npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02349-3

Enhanced language models for predicting and understanding HIV care disengagement: a case study in Tanzania

Waverly Wei[1,10], Junzhe Shao[2,10], Rita Qiuran Lyu[2,10], Rebecca Hemono[2], Xinwei Ma[3], Joseph Giorgio[4], Zeyu Zheng[5], Feng Ji[6], Xiaoya Zhang[7], Emmanuel Katabaro[8], Matilda Mlowe[8], Amon Sabasaba[8], Caroline Lister[8], Siraji Shabani[9], Prosper Njau[9], Sandra I. McCoy[2] & Jingshen Wang[2] ✉

Sustained engagement in HIV care and adherence to ART are crucial for meeting the UNAIDS "95-95-95" targets. Disengagement from care remains a significant issue, especially in sub-Saharan Africa. Traditional machine learning (ML) models have had moderate success in predicting disengagement, enabling early intervention. We developed an enhanced large language model (LLM) fine-tuned with electronic medical records (EMRs) to predict individuals at risk of disengaging from HIV care in Tanzania. Using 4.8 million EMR records from the National HIV Care and Treatment Program (2018–2023), we identified risks of ART non-adherence, non-suppressed viral load, and loss to follow-up. Our enhanced LLM may outperform traditional machine learning models and zero-shot LLMs. HIV physicians in Tanzania evaluated the model's predictions and justifications, finding 65% alignment with expert assessments, and 92.3% of the aligned cases were considered clinically relevant. This model can support data-driven decisions and may improve patient outcomes and reduce HIV transmission.

The global effort to control the HIV epidemic critically depends on sustained engagement in care and adherence to antiretroviral therapy (ART). The Joint United Nations Programme on HIV/AIDS (UNAIDS) has set ambitious "95-95-95" targets for 2030[1], aiming for 95% of people living with HIV (PLHIV) to know their status, 95% of those diagnosed to receive ART, and 95% of those on ART to achieve viral suppression. Central to these goals is the widespread and effective use of ART, which not only improves the quality of life for PLHIV but also reduces the risk of transmission. Despite significant progress in scaling up ART access, challenges such as ART non-adherence and poor clinical visit adherence persist, particularly in sub-Saharan Africa. For example, studies[2,3] have shown that a substantial proportion of patients discontinue care within the first two years of treatment initiation, clearly indicating the need for interventions to improve retention.

To maximize the impact of HIV care interventions, a shift toward more targeted approaches is warranted, as current evidence indicates that one-size-fits-all interventions may not effectively allocate resources to those most in need. For instance, Fahey et al.[4] found that small, short-term financial incentives improved viral suppression six months after ART initiation by 13 percentage points. However, 73% of the comparison group achieved viral suppression without the intervention, suggesting that more efficient targeting could enhance population-level impacts. In environments with scarce resources, ensuring that programs are efficiently tailored to populations at greatest risk is critical for optimizing outcomes. Evidence like this highlights the importance of identifying individuals most likely to experience poor outcomes and directing limited resources to where they can produce the greatest benefit.

The increasing implementation of electronic medical record (EMR) systems in sub-Saharan Africa, coupled with advancements in machine learning (ML) and artificial intelligence (AI), offers the potential to direct interventions more precisely[5,6]. This is because ML and AI methods can potentially leverage HIV-related EMR data to predict which people face elevated risks of key outcomes by forecasting virologic outcomes, future missed visits, and attrition from care[7]. As a result, healthcare providers can proactively engage at-risk individuals with supportive interventions, enhancing the efficiency of HIV care programs and maximizing the impact of limited resources. Nevertheless, existing studies using ML to predict

[1]Marshall School of Business, University of Southern California, Los Angeles, CA, USA. [2]School of Public Health, University of California, Berkeley, CA, USA. [3]Department of Economics, University of California, San Diego, CA, USA. [4]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA. [5]College of Engineering, University of California, Berkeley, CA, USA. [6]Department of Applied Psychology, University of Toronto, Toronto, ON, Canada. [7]Department of Family, Youth and Community Sciences, University of Florida, Gainesville, FL, USA. [8]Health for a Prosperous Nation, Dar es Salaam, Tanzania. [9]Ministry of Health, Dodoma, Tanzania. [10]These authors contributed equally: Waverly Wei, Junzhe Shao, Rita Qiuran Lyu. ✉e-mail: jingshenwang@berkeley.edu

PLHIV at risk of care disengagement have shown only moderate predictive power[8].

In this study, we develop a novel AI model by enhancing a pre-trained large language model (LLM), specifically Llama 3.1[9]–an open-source pre-trained LLM released by Meta, using routinely collected EMR data from Tanzania's National HIV Care and Treatment Program to predict PLHIV who are at risk of disengaging from care (i.e., becoming ART non-adherent or lost-to-follow-up) or experiencing adverse clinical outcomes (i.e., having an unsuppressed viral load). Since the introduction of GPT-3 in 2020, LLMs have emerged as transformative tools with the potential to revolutionize healthcare due to their ability to process and understand large volumes of textual data and capture complex patterns within it[10,11]. Unlike traditional ML, which often requires structured data and extensive feature engineering, LLMs can leverage unstructured data, natural language inputs, and an extensive training knowledge base, enabling them to uncover nuanced relationships and develop context-specific insights. To our knowledge, this is the first study to develop an enhanced LLM-based AI tool to predict PLHIV at risk of disengagement or developing adverse outcomes using large-scale EMR data collected in sub-Saharan Africa, along with novel insights into individual risk profiles that could inform targeted follow-up interventions.

Results

Table 1 summarizes the multi-class AUC values for predicting the risk of ART non-adherence category, viral suppression, and LTFU using different models on internal (Kagera region 20% subset) and external (Geita region) validation sets. The enhanced LLM fine-tuned with both summarized and natural language descriptions of EMRs consistently achieved improved AUCs. For the risk of ART non-adherence category, it attained AUCs of 0.74 (internal, 95% CI: 0.72–0.76) and 0.67 (external, 95% CI: 0.65–0.69). The enhanced LLM using only summarized data closely followed with AUCs of 0.73 (internal, 95% CI: 0.71–0.75) and 0.66 (external, 95% CI: 0.63–0.69), while the supervised machine learning model yielded AUCs of 0.72 (internal, 95% CI: 0.70–0.74) and 0.66 (external, 95% CI: 0.63–0.69). The zero-shot LLMs performed significantly worse. The zero-shot LLM with summarized and natural language descriptions of EMRs attained AUCs of 0.54 (internal, 95% CI: 0.52–0.56) and 0.54 (external, 95% CI: 0.51–0.57). The zero-shot LLM with summary only attained AUCs of 0.53 (internal, 95% CI: 0.51–0.55) and 0.55 (external, 95% CI: 0.52–0.58). In predicting viral suppression, the enhanced LLM with both data types achieved AUCs of 0.68 (internal, 95% CI: 0.67–0.69) and 0.64 (external, 95% CI: 0.62–0.66), may outperform the summary-only LLM (internal: 0.65, 95% CI: 0.64–0.66; external: 0.62, 95% CI: 0.60–0.64) and the supervised model (internal: 0.68, 95% CI: 0.67–0.69; external: 0.64, 95% CI: 0.62–0.66. The zero-shot LLM with summary only attained AUCs of 0.51 (internal: 95% CI: 0.49–0.53; external: 95% CI: 0.49–0.53) for both validation sets. The zero-shot LLM with both data types attained AUC of 0.51 (internal, 95% CI: 0.48–0.54) and 0.52 (external, 95% CI: 0.50–0.54). For LTFU, the enhanced LLM with combined inputs achieved the highest AUCs of 0.77 (internal, 95% CI: 0.75–0.79) and 0.71 (external, 95% CI: 0.69–0.73). The summary-only LLM followed with AUCs of 0.76 (internal, 95% CI: 0.74–0.78) and 0.68 (external, 95% CI: 0.65–0.71), and the supervised model obtained 0.71 (internal, 95% CI: 0.69–0.73) and 0.65 (external, 95% CI: 0.62–0.68). Zero-shot LLMs lagged with AUCs. The summary-only zero-shot LLM achieved AUCs of 0.55 (internal, 95% CI: 0.52–0.58) and 0.56 (external, 95% CI: 0.54–0.58). The zero-shot LLM with both summary and raw data achieved AUCs of 0.57 (internal, 95% CI: 0.55–0.59) and 0.57 (external, 95% CI: 0.55–0.59). Overall, the enhanced LLMs, especially when fine-tuned with both summarized and natural language descriptions of EMRs, demonstrate improved performance, particularly for the LTFU outcome.

The effort-benefit trade-off curves in Fig. 1 illustrate the cumulative TPR against the proportion of PLHIV predicted to be at the highest risk of LTFU who can be prioritized for proactive support interventions. The enhanced LLM using both summarized and natural language descriptions of EMRs consistently achieved the highest cumulative TPR across all proportions. For instance, when defining 25% of patients identified as will be LTFU, this model achieved TPRs of 78% in internal validation and 73% in external validation. The LLM using only summarized data also performed well, achieving high TPRs (72% and 64% for internal and external validation). In comparison, the supervised ML exhibited lower cumulative TPRs across both internal (46%) and external (39%) validation, and the zero-shot LLM had the lowest TPR values (22% and 21% for internal and external validation).

The subgroup analysis in Fig. 2 shows that the enhanced LLM using both summarized and natural language descriptions of EMRs achieved high AUC across most subgroups, particularly excelling in patients not in care in 2021 and those over 42 years of age, with AUCs up to 0.89 in predicting LTFU. The LLM using summarized data alone also performed well, though slightly lower, achieving similar scores for patients not in care, but with reduced performance for younger age groups. The supervised ML model performed moderately, achieving AUCs up to 0.75 for various age groups, but it struggled with LTFU prediction for patients not in care (AUCs ranging 0.53–0.56). Zero-shot LLMs performed poorly across all subgroups, with AUCs generally below 0.55.

Table 1 | Comparison of model performance AUC (95% CI) on internal and external validation datasets

Internal validation in the Kagera Region			External validation in the Geita Region		
Risk category	Viral suppression	LTFU	Risk category	Viral suppression	LTFU
Enhanced LLM with **summarized** CTC3 medical records					
0.73 (0.71–0.75)	0.65 (0.64–0.66)	0.76 (0.74–0.78)	0.66 (0.63–0.69)	0.62 (0.60–0.64)	0.68 (0.65–0.71)
Enhanced LLM with **summarized and raw** CTC3 medical records					
0.74 (0.72–0.76)	**0.68 (0.67–0.69)**	**0.77 (0.75–0.79)**	**0.67 (0.65–0.69)**	**0.64 (0.62–0.66)**	**0.71 (0.69–0.73)**
Supervised machine learning (gradient boosting)					
0.72 (0.70–0.74)	0.68 (0.67–0.69)	0.71 (0.69–0.73)	0.66 (0.63–0.69)	0.64 (0.62–0.66)	0.65 (0.62–0.68)
Zero-shot LLM with summarized CTC3 medical records					
0.53 (0.51–0.55)	0.51 (0.49–0.53)	0.55 (0.52–0.58)	0.55 (0.52–0.58)	0.51 (0.49–0.53)	0.56 (0.54–0.58)
Zero-shot LLM with summarized and raw CTC3 medical records					
0.54 (0.52–0.56)	0.51 (0.48–0.54)	0.57 (0.55–0.59)	0.54 (0.51–0.57)	0.52 (0.50–0.54)	0.57 (0.55–0.59)

The model demonstrating the overall best performance is highlighted in bold.

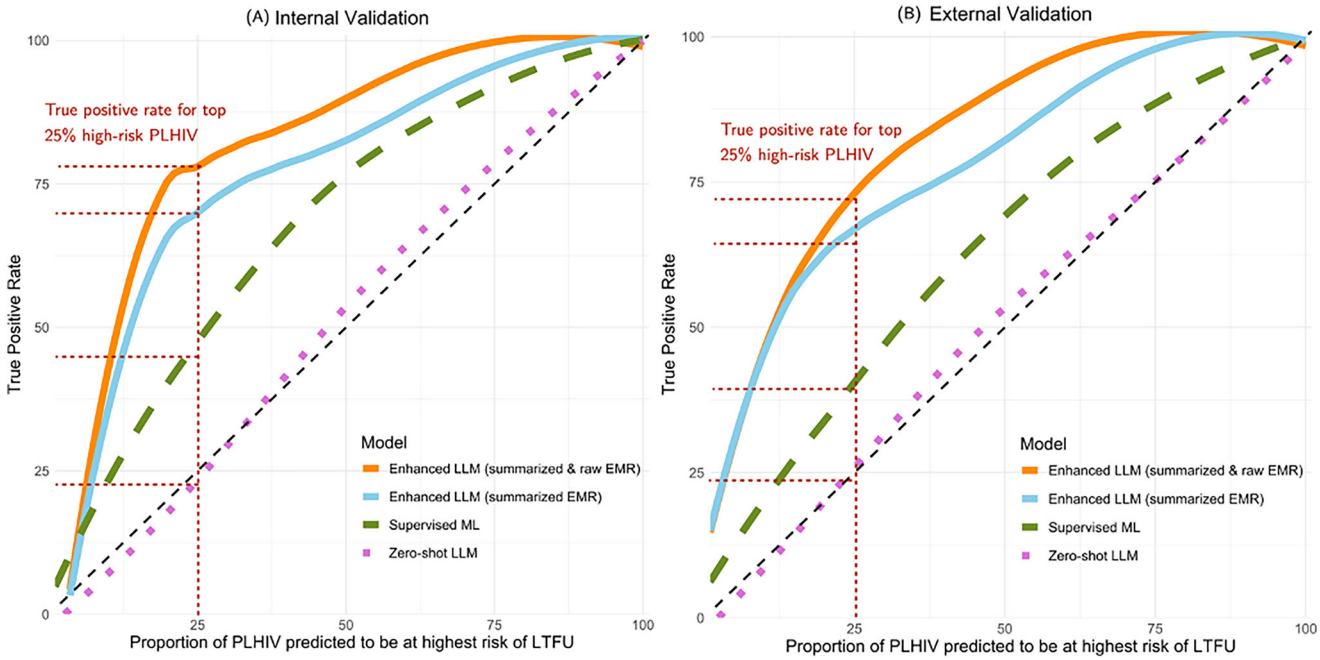

Fig. 1 | Effort-benefit trade-off curves. The curves describe the cumulative TPRs (the proportion of correctly identified at-risk patients) against the proportion of PLHIV predicted to be at highest risk of LTFU who can be prioritized for proactive support interventions within resource constraints. **A** presents the AUC curve for the internal validation set, while (**B**) presents the AUC curve for the external validation set.

Figure 3 presents the attention scores of enhanced LLMs and feature importance for the supervised ML (gradient boosting) model. For the enhanced LLM using summarized EMRs (Panel A), the highest attention scores were given to the keywords "sex", "age", "gaps in follow-up care," "delayed appointments," "missed appointments", and "ART adherence," suggesting that these factors are most critical for predicting defined HIV risk outcomes. In the enhanced LLM using both summarized and raw EMR data (Panel B), similar variables were highlighted, with the "summary" feature receiving the highest attention, indicating that the summarized information played a crucial role in model prediction. Additionally, keywords like "# days since HIV" and "WHO stage" also received high attention scores, emphasizing their importance for accurate prediction. The supervised ML model (Panel C) showed the highest variable importance for "age" and "# pills," followed by "# days since HIV diagnosis" and "weight."

Lastly, the agreement between human and LLM decisions—defined as both arriving at the same prediction regarding patient LTFU status—was observed in 65% of cases (13 out of 20). When aligned, Tanzanian HIV care providers found the AI-provided justifications to be reasonable, with 92.3% being rated as sensible. In cases where physician and AI decisions were misaligned (7 out of 20), the AI provided correct answers 57% of the time (4 out of 7), while humans provided correct answers in 43% of those cases (3 out of 7). However, in instances of misalignment, experts judged the LLM's justifications as not helpful.

Discussion

We developed and validated two enhanced LLMs based on Llama 3.1, fine-tuned with EMRs from Tanzania's National HIV Care and Treatment Program, to predict PLHIV at risk of ART non-adherence, LTFU, and developing non-suppressed viral load. Our findings demonstrate that the enhanced LLMs, particularly when utilizing both summarized and natural language descriptions of EMR data, may perform better than traditional supervised ML models and zero-shot LLM in all outcomes. Our enhanced LLM not only achieved higher AUC values but also maintained robust performance across various subgroups, suggesting broad applicability in diverse patient populations.

Compared to previous studies using ML to predict HIV care disengagement, our enhanced LLMs demonstrated both superior predictive performance and the ability to address multiple clinically meaningful HIV outcomes within a single framework. For example, prior ML-based approaches have typically focused on a single outcome. Esra et al.[12] developed an ML model in South Africa to predict subsequent missed ART visits, achieving a sensitivity of 61.9% and a PPV of 19.7%, while Maskew et al.[13] applied ML algorithms in two South African districts to predict the risk of LFTU, reporting an AUC of 0.68. Similarly, Xie et al.[14] leveraged EMR data from two regions in Tanzania to forecast LTFU, achieving an AUC of 0.65. When we compare the performance on the same outcome of LTFU, our LLMs achieved more powerful predictive performance, with AUCs of 0.77 for internal validation and 0.71 for external validation, surpassing the results reported by both Maskew et al.[13] (AUC = 0.68) and Xie et al.[14] (AUC = 0.65). This combination of improved predictive power and a broader scope of clinically relevant outcomes demonstrates the unique potential of LLM-based methods to advance HIV care.

Our findings have potential implications for HIV care programs in sub-Saharan Africa. By accurately identifying patients at higher risk of disengagement or adverse outcomes, healthcare providers can allocate resources more efficiently and implement targeted interventions to groups who are most in need. For example, when focusing on the top 25% of patients identified by our model as became LTFU, we found that 78% of these people in internal validation and 73% in external validation were genuinely LTFU (Fig. 2). This high level of accuracy enables healthcare teams to concentrate their efforts on a smaller subset of PLHIV who are most likely to benefit from additional support, thereby maximizing the impact of limited resources. In addition, our enhanced LLM works in a manner akin to a clinician by reading patient summaries, providing risk predictions, and offering reasoning for its assessments. In the future, such AI models could be integrated into clinical workflows to collaborate side by side with human experts, augmenting their ability to monitor patient progress and intervene promptly when necessary. This human-AI collaboration not only facilitates more informed clinical evaluations and decision-making but also aligns with the UNAIDS "95-95-95" targets, potentially contributing to reduced HIV transmission and improved health outcomes for PLHIV.

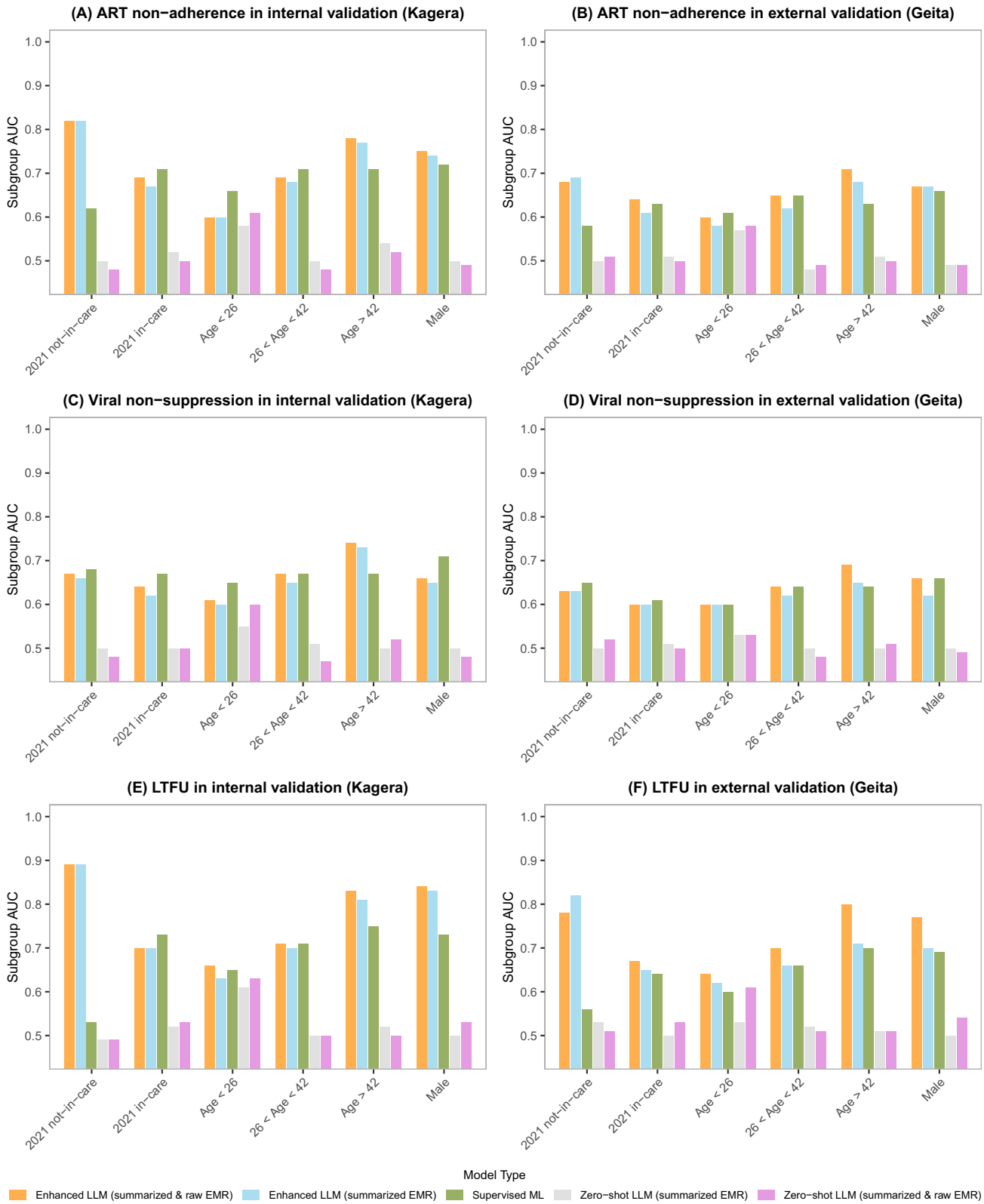

Fig. 2 | Subgroup analysis of AUC comparisons across different in-care, age, and sex groups. A, **C**, **E** show the AUCs for predicting ART non-adherence, viral non-suppression, and LTFU in the internal validation set. **B**, **D**, **F** show the subgroup AUCs in the external validation set.

An innovative technical aspect of our enhanced LLM is the use of textual summaries of longitudinal EMRs, which not only improved the model's predictive performance but also provided a valuable tool for HIV clinicians to monitor patient health and to provide timely interventions. Traditionally, clinicians must sift through large volumes of patient medical records, a task that is both time-consuming and prone to oversight. By providing interpretable summaries, our model helps physicians quickly identify critical changes and emerging health issues, effectively augmenting

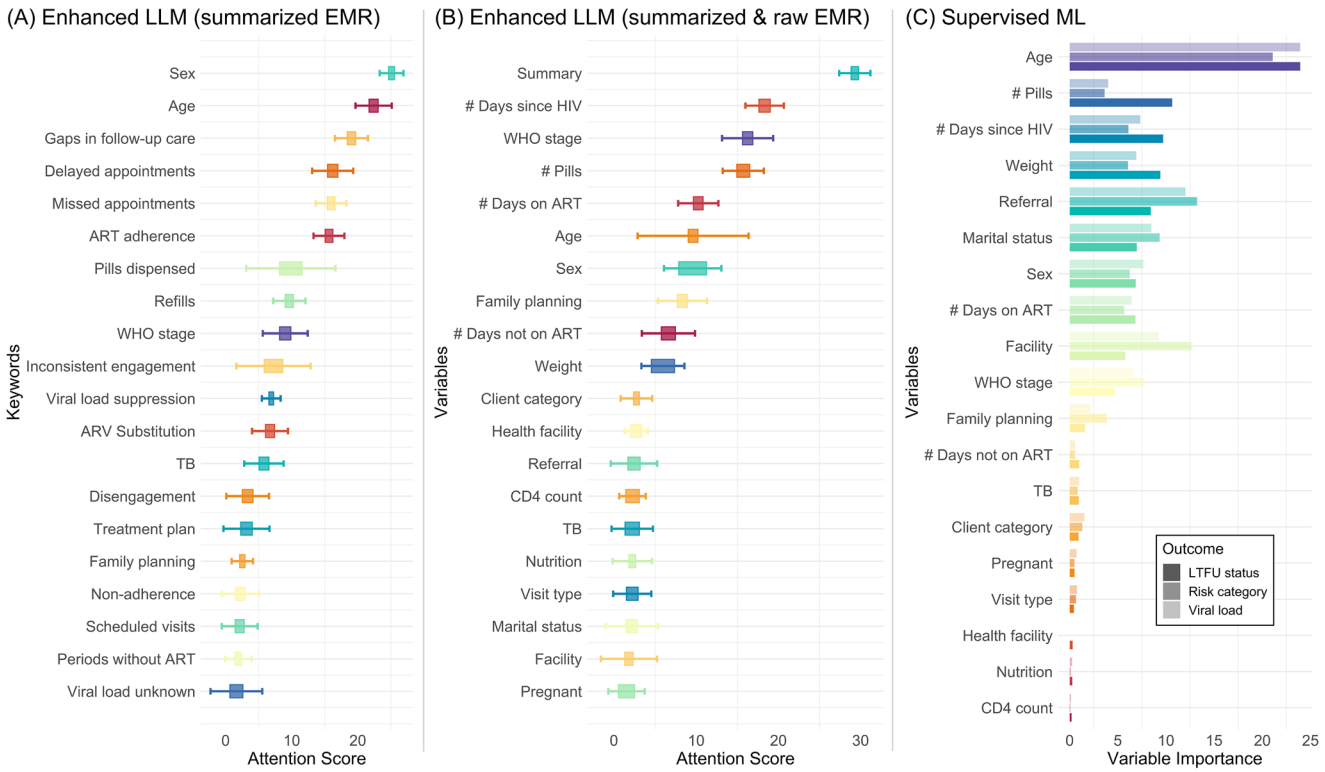

Fig. 3 | Attention scores of enhanced LLMs and feature importance for supervised ML. A displays the attention scores for the summarized EMR model. **B** displays the attention scores for the summarized and raw EMR model. **C** displays the attention scores for the supervised machine learning model.

their capacity to manage complex individual medical histories. Additionally, whereas conventional ML approaches may appear as "black boxes" that are difficult to interpret, our LLMs provide the underlying reasoning and temporal patterns captured by the model (Supplementary Table 1). Compared to LLMs, the ML model relied more on demographic and treatment duration features (Fig. 3C). The enhanced LLMs, especially with the summarized feature included, appeared to focus more on care continuity and adherence-related aspects, highlighting their nuanced understanding of patient engagement in HIV care (Fig. 3A). The LLM's dual focus on interpretability and clinical relevance enables timely interventions and supports more informed decision-making. Independent evaluations by clinical experts confirm that these summaries are both clinically accurate and highly actionable, illustrating the potential of this technique to advance HIV care and other chronic conditions requiring sustained engagement.

Our results suggest that LTFU, as defined by a 28-day or greater disruption in care, may be the most actionable outcome for identifying higher-risk individuals who can be reached with personalized HIV interventions, as it achieved the highest predictive performance across all models under comparison. Recall that the enhanced LLM reached an AUC up to 0.77 in internal validation and 0.71 in external validation for LTFU prediction, surpassing its performance on other outcomes like risk category and viral non-suppression. The ability to accurately predict patients who will be LTFU is valuable, as it enables healthcare providers to implement timely interventions aimed at retaining patients in care, thereby improving adherence to ART before the patient experiences viral rebound, thereby reducing the risk of onward transmission.

Despite these promising results, there are limitations to consider. While the enhanced LLMs outperformed traditional ML models, there remains room for models' further improvement. Future research could explore integrating additional data sources, such as social determinants of health or patient-reported outcomes, to enhance model performance. Additionally, our study utilized data from two regions in Tanzania, and although we conducted external validation, the generalizability to other settings may be limited. Evaluating the models in different geographic and healthcare contexts is necessary to confirm their broader applicability. Currently, the model is hosted locally at UC Berkeley in a secure data environment, ensuring robust data privacy protections. Looking ahead, our private LLM environment at Berkeley provides a secure platform for expanding these capabilities to other parts of sub-Saharan Africa. This effort will require robust data-sharing agreements, strict adherence to privacy regulations, and strong encryption measures to safeguard patient information. The data utilized in this analysis was obtained through a formal Data Transfer and Use Agreement established between the Tanzania Ministry of Health and UC Berkeley. This agreement explicitly permits only the transfer of de-identified HIV Care and Treatment data, ensuring that explicitly identifiable patient information, such as patient names, is fully removed prior to transfer, thus preventing direct patient identification without consent. Data access at UC Berkeley is strictly limited to authorized researchers bound by confidentiality obligations specified in the agreement. The data was transferred securely via a SharePoint folder approved for P4-level data handling, meeting international security standards and the requirements of HIPAA and Tanzanian privacy laws, thus ensuring patient confidentiality and regulatory compliance.

Another limitation of our work is that more rigorous de-identification procedures might be warranted to further minimize re-identification risk. As kindly suggested by our reviewers, we have implemented and tested an enhanced de-identification procedure that randomly shifts each participant's visit dates by a few days while maintaining internal consistency (see results in Supplementary Table 5). Beyond date shifting, additional privacy-preserving steps—such as removing free-text fields, replacing exact dates of birth with month–year information, and removing all original unique identifiers—could further strengthen data protection without compromising analytic utility. We shall keep exploring robust de-identification approaches in future applications of our model to further strengthen patient privacy protection.

Lastly, as kindly suggested by our reviewer, we have reported the PRAUC results of the methods in comparison (Supplementary Table 6) and the prevalence of the rare events classes (Supplementary Table 7). Our

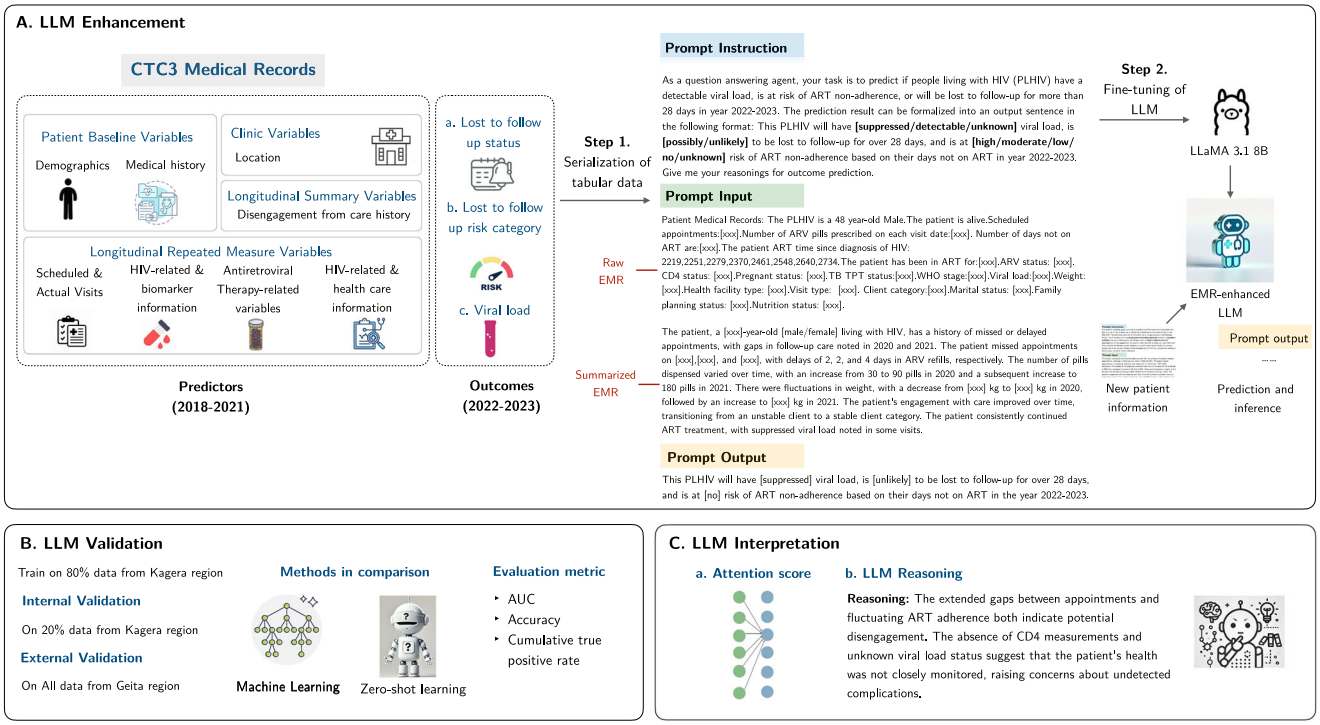

Fig. 4 | Method for open-source language model (LLaMA 3.1) enhancement pipeline. A illustrates the procedure for enhancing the LLM with CTC3 medical records, including the types of variables in the CTC3 dataset, the data serialization process, and the prompt design. **B** depicts the validation workflow, comprising internal validation using data from the Kagera region and external validation using data from the Geita region. **C** presents two approaches for interpreting LLM predictions: attention score analysis and LLM reasoning extraction. Specific words are highlighted in red to illustrate how standardization of language and format reduces model hallucination and improves prediction consistency.

enhanced LLM using summarized and raw CTC3 records achieves a higher PRAUC than the machine learning method, indicating that it is more effective at identifying high-risk individuals under class imbalance. This improvement is practically meaningful because PRAUC captures the model's ability to prioritize true positives while limiting false positives, an important feature in the setting where outreach resources are limited. Therefore, a higher PRAUC of our method may translate into more efficient allocation of follow-up resources. However, the benefit of improved PRAUC must be weighed against the trade-offs inherent to LLM deployment. The enhanced LLM model may introduce additional complexity in implementation and ongoing maintenance. Thus, while the improved PRAUC demonstrates the promise of LLM-augmented prediction, the choice between an LLM model and machine learning models should consider the trade-off between predictive gains and the practical cost.

In conclusion, this study demonstrates the potential of enhanced LLMs in predicting the risk profiles of PLHIV using large-scale EMR data and in supporting clinicians through interpretable summaries. The improved predictive performance over traditional ML models suggests that advanced AI approaches can play a pivotal role in optimizing future HIV care. Future research could focus on refining these models, integrating them into clinical workflows, and assessing their impact on patient outcomes.

Methods

Methods for developing the enhanced LLM; Fig. 4 for method overview

Data source and ethical statement. We utilized data from two regions in Tanzania from the national EMR repository, which includes digitized medical records manually entered after clinical visits and maintained by the Ministry of Health (MoH) in Tanzania and the National AIDS, STIs, and Hepatitis Control Programme (NASHCoP), which are referred to as CTC3 medical records. Our analysis adopted the EMRs from January 1, 2018, to June 30, 2023, resulting in 4,809,765 records for 261,192 PLHIV. These medical records include 99 variables related to demographics, clinical visit details, biomarkers and lab tests, medical history, and health facility information (Supplementary Table 4).

Ethical approval. The Tanzania National Institute for Medical Research (NIMR) and the Committee for Protection of Human Subjects at the University of California, Berkeley (UC Berkeley) provided ethical review and approval for this study (UC Berkeley ethics approval number: 2021-10-14723, renewed on October 23, 2024; [Tanzania NIMR ethics approval number: NIMR/HQ/R.8a/Vol.X/3992, obtained on 5 May 2022.]) The study protocol titled, "Strengthening the continuity of HIV care in Tanzania with economic support" received expedited review and approval. With approval from both ethical review boards, we received a waiver of written informed consent for this part of the study in accordance with FDA guidelines, according to 45 CFR 46.116(f). Our research also complies with Tanzania's Personal Data Protection Act (PDPA).

Specific permission was granted to use routinely collected demographic and clinical information, as well as the minimal amount of identifiable private information to identify participants for potential future trials. These EMRs are used to enhance the pre-trained open-source LLM, Llama 3.1 (an 8-billion-parameter version of Meta's pre-trained open-source large language model), which has been securely downloaded to our GPU cluster. The cluster is housed within UC Berkeley's campus data center, with full encryption implemented to ensure the highest level of data protection. To protect the confidentiality of the data, we use a locally hosted GPU cluster securely stored at the UC Berkeley data center for this analysis. This means that the input data remains entirely within our own secured computing environment. At no point is the data uploaded, shared, or otherwise accessible to Meta or any other external provider or developer of LLMs. All analysis occurs strictly within our local infrastructure, ensuring complete confidentiality.

Table 2 | Predictors and their serialization for LLM fine-tuning

Description		Serialization for LLM fine-tuning
Demographics		
Sex	Male/Female	The patient is a [37] year-old [female].
Age	Age at recruitment	
Marital status	Record of marital status, including changes over time	Marital status: [divorced, divorced, divorced, divorced].
Clinical visits details		
Date of scheduled appointment	Historical records of the patient's scheduled appointment dates	Scheduled appointments: [NA,2020-04-03,2020-05-04,2020-06-04].
Visit date	Date of patient's clinic visit	Actual visit dates: [2020-03-02,2020-04-03,2020-05-05,2020-06-04].
Visit type	Type of clinic visit	Visit type: [Unscheduled visit at this clinic, Scheduled Visit at this clinic, Scheduled Visit at this clinic, Scheduled Visit at this clinic].
Client category	Client type	Client category: [Unstable Client, NA, Stable Client, Unstable Client].
Medical history		
Number of pills prescribed	History of the number of pills prescribed	Number of ARV pills prescribed on each visit date: [30,30,30,30].
Number of days since first ART treatment	Number of days the patient has been on ART treatment	The patient has been on ART for [0,32,64,94 days].
Number of days since HIV diagnosis	Number of days since the patient diagnosed with HIV	The number of days since HIV diagnosis: [2131,2163,2195,2225] days.
TB preventive therapy	The status of the tuberculosis preventive therapy	TB TPT status: [start using 3HP TB preventive treatment, start using INH (Isoniazid) and IPT (Isoniazid Preventive Therapy), start using INH (Isoniazid) and IPT (Isoniazid Preventive Therapy), start using INH (Isoniazid) and IPT (Isoniazid Preventive Therapy)].
Biomarkers		
ART status	Status of the antiretroviral therapy	ART status: [continue ART, continue ART, continue ART, continue ART].
CD4 count	History of CD4 count measurements	CD4 status: [No CD4 measured, No CD4 measured, No CD4 measured, No CD4 measured].
Nutrition status	History of nutrition status	Nutrition status: [Not malnourished, Not malnourished, Not malnourished, Not malnourished].
Pregnant status	Whether the patient is pregnant	Pregnant status: [not pregnant, not pregnant, not pregnant, not pregnant].
WHO stage	History of WHO stage	WHO stage: [WHO stage is stage 1 (asymptomatic), WHO stage is stage 1 (asymptomatic), WHO stage is stage 1 (asymptomatic), WHO stage is stage 1 (asymptomatic)].
Viral load history	History of viral load status	Viral load: [suppressed, unknown, unknown, unknown].
Weight	Measurements of the patient's weight	Weight: [68,66,66,65].
Family planning	Family planning methods, including changes in methods	Family planning status: [NOT USING, NA, NOT USING, NA].
Health facility information		
Health facility type	Type of health facility attended by the patient	Health facility type: [Health Center].

Table 3 | Composite outcomes and their textual serialization for LLM fine-tuning

Outcomes	Levels	Serialization for LLM fine-tuning
Risk profile	High/moderate/low/no risk	This PLHIV will have [detectable] viral load, is [possibly] to be lost to follow-up for over 28 days, and is at [high] risk of ART non-adherence based on their days not on ART in the year 2022–2023.
High viremia	Detectable/suppressed/unknown	
Ever lost-to-follow-up	Possibly/unlikely	

Predictors. To ensure the practicality of our enhanced LLM model, we employed a bi-directional, participatory approach to predictor selection—this collaborative process involved close coordination with local clinicians and MoH representatives, drawing on CTC3 medical records. The selected predictors listed in Table 2 were categorized into six groups: (1) Demographics (sex, age, marital status), (2) Clinical visit details (appointment date, visit date, visit type, client category), (3) Medical history (number of pills, duration on ART, time since HIV diagnosis, tuberculosis preventive therapy), (4) Biomarkers (antiretroviral treatment status, CD4 count, nutrition status, pregnancy status, WHO stage, viral load measurements, weight, family planning), and (5) Health facility information (facility type). To ensure relevance to current practice, we used medical records from January 1, 2018, to December 31, 2021, when constructing the predictor set (Supplementary Table 3).

Outcomes. Aiming to guide the targeted implementation of proactive interventions for PLHIV to improve retention in care, reduce morbidity, and prevent onward transmission, we define a comprehensive composite outcome using data from January 1, 2022, to June 30, 2023 (Table 3). This composite outcome includes three components: (1) the risk of ART non-adherence, determined by the number of days not on ART (details below); (2) high viremia, defined as a viral load exceeding 1000 copies/mL at least once; and (3) lost-to-follow-up (LTFU), defined as missing a scheduled clinical visit by 28 days or more (following the PEPFAR[15] definition of

LTFU from facility-based care). The ART non-adherence risk profile is calculated based on the duration a patient was off ART during the study period. Using data from the CTC3 medical records, which include scheduled visit dates, actual visit dates, and the quantity of ARV pills dispensed at each clinical visit, we estimated the number of days each patient was off ART from January 1, 2022, to June 30, 2023. Patients off ART for fewer than 28 days are classified as "no risk," aligning with the PEPFAR definition. Those off ART for more than 28 days are categorized into three risk groups based on the distribution of the number of days off ART: low risk (28 to less than 33 days, below the 25th percentile), moderate risk (33 to less than 44.5 days, between the 25th and 50th percentiles), and high risk (44.5 days or more, above the 50th percentile).

Fixed formatting prompt instruction. Fixed formatting and task-specific instructions are applied to facilitate downstream validation. The LLM was instructed as follows:

"*As a question-answering agent, your task is to predict if people living with HIV (PLHIV) have a detectable viral load, is at risk of ART non-adherence, or will be lost to follow-up for more than 28 days in year 2022–2023. The prediction result can be formalized into an output sentence in the following format: This PLHIV will have [suppressed/detectable/unknown] viral load, is [possibly/unlikely] to be lost to follow-up for over 28 days, and is at [high/moderate/low/no/unknown] risk of ART non-adherence based on their days not on ART in the year 2022–2023. Give me your reasoning for outcome prediction.*"

This instruction guides the LLM in generating outcomes in a standardized format, using predefined language to ensure accuracy and focus. This approach also allows us to extract token probabilities for specific components of predictions, which enables the application of classical multi-class classification metrics to validate model performance.

Natural language description of predictors and outcomes in the prompt input and outcome. To adapt pre-trained LLMs for processing longitudinal EMRs, we performed feature engineering to convert predictors and outcomes into natural language—a process known as textual serialization[16–19]. This conversion makes the data compatible with LLMs, which are inherently designed to process textual information. Table 2 details the serialization process for the predictors extracted from the CTC3 medical records. Our serialization approach varied depending on the type of variable. Some variables, such as demographics (e.g., sex, age), were converted into natural language formats. In contrast, others, like clinical visits and biomarker data, were serialized as structured text sequences (e.g., dates and weight measurements). Baseline demographics and health facility information were serialized into interpretable text-based formats, enabling the LLM to process tabular data effectively. Longitudinal variables, such as clinical visit data and biomarkers, were transformed into narrative sequences that preserve the temporal dynamics essential for modeling patient health trajectories. Missing data was represented as "NA," allowing the LLM to interpret the absence of information without requiring imputation. See Table 3 for details following the prompt instruction.

Utilizing summarized longitudinal predictors. To capture temporal trends in longitudinal EMRs, we instructed Llama 3.1 to generate narrative summaries of individual patient predictors with the following prompt: "*Please review the patient's medical records from 2018 to 2021. For example, the summary could include how patient profiles change over the years, periods of missed or delayed appointments, changes or gaps in follow-up care related to ART adherence, patterns in the number of pills dispensed during this time, and fluctuations in CD4 count, viral load, or weight.*"

Enhance pre-trained LLM with fine-tuning. To improve the predictive accuracy of a pre-trained LLM in predicting comprehensive outcomes for PLHIV, we fine-tuned Llama 3.1 using fixed formatting instructions and natural language descriptions of predictors and outcomes as input training prompts. Fine-tuning involves adapting a pre-trained LLM with a specialized dataset, enhancing its ability to perform specific tasks while minimizing computational demands. In this study, we used Low-Rank Adaptation (LoRA)[20], an efficient method for fine-tuning that optimizes only a subset of model parameters. We developed two fine-tuned LLMs using the same LoRA-based approach, with differences in the structure and focus of their input data. Both models were trained with a fixed-format prompt instruction, guiding the LLM to predict the composite outcome comprising three components (the risk of ART non-adherence, non-suppressed viral load, and LTFU). The first model utilized only the generated longitudinal summaries of patient data as inputs. In contrast, the second model incorporated both the longitudinal summaries and detailed natural language descriptions of the predictors. While the underlying prediction task remained consistent across both models, the second model was designed to leverage the richer contextual information embedded in the expanded input structure, enabling a more nuanced assessment of patient outcomes.

Methods for evaluating, validating, and interpreting the enhanced LLMs

External and internal validation datasets. The enhanced LLMs underwent both internal and external validation. For internal validation, 80% of the CTC3 medical records from Kagera region from January 1, 2022, to June 30, 2023 were randomly selected for model development, with the remaining 20%, which were completely withheld during training, used for validation. The model trained on 80% of CTC3 data from Kagera region was then externally validated on CTC3 data from Geita region. Summary distributions of predictors in both the training and validation datasets are provided in Table S3.

Evaluation metrics. We compared the performance of our enhanced LLM with a state-of-art ML method[8,14] and zero-shot Llama 3.1. For the supervised ML method, we employ the gradient boosting method, an ensemble learning method known to have robust performance in prediction tasks. Specifically, we utilize the *CatBoost*[21] Python package that implements the gradient boosting method while optimizing computational efficiency. For the zero-shot method, we instructed Llama 3.1 to make predictions directly based on natural language descriptions of patient predictors and outcomes.

Our prompt instructions specified a fixed format for prompt outputs, facilitating easy extraction of token probabilities and enabling us to apply classical methods to evaluate model performance. We assessed the performance of various models using three metrics: the classical Area Under the Receiver Operating Characteristic Curve (AUC), effort-benefit trade-off curves, and subgroup analyses (Supplementary Table 2). Receiver operating characteristic (ROC) curves were plotted to visualize the trade-off between true positive rates (TPR, sensitivity) and false positive rates (1 – specificity) at various thresholds. For the multi-class outcome variable, we reported multi-class AUC[22] using a one-vs-rest approach. Effort-benefit trade-off curves plot the cumulative TPRs, representing the proportion of correctly identified at-risk patients, against the proportion of PLHIV predicted to be at the highest risk of LTFU—defined as those with calculated risk scores exceeding a given threshold—who can be prioritized for proactive support interventions under resource constraints. Lastly, subgroup analyses examined how model performance varied across different patient groups as defined by age and sex.

Attention scores for model interpretation. Attention scores are weights in a language model assigned to each token in the input sequence, indicating the relative importance of each token when predicting the output[23,24]. Higher attention scores often imply that a group of tokens is more influential for the model's prediction. See technical details for attention score extraction in supplements.

Human expert evaluation. Two local HIV physician researchers in Tanzania evaluated the prediction results of the enhanced LLM for 20 patient medical records. They independently reviewed the LLM-summarized longitudinal patient medical records (a synthetic summary provided in the Supplementary Materials to ensure patient privacy) to predict the patient LTFU status. Additionally, they were provided with the enhanced LLM's justifications and assessed whether the reasoning was reasonable and aligned with the information in the reviewed medical records.

Data availability

The electronic medical records data from Tanzania's National HIV Care and Treatment Program used in this study contains protected health information and, therefore, cannot be shared publicly.

Received: 25 February 2025; Accepted: 7 January 2026;
Published online: 21 January 2026

References

1. Frescura, L. et al. Achieving the 95 95 95 targets for all: a pathway to ending AIDS. *PLoS ONE* **17**, e0272405 (2022).
2. Plazy, M., Orne-Gliemann, J., Dabis, F. & Dray-Spira, R. Retention in care prior to antiretroviral treatment eligibility in sub-Saharan Africa: a systematic review of the literature. *BMJ Open* **5**, e006927 (2015).
3. Penn, A. W. et al. Supportive interventions to improve retention on ART in people with HIV in low- and middle-income countries: a systematic review. *PLOS ONE* **13**, e0208814 (2018).
4. Fahey, C. A. et al. Financial incentives to promote retention in care and viral suppression in adults with HIV initiating antiretroviral therapy in Tanzania: a three-arm randomised controlled trial. *Lancet HIV* **7**, e762–e771 (2020).
5. Olatosi, B., Vermund, S. H. & Li, X. Power of Big Data in ending HIV. *AIDS* **35**, S1–S5 (2021).
6. Akanbi, M. O. et al. Use of electronic health records in sub-Saharan Africa: progress and challenges. *J. Med Trop.* **14**, 1–6 (2012).
7. Audere & Desmond Tutu Health Foundation. Artificial intelligence to enhance HIV prevention in age ofdisruptions: Recommendations from the Audere and Desmond Tutu Health Foundation expert consultation. Audere Africa (2025).
8. Fahey, C. A. et al. Machine learning with routine electronic medical record data to identify people at high risk of disengagement from HIV care in Tanzania. *PLOS Glob. Public Health* **2**, e0000720 (2022).
9. Meta AI. The Llama 3 herd of models. https://ai.meta.com/research/publications/the-llama-3-herdof-models/ (2024).
10. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. https://doi.org/10.1001/jama.2024.21700. (2024).
11. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
12. Esra, R. T. et al. Historical visit attendance as predictor of treatment interruption in South African HIV patients: extension of a validated machine learning model. *PLOS Glob. Public Health* **3**, e0002105 (2023).
13. Maskew, M. et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. *Sci. Rep.* **12**, 12715 (2022).
14. Xie, Z. et al. Prevention of adverse HIV treatment outcomes: machine learning to enable proactive support of people at risk of HIV care disengagement in Tanzania. *BMJ Open* **14**, e088782 (2024).
15. PEPFAR Fiscal Year 2023 Monitoring, Evaluation, and Reporting (MER) Indicators. United States Department of State. https://www.state.gov/pepfar-fy-2023-mer-indicators/ (accessed Dec 20, 2024).
16. Fang, X. et al. Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding -- A Survey. https://doi.org/10.48550/arXiv.2402.17944. (2024).
17. Yu, B., Fu, C., Yu, H., Huang, F. & Li Y. Unified language representation for question answering over text, tables, andimages. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4756–4765, https://doi.org/10.18653/v1/2023.findings-acl.292 (Toronto, Canada, 2023).
18. Gong, H. et al. TableGPT: Few-shot Table-to-Text Generation with Table Structure Reconstruction and Content Matching. In *Proc. 28th International Conference on Computational Linguistics. Barcelona, Spain (Online)* (eds Scott D., Bel N., Zong C.) 1978–1988 (International Committee on Computational Linguistics, 2020).
19. Hegselmann, S. et al. TabLLM: few-shot classification of tabular data with large language models. In *Proc. 26th International Conference on Artificial Intelligence and Statistics* (PMLR, 2023).
20. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. https://doi.org/10.48550/arXiv.2106.09685. (2021).
21. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. In *Proc. 32nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2018).
22. Hand, D. J. & Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **45**, 171–186 (2001).
23. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing System*. 6000–6010 (Curran Associates Inc., 2017).
24. Yeh, C. et al. AttentionViz: a global view of transformer attention. In *IEEE Trans Visual Comput Graphics* (IEEE, 2023).

Acknowledgements

The study was supported by a grant from the US National Institutes of Health (NIH): NIH 1R01MH125746. We thank the review team for their valuable suggestions and comments, which significantly improved our paper.

Author contributions

W.W., J.S., and R.Q.L. conducted the model training and data analysis and organized the study results, including figures and tables. X.M., J.G., Z.Z., X.Z., R.H., and F.J. provided guidance on methodology development. E.K., M.M., A.S., C.L., S.S., P.N., and S.I.M. facilitated data access, offered administrative support, and contributed domain knowledge and insights during manuscript development. J.W. drafted the manuscript, guided the study design, and supervised the project. All authors contributed to study results interpretation and major revision of the manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02349-3.

Correspondence and requests for materials should be addressed to Jingshen Wang.