HealthLit: A Large Language Model Driven Health Literacy Fidelity Audit and Feedback System

Hang Tran*[1], Sameep Shah*[2], Donger Chen*[1], Jixin Wang[1], Yunhe Feng[1], Carol Howe[2], Lindsey Patton[3], Liran Ma[4], Song Fu[1]
[1]University of North Texas, [2]Texas Christian University, [3]Children's Health, [4]Miami University
USA

ABSTRACT

Effective health communication is crucial for patient outcomes, but clinicians often struggle to consistently apply health literacy best practices. Audit and Feedback (A&F) is a proven method for improving adherence to guidelines, but traditional manual audits are costly and unscalable. This paper introduces *HealthLit*, an AI-driven A&F system that uses large language models (LLMs) to automatically assess clinician fidelity to health literacy principles in healthcare interactions. We fine-tuned Mistral 7B and Llama3 8B on a dataset of 212 annotated healthcare interactions, incorporating Retrieval-Augmented Generation (RAG) for enhanced contextual understanding. *HealthLit* demonstrates strong performance in identifying key health literacy practices, with Llama3 8B achieving higher accuracy than Mistral 7B. User studies with healthcare professionals indicate that *HealthLit*'s feedback is acceptable, appropriate, and feasible for real-world application. Our results highlight the potential of LLMs to significantly improve the scalability and accessibility of health literacy training and practice.

KEYWORDS

Health Literacy, Large Language Model, Retrieval Augmented Generation, Natural Language Processing, AI.

ACM Reference Format:
Hang Tran*[1], Sameep Shah*[2], Donger Chen*[1], Jixin Wang[1], Yunhe Feng[1], Carol Howe[2], Lindsey Patton[3], Liran Ma[4], Song Fu[1]. 2025. *HealthLit*: A Large Language Model Driven Health Literacy Fidelity Audit and Feedback System. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3721201.3725519

CHASE '25, June 24–26, 2025, New York, NY, USA

ACM ISBN 979-8-4007-1539-6/2025/06
https://doi.org/10.1145/3721201.3725519

1 INTRODUCTION

Effective communication is crucial in healthcare. Low health literacy – the ability to find, understand, and use health information – is a significant barrier to positive patient outcomes [11]. While *Healthy People 2030* [12] emphasizes the importance of organizational health literacy, many healthcare systems struggle to consistent implement best practices. Research shows that specific communication strategies, such as limiting jargon, using teach-back, and structuring information clearly, can improve patient comprehension and adherence [2]. However, widespread adoption is hindered by factors such as time constraints, insufficient training, and limited resources.

Audit and Feedback (A&F) is a well-established intervention to promote guideline adherence by providing clinicians with performance feedback relatively to established standards[6]. Tradditional A&F, relies on manual reviews by health literacy experts (HLEs), which is labor-intensive, expensive, and difficult to scale. Recent advances in natural language processing (NLP), particularly large language model (LLM) offer promising methods to automate A&F. Prior work has applied ML models in automating A&F in contexts like motivational interviewing [7], however, their use in health literacy remains unexplored.

This paper introduces *HealthLit*, an LLM-driven A&F system designed to automatically audit and provide feedback on clinician fidelity to established health literacy practices. We address the limitations of manual audits by leveraging the capabilities of LLMs to analyze healthcare interactions and generate actionable feedback. *HealthLit* employs Retrieval-Augmented Generation (RAG) to provide the fine-tuned models with relevant context from healthcare interactions, ensuring an accurate assessment of health literacy practices. Structured prompts, crafted using prompt engineering techniques, guide the model to perform a step-by-step evaluation, mirroring the approach of human experts.

*These authors contributed equally to this work.

By automating the health literacy assessment process, *HealthLit* aims to reduce the burden on human experts, increase scalability, and ultimately promote the wider adoption of evidence-based communication strategies in healthcare settings. The key contributions of this work are:

- The development and evaluation of *HealthLit*, an AI-driven A&F system for health literacy, utilizing fine-tuned LLMs (Mistral 7B [10] and Llama3 8B [3]) and Retrieval-Augmented Generation (RAG).
- A comparative analysis of *HealthLit*'s AI-generated audits against expert HLE assessments, focusing on completeness, accuracy, and efficiency.
- An evaluation of the usability of *HealthLit*'s feedback through user studies with healthcare professionals.

2 BACKGROUND AND RELATED WORK

This section provides background on health literacy, describes key health literacy practices, and reviews relevant work on A&F.

Health Literacy and its Importance

Health literacy refers to an individual's ability to obtain, process, and understand basic health information to make appropriate health decisions [11]. Low health literacy associated with higher hospitalization rates, poor adherence to medical guidance, and worse health outcomes [13]

Key Health Literacy Practices Research has identified several key communication strategies to enhance patient understanding and engagement in healthcare. The teach-back method-asking patients to reiterate information, helps identify and correct misunderstandings [2, 4]. Using open-ended questions, rather than a simple yes/no answer, encourages patients to express their concerns and understanding more fully. Furthermore, employing plain language, free of medical jargon, is crucial for improving comprehension, especially for those with lower health literacy [1]. Other supporting strategies involve chunking and checking information as well as using visual aids.

Audit and Feedback (A&F) A&F is a widely used quality improvement strategy that involves providing healthcare professionals with summaries of their clinical performance over time[8, 9]. It has proven effective in improving adherence to guidelines across various healthcare domains, including medication prescribing, diagnostic testing, and preventative care. The effectiveness of A&F depends on several factors, including the source, the format, and the frequency of feedback delivery [8].

3 SYSTEM DESIGN

3.1 *HealthLit* System Overview

HealthLit is an LLM-driven Audit and Feedback system designed to assess clinician adherence to health literacy best practices. As depicted in Figure 1, the system architecture

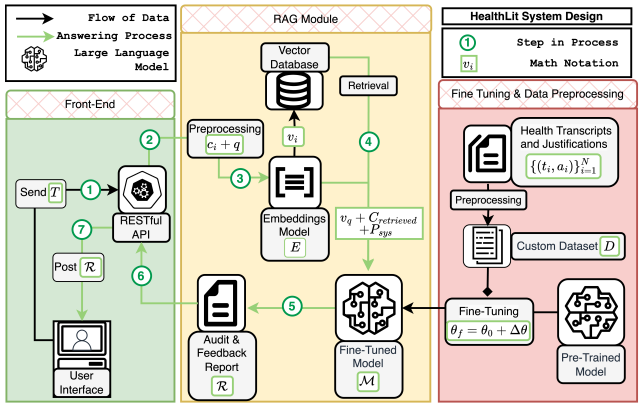

Figure 1: *HealthLit* System.

comprises three primary components: fine-tuned LLMs, a Retrieval-Augmented Generation (RAG) module, and a prompt engineering framework. We represent the overall system as a function $\mathcal{H}$ that takes a healthcare interaction transcript T as input and outputs an audit and feedback report R:

$$R = \mathcal{H}(T) \tag{1}$$

The system $\mathcal{H}(T)$ can be further decomposed into its core components:

$$\mathcal{H}(T) = \mathcal{P}(\mathcal{R}(T, \mathcal{M}(T; \theta_f))) \tag{2}$$

where $\mathcal{M}$ represents the fine-tuned LLM (either Mistral 7B [10] or Llama3 8B [3]), parameterized by θ_f after fine-tuning; $\mathcal{R}$ represents the RAG module; and $\mathcal{P}$ represents the prompt engineering framework.

3.2 Fine-Tuning LLMs

LLMs are typically pre-trained on vast, general-domain corpora, which may limit their performance on specialized tasks requiring domain-specific knowledge, such as health literacy assessment. To address this, we fine-tune pre-trained LLMs on a curated dataset $D = \{(t_i, a_i)\}_{i=1}^{N}$, where t_i is a healthcare interaction transcript and a_i represents the corresponding expert annotations for health literacy practices. We employ Low-Rank Adaptation (LoRA) [5] using LLaMA-Factory [18] for efficient fine-tuning. LoRA adapts a pre-trained LLM with parameters θ_0 by introducing trainable rank decomposition matrices. Instead of updating all parameters in θ_0, LoRA updates a small set of parameters $\Delta\theta$, such that the updated parameters θ_f are given by:

$$\theta_f = \theta_0 + \Delta\theta \tag{3}$$

where $\Delta\theta$ represents a low-rank update. This significantly reduces the number of trainable parameters, memory requirements, and training time. Specifically, we fine-tuned the following models (although HealthLit's modular design allows for integration of other LLMs):

- **Mistral 7B** [10]: A lightweight yet powerful model optimized for efficiency and performance.

- **Llama3 8B** [3]: A larger model with enhanced contextual understanding and reasoning capabilities.

3.3 RAG Module

While fine-tuning improves LLM performance, limitations such as potential hallucinations and difficulties with nuanced medical terminology remain, particularly critical in healthcare. To mitigate these, we incorporate a Retrieval-Augmented Generation (RAG) module. The RAG module operates as follows:

Preprocessing Input healthcare transcripts T are segmented into smaller, semantically coherent chunks $\{c_1, c_2, ..., c_n\}$. These chunks are then transformed into vector embeddings using an embedding function E:

$$v_i = E(c_i), \quad \forall i \in \{1, 2, ..., n\} \tag{4}$$

The set of embeddings $V = \{v_1, v_2, ..., v_n\}$ forms the knowledge base for the RAG module.

Query Embedding Given a query q (e.g., a question about health literacy practices in the transcript), it is also embedded using the same embedding function:

$$v_q = E(q) \tag{5}$$

Retrieval The similarity between the query embedding v_q and each chunk embedding v_i is calculated using a similarity function S (e.g., cosine similarity). The top-k most relevant chunks are retrieved:

$$C_{retrieved} = \text{top-}k(\{c_i | \forall i \in \{1, 2, ..., n\}\}, S(v_q, v_i)) \tag{6}$$

Augmented Generation The retrieved chunks $C_{retrieved}$ are provided to the fine-tuned LLM $\mathcal{M}$ as additional context, along with the original query q. The LLM then generates the response r:

$$r = \mathcal{M}(q, C_{retrieved}; \theta_f) \tag{7}$$

The RAG module, therefore, enhances the LLM's responses by grounding them in the specific content of the transcript, reducing the risk of hallucinations and improving accuracy. The direct link between retrieved chunks and their embeddings enables source attribution, enhancing transparency.

3.4 Prompt Engineering

Effective prompt engineering is critical for guiding LLMs. We employ a two-pronged approach:

Fine-Tuning Prompts: The fine-tuning dataset D includes prompts designed to evaluate specific health literacy techniques (e.g., Teach-Back). Let p_i be a prompt paired with a transcript segment t_i, forming an input-output pair (p_i, t_i, a_i) used for fine-tuning. These prompts are designed following principles from healthcare NLP research [15], focusing on objectives like patient understanding and comfort.

System Prompt for Inference: At inference time, a structured system prompt P_{sys} guides the LLM's evaluation. This prompt defines the LLM's role as a "skilled language evaluator" and enforces a structured reasoning process, $P_{sys} = (R, I, O)$, where R (Role): Defines the LLM's role (e.g., "skilled language evaluator"); I (Instructions): Specifies the task, including referencing evidence-based health literacy practices like the Teach-Back method [17]; and O (Output Format): Defines the desired output structure (e.g., identifying strengths, areas for improvement, and specific suggestions).

The final output R is then generated by applying the system prompt to the LLM's output with the RAG context:

$$R = \mathcal{P}(\mathcal{M}(q, C_{retrieved}; \theta_f)) = P_{sys}(\mathcal{M}`(q, C_{retrieved}; \theta_f))$$

This approach reduces output variability, a key requirement for reliable deployment in healthcare settings [16].

4 EXPERIMENTAL EVALUATION

4.1 Dataset.

To train and evaluate *HealthLit*, we curated a dataset of simulated healthcare interactions designed to exemplify varying levels of adherence to health literacy best practices. The dataset was derived from 33 audio recordings of healthcare interactions, covering diverse clinical specialties. While the specific medical content might vary across recordings, the underlying principles of effective health communication remained consistent.

The recordings were professionally transcribed using TranscribeMe [14]. Initially, we used the full transcripts and corresponding expert evaluations as input-output pairs for model training. However, preliminary experiments revealed that the length of these transcripts hindered the LLMs' ability to effectively identify and learn key health literacy concepts.

To address this challenge, we segmented the transcripts into smaller, semantically coherent blocks. Health Literacy Experts (HLEs) then evaluated each segment independently. The HLE evaluation for each segment consisted of:

- Categorical ratings: "Agree", "Disagree", or "Neutral", depending on whether the HLE believed the nurse in the conversation adhered to health literacy principles.
- Suggestion for improvement: HLE provided insights on how the healthcare provider could better incorporate health literacy practices.

To promote consistency, key health literacy concepts (e.g., Teach-Back, open-ended questions, plain language) were embedded in the LLM prompts during fine-tuning and inference, enabling the model to better recognize and assess these practices.

The final dataset comprises 212 data points, each consisting of three components:

- Instruction: An embedded prompt guiding the model's response.
- Input: The segmented transcribed healthcare interaction.
- Output: The expert evaluation of the transcript.

Script name	HL Practice	*HealthLit* A							
		HLEs rating	HL-rating	Quality Feedback					
				Specificity	Actionability	Constructiveness	Accuracy	Quality score	**Total Quality Score**
24 hr urine. Health literacy expert	Open Ended	0	1	1	2	2	0	5	15
	plain language	2	2	0	2	2	0	4	
	Teach Back	2	1	2	2	2	0	6	
Biofeedback Data Sheet	Open Ended	2	0	2	2	2	0	6	19
	plain language	0	2	2	2	2	1	7	
	Teach Back	2	0	2	2	2	0	6	
Catheter Removal. Data Sheet	Open Ended	0	2	2	2	2	0	6	17
	plain language	2	2	2	2	2	2	8	
	Teach Back	0	2	1	0	2	0	3	
Circumcision_ Health literacy expert	Open Ended	2	2	1	2	2	0	5	16
	plain language	2	2	1	2	2	0	5	
	Teach Back	2	2	2	2	2	0	6	
Epi Pen Data Sheet	Open Ended	0	2	2	2	2	0	6	19
	plain language	2	2	2	2	2	2	8	
	Teach Back	1	2	1	2	2	0	5	
Nasogastric Tube Placement Data Sheet	Open Ended	0	2	0	0	2	0	2	18
	plain language	1	2	2	2	2	2	8	
	Teach Back	2	2	2	2	2	2	8	
Neurogenic bowel _Health literacy	Open Ended	1	2	2	2	1	0	5	18
	plain language	2	1	2	2	2	1	7	
	Teach Back	2	2	1	2	2	1	6	
RVP instructions _Health literacy	Open Ended	0	2	2	2	2	0	6	19
	plain language	0	2	2	2	2	2	8	
	Teach Back	0	1	1	2	2	0	5	
UTI Antibiotics Data Sheet	Open Ended	0	2	2	2	2	0	6	20
	plain language	1	2	2	2	2	2	8	
	Teach Back	1	2	2	2	2	0	6	

Table 1: Peformance of HealthLit with Mistral 7B Fine-tuned Model Integrated in RAG.

Script name	HL Practice	*HealthLit* B							
		HLEs rating	HL-rating	Quality Feedback					
				Specificity	Actionability	Constructiveness	Accuracy	Quality score	**Total Quality Score**
24 hr urine. Health literacy expert	Open Ended	0	1	1	0	0	2	3	15
	plain language	2	1	2	2	2	0	6	
	Teach Back	2	1	2	2	2	0	6	
Biofeedback Data Shee	Open Ended	2	2	2	2	2	0	6	20
	plain language	0	1	1	2	2	1	6	
	Teach Back	2	1	2	2	2	2	8	
Catheter Removal. Data Sheet	Open Ended	0	1	2	2	2	2	8	22
	plain language	2	2	2	1	2	1	6	
	Teach Back	0	1	2	2	2	2	8	
Circumcision_ Health literacy expert	Open Ended	0	1	2	2	2	0	6	18
	plain language	2	2	2	2	2	0	6	
	Teach Back	0	1	2	2	2	0	6	
Epi Pen Data Sheet	Open Ended	0	2	2	2	2	2	8	21
	plain language	2	2	2	2	2	2	8	
	Teach Back	1	2	1	2	2	0	5	
Nasogastric Tube Placement Data Sheet	Open Ended	0	2	1	1	2	1	5	15
	plain language	1	1	1	1	1	0	3	
	Teach Back	2	1	2	2	2	1	7	
Neurogenic bowel _Health literacy	Open Ended	1	0	2	2	2	1	7	20
	plain language	2	1	2	2	2	2	8	
	Teach Back	2	0	1	2	2	0	5	
RVP instructions _Health literacy	Open Ended	0	0	2	2	2	1	7	23
	plain language	0	1	2	2	2	2	8	
	Teach Back	0	1	2	2	2	2	8	
UTI Antibiotics Data Sheet	Open Ended	0	1	2	2	2	2	8	24
	plain language	1	1	2	2	2	2	8	
	Teach Back	1	1	2	2	2	2	8	

Table 2: Peformance of HealthLit with Llama3 8B Fine-tuned Model Integrated in RAG.

4.2 Quantitative Evaluation

We evaluated *HealthLit*'s performance in identifying the presence or absence of key health literacy practices in transcribed healthcare interactions. Two fine-tuned LLMs, Mistral 7B [10] and Llama3 8B [3], both integrated with Retrieval-Augmented Generation (RAG), were used. The evaluation framework compared the automated assessments of *HealthLit* (HL) with those of a Health Literacy Expert (HLE). Both *HealthLit* and the HLE rated each transcript segment on a three-point scale: 0 (Disagree), 1 (Neutral), and 2 (Agree). These ratings indicate whether a specific health literacy practice (e.g., plain language, teach-back, open-ended questions) was absent (Disagree), partially present (Neutral), or present (Agree) in the segment.

In addition to the practice-specific ratings, we assessed the quality of the feedback generated by *HealthLit* using a comprehensive Quality Feedback Evaluation Rubric. This rubric measures five dimensions: Specificity, Actionability, Constructiveness, and Accuracy. Each dimension is rated on a scale of 0 (Poor), 1 (Fair), to 2 (Good). A total Quality Score (ranging from 0 to 8) is calculated, providing a holistic measure of feedback quality. A higher score indicates that the feedback is specific, actionable, constructive, and accurate. It is important to note that the Quality Score is independent of the presence/absence ratings of the health literacy practices themselves; it assesses the quality of the generated feedback, not the correctness of the practice identification. The HLEs compared and evaluated both types of assessments (practice presence and feedback quality). Table 1 (Mistral 7B, *HealthLit* A) and Table 2 (Llama3 8B, *HealthLit* B) present the results of the HLE evaluations.

Analysis of the results revealed distinct performance characteristics between the two models. Mistral 7B showed good agreement with the HLE when health literacy practices were clearly present. For example, segments demonstrably using plain language or the teach-back method were typically rated "Agree (2)" by both HL-Mistral 7B and the HLE. However, in ambiguous or borderline cases, Mistral 7B exhibited a tendency towards "Neutral (1)" or, less frequently, an incorrect "Agree (2)" rating when the HLE assessed the practice as absent. This suggests a potential limitation in Mistral 7B's ability to discern subtle contextual cues, possibly due to its smaller model size.

In contrast, Llama3 8B, with its larger model architecture, demonstrated greater sensitivity to nuanced linguistic features. Llama3 8B's ratings more frequently aligned with the HLE's assessments, particularly in assigning "Agree (2)" and "Neutral (1)" ratings. This closer alignment indicates that Llama3 8B is more effective at capturing the presence or partial presence of key health literacy practices, even when indicators are subtle. The improved performance of Llama3 8B in borderline cases suggests a superior ability to differentiate between clearly present, ambiguous, and absent health literacy features.

Regarding feedback quality, Llama3 8B consistently outperformed Mistral 7B. Llama3 8B achieved an average overall Quality Score of 6.59 across the three key health literacy practices, while Mistral 7B scored 5.96. Notably, Mistral 7B scored 0 for 18 times on the Accuracy criterion, indicating significant limitations in providing precise and accurate feedback. The Llama3 8B model, however, consistently provided more accurate, actionable, and specific recommendations.

In conclusion, Llama3 8B demonstrated superior performance compared to Mistral 7B in both accurately identifying key health literacy practices and generating high-quality, actionable feedback. These findings highlight Llama3 8B's greater potential for supporting effective improvements in clinical communication.

4.3 Qualitative Evaluation

Besides the quantitative evaluation, we performed a qualitative analysis to understand how the RAG module and structured system prompt enhance *HealthLit*'s ability to identify and provide feedback on health literacy practices. Examples from a transcript segment about UTI antibiotic treatment ("UTI Antibiotics V1"), detailing nurse-parent interactions, illustrate the benefits of these components.

To demonstrate the effectiveness of *HealthLit*'s components (fine-tuning, RAG, and structured system prompts), we conducted an ablation study comparing its performance against baseline models. We compared outputs from *HealthLit* (both the Mistral 7B and Llama3 8B versions) with those from the pre-trained Llama3 8B and Mistral 7B models without fine-tuning, RAG, or structured system prompts. The pre-trained models were provided with the same medical transcript and the query: *"Did the nurse use effective Teach Back questions to gauge patient understanding?"*. The *HealthLit* outputs were derived from the system's generated audit reports.

The comparison reveals significant differences. *HealthLit*'s outputs exhibit a clearly defined analytical structure, directly attributable to the structured system prompt and fine-tuning. This structure is crucial for providing effective feedback to healthcare practitioners, as it explicitly identifies strengths, weaknesses, and suggestions for improved phrasing, with supporting examples from the transcript. In contrast, the outputs from the pre-trained models are primarily descriptive, focusing on extracting examples without providing a structured analysis of adherence to health literacy best practices. Furthermore, the pre-trained models frequently misclassify conversational elements. For instance, questions like "Any questions?" and "Does he have any allergies...?" were incorrectly identified as open-ended questions. This inaccurate

feedback limits its usefulness for training and improvement, as clinicians may receive incorrect or insufficiently focused guidance. The structured, analytical output of *HealthLit*, enabled by fine-tuning, RAG, and prompt engineering, provides significantly more valuable and actionable feedback for improving health literacy practices. By selectively retrieving relevant segments (especially nurse questions), RAG reduces the LLM context window while maintaining accurate health literacy assessment, yielding practical outputs for improved communication.

5 CONCLUSION

This paper proposes *HealthLit*, an AI-driven Audit and Feedback (A&F) system leveraging fine-tuned LLMs (using LoRA) to efficiently assess healthcare providers' adherence to health literacy best practices in clinician-patient interactions. Initial development revealed challenges including limited data, computational constraints, and aligning AI with expert evaluations. We addressed these using prompt engineering (in-context learning, CoT, role-playing), transcript segmentation, and RAG (though this introduced some consistency concerns). *HealthLit* represents a significant step towards scalable, AI-driven A&F systems promoting health literacy best practices. By combining human expertise with AI (especially LLMs), we aim to improve clinician training, patient education, and healthcare communication, ultimately enhancing patient outcomes.

To enhance *HealthLit*'s capabilities and address remaining challenges, future work will prioritize: (1) a hybrid AI-human framework where AI pre-evaluates high-impact segments for expert review, reducing HLE workload while maintaining quality; (2) exploring Knowledge-Augmented Generation (KAG) to improve understanding of health literacy concepts within professional healthcare contexts; and (3) expanding the dataset's size and diversity. While larger LLMs are a consideration, computational resource limitations remain a practical constraint.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Centers for Disease Control and Prevention (CDC). 2024. *Develop and Test Materials: Plain Language.* https://www.cdc.gov/health-literacy/php/develop-materials/plain-language.html

[2] Thi Thuy Ha Dinh, Ann Bonner, Robyn Clark, Joanne Ramsbotham, and Sonia Hines. 2016. 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* 14, 1 (2016), 210–247.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[4] Richard T Griffey, Nicole Shin, Solita Jones, Nnenna Aginam, Maureen Gross, Yonitte Kinsella, Jennifer A Williams, et al. 2015. The impact of teach-back on comprehension of discharge instructions and satisfaction among emergency patients with limited health literacy: A randomized, controlled study. *Journal of communication in healthcare* 8, 1 (2015), 10–21.

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[6] Sylvia J Hysong, Richard G Best, and Jacqueline A Pugh. 2006. Audit and feedback and clinical practice guideline adherence: making feedback actionable. *Implementation science* 1 (2006), 1–10.

[7] Zac E Imel, Brian T Pace, Christina S Soma, Michael Tanana, Tad Hirsch, James Gibson, Panayiotis Georgiou, Shrikanth Narayanan, and David C Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* 56, 2 (2019), 318.

[8] Noah Ivers, Gro Jamtvedt, Signe Flottorp, Jane M Young, Jan Odgaard-Jensen, Simon D French, Mary Ann O'Brien, Marit Johansen, Jeremy Grimshaw, and Andrew D Oxman. 2012. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane database of systematic reviews* 6 (2012).

[9] Noah M Ivers, Anne Sales, Heather Colquhoun, Susan Michie, Robbie Foy, Jill J Francis, and Jeremy M Grimshaw. 2014. No more 'business as usual'with audit and feedback interventions: towards an agenda for a reinvigorated intervention. *Implementation Science* 9 (2014), 1–8.

[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[11] Don Nutbeam. 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health promotion international* 15, 3 (2000), 259–267.

[12] Nico Pronk, Dushanka V Kleinman, Susan F Goekler, Emmeline Ochiai, Carter Blakey, and Karen H Brewer. 2021. Promoting health and well-being in healthy people 2030. *Journal of Public Health Management and Practice* 27, Supplement 6 (2021), S242–S248.

[13] Rabia Shahid, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. 2022. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC health services research* 22, 1 (2022), 1148.

[14] TranscribeMe. 2024. *TranscribeMe: Fast, Accurate, Affordable Transcription Services.* https://www.transcribeme.com/

[15] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670* (2023).

[16] Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* 7, 1 (2024), 41.

[17] Peggy H Yen and A Renee Leasure. 2019. Use and effectiveness of the teach-back method in patient education and health outcomes. *Federal practitioner* 36, 6 (2019), 284.

[18] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. arXiv:2403.13372 [cs.CL] https://arxiv.org/abs/2403.13372