for the model.

Next, we provided few-shot exemplars. Two contrasting conversation examples are provided. Each example is followed by a detailed 'Analysis' section, demonstrating the desired reasoning process and output format for all three HL concepts (classification, justification, and actionable feedback with examples). These exemplars serve as in-context learning guides. In curating these examples, we adopted a strategy inspired by Active-Prompt, which emphasizes selecting impactful questions for annotation [21]. While Diao et al. [21] utilize metrics for uncertainty-based selection, our work simply employed ChatGPT-4o to generate few-shot examples that reflect high uncertainty within our health communication task domain. This approach is guided by the observation that focusing on uncertain instances can enhance performance, and our method utilizes an LLM's generative capabilities for this example selection.

Lastly, a "Your Task" section clearly outlines the model's objective for new conversations:

1) Analyze each health literacy concept sequentially.
2) Classify the conversation's adherence to each concept (Agree/Disagree/Neutral).
3) Provide specific feedback detailing what worked or did not, offering concrete improvement examples, and suggesting alternative phrasing.

A crucial directive is to "Focus on actionable feedback clinicians can implement and make the patient or caregiver feel comfortable and respected."

Each entry in $\mathcal{D}_{\text{merged}}$ is structured in the Alpaca instruction format, consisting of a detailed `instruction`, a conversational snippet as the `input` (x_i), and a corresponding composite `output` (Y_i) representing the multi-concept health literacy assessment.

b) System Prompt Design: To further guide the LLM's behavior and persona, a system prompt was utilized. Consistent with our overall approach, this prompt also incorporates a two-step design involving concept abstraction followed by task specification and includes phrasing designed to trigger CoT reasoning. Its components are:

- **Role Assignment:** The model is explicitly assigned the role of a "Health Literacy Coach trained to evaluate nurse-patient conversations" with the goal of providing "actionable, respectful feedback."
- **Core Instructions and CoT Trigger (Abstraction & Reasoning):** It reiterates the three key concepts (Teach-Back, Plain Language, Open-Ended Questions) and their essential evaluation criteria. A sentence "Analyze the provided conversation step-by-step" is included to encourage CoT reasoning.
- **Classification Guidance:** It restates the classification schema: "Agree (Met criteria), Disagree (Did not meet), or Neutral (Partial/Unclear)."
- **Feedback Structure and Quality:** The system prompt details the expected components of the feedback: identifying what worked/didn't, providing concrete examples, and suggesting alternative phrasing.
- **Tone Specification:** It mandates a "Supportive", "Specific", and "Actionable" tone for the feedback, focusing on improvement rather than criticism.
- **Output Template:** Finally, it provides a desired structured output format to ensure consistency.

This dual-prompt strategy, with its structured, two-step design, CoT triggers, and explicit output formatting, is intended to elicit reliable, actionable, and contextually appropriate multifaceted evaluations from the fine-tuned LLM, aligning its behavior with the nuanced requirements of HL assessment.

C. Model Fine-tuning

Our experiments utilized two open-source LLMs: **Mistral-7B-Instruct-v0.3** and **Llama 3.1-8B-Instruct**. Mistral-7B-Instruct-v0.3 was prioritized for its lightweight architecture and computational efficiency, making it ideal for deployment in resource-constrained pediatric care environments. Its 7B-parameter design balances performance with practical considerations such as GPU memory requirements (e.g., 16GB VRAM compatibility) and rapid inference speeds, critical for real-time audit workflows. Llama 3.1-8B-Instruct was selected as a complementary model for its status as a state-of-the-art model with robust performance on conversational tasks. This dual-model approach enabled comparative analysis of accuracy-efficiency trade-offs while ensuring compliance with open-source licensing requirements. Both models were sourced from the Hugging Face repository to ensure reproducibility [26], [27].

We fine-tuned the models using the following hyperparameters: a cutoff length of 2048 tokens, a learning rate of 2×10^{-5}, a batch size of 2, and 9 training epochs. The training was conducted on a Linux-based GPU node with three NVIDIA A100 PCIe GPUs (40 GB each) on the Lonestar6 system at the Texas Advanced Computing Center.

D. Evaluation Framework

To assess the efficacy of our Merged Finetuning model, we conducted a comparative evaluation using a dedicated test set of 9 long clinical dialogue scripts. The performance of the following methods was compared:

- **Human Expert Ratings:** Assessments provided by health literacy experts, serving as the reference standard. Ratings were on a scale where Agree = 1, Neutral = 0, and Disagree = -1 for each HL concept.
- **Baseline HealthLit System:** The performance of our original RAG-based system using Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct models.
- **Merged Finetuning Model:** The performance of the same models after fine-tuning on the $\mathcal{D}_{\text{merged}}$ dataset without RAG.

The primary evaluation metric was the level of agreement between model outputs and human expert ratings for each HL concept. This was quantified using two complementary measures: agreement rate and Cohen's Kappa.

Agreement rate refers to the proportion of instances where a model's categorical rating exactly matched the expert's rating for a given transcript segment and HL practice. However, the agreement rate does not account for chance agreement and may overestimate reliability in cases with unbalanced label distributions. To address this, we also computed Cohen's Kappa, a standard statistical measure of inter-rater reliability that corrects for the level of agreement expected by random chance. Both metrics were reported for each model configuration and HL concept to holistically evaluate model performance.

IV. Results

This section presents the performance of the Merged Finetuning models compared to our Baseline HealthLit system and human health literacy expert ratings. The evaluation was conducted on 9 distinct clinical dialogue scripts, assessing three key HL practices: Teach-Back, Plain Language, and Open-Ended Questions. Performance is reported based on the agreement with expert ratings.

A. Overall Model Performance Comparison

We first evaluated the overall performance of each model by calculating the agreement rate and Cohen's Kappa coefficient against human expert ratings across all scripts and all three HL practices.

TABLE I: Overall Agreement Rate and Cohen's Kappa with Human Expert Ratings Across All HL Practices.

Model Configuration	Agreement Rate	Cohen's Kappa
Baseline Llama 8B (RAG)	25.93%	-0.040
Merged Finetuning Llama 8B	62.96%	0.357
Baseline Mistral 7B (RAG)	29.63%	-0.179
Merged Finetuning Mistral 7B	37.04%	0.084

TABLE II: Comparison of Model Outputs with Expert Ratings for the "24 hr urine collection" Script

		Baseline [1]		Merged Finetuning	
HL	**Exp**	**Llama 8B**	**Mistral 7B**	**Llama 8B**	**Mistral 7B**
OE	-1	0	0	-1	0
PL	1	0	1	1	0
TB	1	0	0	1	1

HL = Health Literacy Concepts, Exp = Expert

Table I summarizes these overall performance metrics. The Merged Finetuning Llama 8B model achieved an overall agreement rate of 62.96% with human experts, with a substantial improvement from the 25.93% achieved by the Baseline Llama 8B (RAG) model. Its Cohen's Kappa increased from -0.040 (slight disagreement) for the baseline to 0.357, indicating fair agreement for the fine-tuned version. While the Mistral 7B models showed a similar trend of improvement with fine-tuning, the gains are smaller than LLama model. The Merged Finetuning version reached 37.04% agreement (Kappa: 0.084, slight agreement), up from 29.63% (Kappa: -0.179, slight disagreement). These results strongly suggest that the Merged Fine-tuning approach generally enhanced model alignment with health literacy expert evaluations, with the Llama 8B architecture demonstrating a more pronounced benefit from this strategy.

B. Performance Breakdown by Health Literacy Practice

To further understand the impact of our merged fine-tuning approach, we analyzed model performance for each of the three HL practices as shown in Fig. 2.

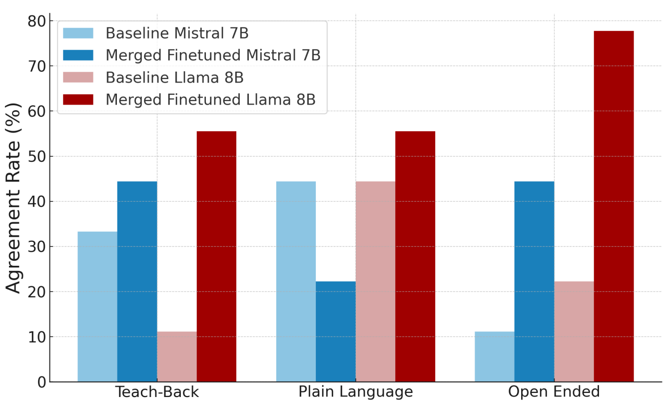

Fig. 2: Agreement Rates By Health Literacy Practice

For the Llama 8B models, the Merged Finetuning approach significantly increased agreement rates. For Teach-Back, it increased fivefold from 11.11% to 55.56%, and for Plain Language, it increased from 44.44% to 55.56%. Most strikingly, for Open-Ended Questions, the fine-tuned Llama 8B model's agreement rate increased more than threefold, from 22.22% to 77.78%. Regarding the Mistral 7B models, the Merged Finetuning model showed mixed results. Agreement for Teach-Back increased from 33.33% to 44.44%. For Open-Ended Questions, it quadrupled from a low score of 11.11% to 44.44%. However, for Plain Language, the agreement rate decreased from 44.44% to 22.22%. The most significant improvements were seen in Open-Ended Questions for both model architectures, suggesting that fine-tuning on the merged dataset particularly enhanced the models' ability to identify and evaluate OE practice.

C. Qualitative Analysis of Model Outputs

While aggregate scores provide a quantitative overview, a qualitative analysis of model outputs for specific scripts can reveal more granular differences in performance and feedback quality. In this section, we focus on the "24hr urine collection" script to illustrate key distinctions between the baseline RAG models and the new Merged Finetuning models. Table II summarizes the ratings from experts, baseline, and our proposed merged fine-tuning model for this script.

A primary operational difference, as noted in our methodology, is the efficiency of output generation. The baseline models required three separate, concept-specific queries to generate a complete assessment for a single script. In contrast, both the Merged Finetuning Llama 8B and Mistral 7B models produced a comprehensive, structured report covering all three HL concepts from a single prompt containing the system instructions and the conversation transcript. Please see the example output from the Llama-based merged fine-tuning model. This represents a significant improvement in usability and system efficiency.

Comparing the content and quality of feedback for the "24hr urine collection" script further highlights these differences:

1) Llama 8B Model Family: The Merged Finetuning Llama 8B model generally demonstrated improved alignment with expert ratings and feedback quality.

Teach-Back: The expert rated this as Agree (1). The Baseline rated it Neutral (0). Although its rating differed, the Baseline RAG model's feedback was educationally rich, detailing why a simple request to repeat instructions was insufficient and offering strong alternative phrasing for true teach-back. The Merged Finetuning model correctly aligned with the expert's Agree (1) rating, identifying that the clinician verified understanding through patient recall. This highlights an interesting trade-off: the merged model achieved better rating accuracy in this instance, while the baseline RAG, despite a differing rating, offered more in-depth formative guidance.

Plain Language: The expert rated this as Agree (1). The Baseline rated it Neutral (0) and seems conservative when rating and the feedback is verbose. In contrast, the Merged Finetuning Llama 8B model achieved an Agree (1) rating, aligning with the expert. Its feedback was concise and effectively praised the clinician's use of clear language and analogies (e.g., "little capsule").

Open-Ended Questions: The expert rated this as Disagree (-1). The Baseline rated it Neutral (0) and its example output revealed a critical reliability issue: it classified the interaction as "Neutral" while its textual explanation stated no open-ended questions were used. The Merged Finetuning Llama 8B model, however, correctly aligned with the expert's Disagree (-1) rating. Furthermore, its feedback was highly specific and actionable, providing excellent examples of alternative open-ended questions (e.g., "What are some potential challenges...").

2) Mistral 7B Model Family: The Mistral 7B models also showed benefits from the merged fine-tuning approach in terms of output structure, though with some variations in assessment accuracy for this specific script.

Teach-Back: Aligning with the expert, the Merged Finetuning Mistral 7B model correctly identified the clinician's attempt to have the patient explain in their own words. The Baseline rated this Neutral (0), but similarly to the Llama Baseline, it provided good suggestions for more explicit teach-back questions.

Plain Language: The expert rated this as Agree (1), and the Baseline Mistral 7B (RAG) also rated as Agree (1), though its example feedback offered unnecessary suggestions like simplifying the term "link." The Merged Finetuning Mistral 7B model, however, diverged with a Neutral (0) rating. Its feedback stated: "There is some use of medical terms, but the clinician also simplifies some terms (e.g., '24-hour day' instead of 'diurnal')." Notably, the term "diurnal" was not present in the original script, indicating that the model introduced this term in its own reasoning. This suggests that while attempting a nuanced assessment, the model's justification for its 'Neutral' rating was based on an internally generated, incorrect comparison, potentially impacting its accuracy for this specific instance.

Open-Ended Questions: Rated Disagree (-1) by the expert. The Baseline provided a Neutral (0) rating, its example output identifying one closed question and one question it considered 'open' ("Do you have any other questions?"), while it is still a yes/no closed question. The Merged Finetuning Mistral 7B model similarly misidentified a declarative statement as a question and also cited "Do you have any questions?" as an open-ended example. This suggests that accurately distinguishing and providing feedback on open-ended questions remained challenging for the Mistral 7B architecture, even after fine-tuning on the merged dataset.

In summary, for the "24hr urine collection" script, the Merged Finetuning Llama 8B model demonstrated strong alignment with expert ratings and provided coherent, specific, and actionable feedback, particularly excelling in the Open-Ended Questions assessment, where its baseline counterpart showed significant inconsistencies. The Baseline models, for both Llama and Mistral, while capable of producing detailed educational insights when their analysis was on point, reasoning errors still exist. The Merged Finetuning Mistral 7B model also benefited from improved output structure but exhibited less accurate reasoning for Open-Ended Questions and an instance of self-generated incorrect comparison for Plain Language in this example. These qualitative insights underscore that while merged fine-tuning enhances efficiency and structural consistency, the degree of improvement in nuanced accuracy and reliability can vary between model architectures and specific HL concepts.

Output from Merged Finetuning Llama 8B Model

Output:
Open-Ended Questions:
Classification: Disagree
Feedback: The clinician did not ask any open-ended questions during the conversation. The conversation was mostly a one-way transmission of information, with the clinician explaining the instructions and the patient and parent repeating them back. One suggestion for improvement is to ask open-ended questions to encourage further discussion and ensure understanding. For example, the clinician could have asked "What are some potential challenges you may face when administering the antibiotic treatment?" or "How will you know if your child is experiencing any side effects from the antibiotic treatment?".
Plain Language:
Classification: Agree
Feedback: The clinician used clear and simple language to explain the information about the antibiotic treatment and the importance of flushing out the bacteria. The clinician avoided using technical jargon and instead used analogies, such as comparing the antibiotic treatment to a "little capsule that he'll take three times a day for seven days", to help the patient and parent understand.
Teach-Back:
Classification: Agree
Feedback: The clinician did an excellent job of teaching the patient and their parent about the antibiotic treatment and the importance of flushing out the bacteria. The clinician asked the patient and their parent to repeat back the information to ensure understanding, and the patient and their parent were able to summarize the instructions correctly.

Future work will further prioritize full-context transcripts to improve accuracy, though human experts' availability remains a constraint. A hybrid AI-Expert review pipeline to assist the human annotation process by leveraging generative AI to pre-select semantically meaningful chunks of conversation and suggest possible feedback that might be helpful. Human experts can then review and refine these AI-generated suggestions, significantly accelerating the annotation process

Beyond data constraints, we are exploring Knowledge-Augmented Generation (KAG) [28] to embed domain-specific health literacy principles (e.g., numerical/temporal reasoning for medication adherence) within the model's vector space. Preliminary results indicate that KAG improves logical reasoning and aligns outputs more closely with expert judgment, particularly for low-health-literacy populations.

V. Conclusion

This study extended our initial HealthLit system by addressing the limitations of isolated concept assessment in clinical dialogue analysis. We introduced a "Merged Finetuning" method, fine-tuning Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct models on a custom dataset and prompt framework designed to jointly evaluate Teach-Back, Plain Language, and Open-Ended Questions from a single conversational input. The Merged Finetuning Llama 8B model significantly improved alignment with expert ratings. Mistral 7B also showed gains, though its qualitative performance lagged behind Llama 8B. The merged finetuning allowed the generation of cohesive feedback from a single prompt, in contrast to the fragmented, multi-query outputs of the baseline. However, the study was fine-tuned and evaluated on a limited amount of data. Broader validation with more data and different languages, for example Spanish, is needed. In addition, consistent, nuanced interpretation across HL domains remains challenging and needs high-quality annotations. Overall, our findings emphasize the value of specialized datasets and targeted fine-tuning in adapting LLMs for complex, real-world healthcare communication, offering a path toward more effective clinician support tools grounded in health literacy principles.

Acknowledgment

This work has been supported in part by NSF grants CNS-2231519 and DUE-2225229.

References

[1] H. Tran, S. Shah, D. Chen, J. Wang, Y. Feng, C. Howe, L. Patton, L. Ma, and S. Fu, "Healthlit: A large language model driven health literacy fidelity audit and feedback system," in *Workshop on Data Quality Aware, High-Performance, and Trustworthy AI Systems for Healthcare, IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, June 2025.

[2] L. Y. Jiang, X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi, *et al.*, "Health system-scale language models are all-purpose prediction engines," *Nature*, vol. 619, no. 7969, pp. 357–362, 2023.

[3] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023.

[4] X. Luo, Z. Deng, B. Yang, and M. Y. Luo, "Pre-trained language models in medicine: A survey," *Artificial Intelligence in Medicine*, p. 102904, 2024.

[5] M. Kim, Y. Kim, H. J. Kang, H. Seo, H. Choi, J. Han, G. Kee, S. Park, S. Ko, H. Jung, *et al.*, "Fine-tuning llms with medical data: Can safety be ensured?," *NEJM AI*, vol. 2, no. 1, p. AIcs2400390, 2025.

[6] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of biomedical informatics*, vol. 156, p. 104662, 2024.

[7] D. Fraile Navarro, E. Coiera, T. W. Hambly, Z. Triplett, N. Asif, A. Susanto, A. Chowdhury, A. Azcoaga Lorenzo, M. Dras, and S. Berkovsky, "Expert evaluation of large language models for clinical dialogue summarization," *Scientific Reports*, vol. 15, no. 1, p. 1195, 2025.

[8] D. Nutbeam, "Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century," *Health promotion international*, vol. 15, no. 3, pp. 259–267, 2000.

[9] R. Shahid, M. Shoker, L. M. Chu, R. Frehlick, H. Ward, and P. Pahwa, "Impact of low health literacy on patients' health outcomes: a multicenter cohort study," *BMC health services research*, vol. 22, no. 1, p. 1148, 2022.

[10] N. Pronk, D. V. Kleinman, S. F. Goekler, E. Ochiai, C. Blakey, and K. H. Brewer, "Promoting health and well-being in healthy people 2030," *Journal of Public Health Management and Practice*, vol. 27, no. Supplement 6, pp. S242–S248, 2021.

[11] T. T. H. Dinh, A. Bonner, R. Clark, J. Ramsbotham, and S. Hines, "The effectiveness of the teach-back method on adherence and self-management in health education for people with chronic disease: a systematic review," *JBI Evidence Synthesis*, vol. 14, no. 1, pp. 210–247, 2016.

[12] Centers for Disease Control and Prevention (CDC), "Develop and test materials: Plain language," 2024.

[13] N. M. Ivers, A. Sales, H. Colquhoun, S. Michie, R. Foy, J. J. Francis, and J. M. Grimshaw, "No more 'business as usual'with audit and feedback interventions: towards an agenda for a reinvigorated intervention," *Implementation Science*, vol. 9, pp. 1–8, 2014.

[14] P. Butow and E. Hoque, "Using artificial intelligence to analyse and teach communication in healthcare," *The breast*, vol. 50, pp. 49–55, 2020.

[15] D. Vithanage, C. Deng, L. Wang, M. Yin, M. Alkhalaf, Z. Zhang, Y. Zhu, and P. Yu, "Adapting generative large language models for information extraction from unstructured electronic health records in residential aged care: A comparative analysis of training approaches," *Journal of Healthcare Informatics Research*, pp. 1–29, 2025.

[16] A. Bodonhelyi, C. Stegemann-Philipps, A. Sonanini, L. Herschbach, M. Szép, A. Herrmann-Werner, T. Festl-Wietek, E. Kasneci, and F. Holderried, "Modeling challenging patient interactions: Llms for medical communication training," *arXiv preprint arXiv:2503.22250*, 2025.

[17] X. Amatriain, "Prompt Design and Engineering: Introduction and Advanced Methods," May 2024. arXiv:2401.14423 [cs].

[18] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," Mar. 2025. arXiv:2402.07927 [cs].

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[20] H. S. Zheng, S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, and D. Zhou, "Take a step back: Evoking reasoning via abstraction in large language models," *arXiv preprint arXiv:2310.06117*, 2023.

[21] S. Diao, P. Wang, Y. Lin, R. Pan, X. Liu, and T. Zhang, "Active prompting with chain-of-thought for large language models," *arXiv preprint arXiv:2302.12246*, 2023.

[22] L. Zhang, T. Ergen, L. Logeswaran, M. Lee, and D. Jurgens, "SPRIG: Improving Large Language Model Performance by System Prompt Optimization," Oct. 2024. arXiv:2410.14826 [cs].

[23] N. Wu, M. Gong, L. Shou, S. Liang, and D. Jiang, "Large language models are diverse role-players for summarization evaluation," in *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 695–707, Springer, 2023.

[24] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong, "Better zero-shot reasoning with role-play prompting," *arXiv preprint arXiv:2308.07702*, 2023.

[25] J. Kim, N. Yang, and K. Jung, "Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks," *arXiv preprint arXiv:2408.08631*, 2024.

[26] M. AI, "Mistral-7b-instruct-v0.3." https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.

[27] Meta, "Llama-3.1-8b-instruct." https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024.

[28] L. Liang, M. Sun, Z. Gui, Z. Zhu, Z. Jiang, L. Zhong, Y. Qu, P. Zhao, Z. Bo, J. Yang, H. Xiong, L. Yuan, J. Xu, Z. Wang, Z. Zhang, W. Zhang, H. Chen, W. Chen, and J. Zhou, "Kag: Boosting llms in professional domains via knowledge augmented generation," 2024.