Submitted to the Annals of Applied Statistics

FEATURE IMPORTANCE AND SELECTION IN LANGUAGE MODELS WITH SEMI-STRUCTURED DATA

BY QIURAN LYU [1,a], XINWEI MA[2,c] AND JINGSHEN WANG[1,b]

[1]*Division of Biostatistics, University of California, Berkeley. ,* [a]*qiuran_lyu@berkeley.edu;* [b]*jingshenwang@berkeley.edu*

[2]*Department of Economics, University of California San Diego,* [c]*x1ma@ucsd.edu*

Large language models (LLMs) are increasingly popular in biomedical prediction tasks, taking as input long serialized prompts that combine heterogeneous data such as demographics, laboratory tests, imaging, genetic profiles, and longitudinal electronic health records. However, it remains unclear which variable groups are relevant for accurate LLM-based predictions. We propose to perform selection in the LLM embedding space: unselected variable groups are replaced by padding-token blocks rather than removed, preserving a fixed input dimension. Combined with a Bernoulli relaxation of the discrete selection problem, our framework enables gradient-based optimization with a relaxed ℓ_0 penalty, thereby circumventing the combinatorial intractability of naive token deletion in the prompt space. We further offer theoretical guarantees on the generalization risk, and demonstrate that sparser selections yield tighter risk bounds. We apply the method to domain-fine-tuned LLMs for two prediction tasks, Alzheimer's disease (AD) pathology and HIV treatment outcomes, and identify sparse, clinically meaningful variable groups. The selected subsets substantially reduce prompt length while maintaining, and in many cases improving, predictive accuracy. In the AD study, selected subsets prioritize genetics and MRI-derived neurodegeneration markers over cognitive diagnosis variables; in the large-scale HIV electronic medical record study from the Tanzanian national CTC3 repository, they prioritize longitudinal treatment adherence and healthcare-engagement variables associated with future disengagement from care. Across both applications, the proposed selections diverge from classical machine learning feature-importance rankings, suggesting that prompt serialization and pretrained language representations reshape feature importance during LLM inference, motivating LLM-aware feature-selection strategies for healthcare deployment and data prioritization.

1. Introduction Large language models (LLMs) are increasingly applied as predictive engines in biomedical settings (Jin et al., 2024; Wang et al., 2025; Wysocki et al., 2024; Naved et al., 2025), where inputs can be naturally organized into structured groups such as demographics, vital signs, laboratory values, medical images, medications, and clinical notes (Hegselmann et al., 2023; Zhu et al., 2024; Wei et al., 2026). Rather than operating on fixed tabular matrices, these heterogeneous inputs are often serialized into textual prompts, enabling the combination of diverse data sources across cohorts and health systems.

In such prompts, however, it is rarely clear *which* variable groups are actually necessary for accurate prediction and which can be safely omitted. Clinical deployments, in particular, require answers to the following questions: *Which parts of electronic health records (EHRs) drive an LLM's predictions? Can one reduce input modalities without degrading LLM performance? Is there scope to shorten prompts to fit tighter context windows and lower computational cost?* (Mienye et al., 2024; Maity and Saikia, 2025; Lee, Wu and Chiang, 2025; Ganjdanesh et al., 2024). Despite the rapid adoption of LLMs for biomedical prediction,

Keywords and phrases: Feature selection in language models, Embedding-level masking, Discrete-continuous surrogate optimization, Alzheimer's disease pathology, HIV treatment engagement and adherence.

there has been no systematic framework for group-wise feature selection at the level of serialized prompt segments, nor for quantifying feature importance aligned with the LLM's own inference mechanism and pretrained knowledge.

Feature selection in language model inference is fundamentally more challenging than in classical tabular settings. LLMs do not take covariates as columns in a fixed design matrix; instead, each variable group is first serialized into text, then mapped through tokenization, one-hot encoding, and embedding into a variable-length block of token vectors (Vaswani, 2017). Dropping a feature group, therefore, changes *both* the semantic content and the *dimension* of the resulting embedding matrix, and subsequent self-attention layers further entangle information across groups. These operations make the prediction loss, as a function of variable group inclusion and selection, fundamentally nonsmooth. Moreover, even with a Bernoulli relaxation of discrete selection variables, evaluating the loss still requires separate forward passes over exponentially many subsets combinations. As a result, naively adapting classical sparsity-inducing penalties to the LLM input space leads to a non-smooth, combinatorially complex, and intractable optimization problem.

This paper addresses these challenges by developing a principled feature-selection framework for serialized LLM prompts, analyzing its generalization properties, and demonstrating its utility in Alzheimer's disease (AD) prediction, HIV treatment outcome prediction, and controlled synthetic studies. Conceptually, our framework plays a role in LLM-based prediction analogous to that of LASSO and group LASSO (Yuan and Lin, 2006): it induces sparsity over variable groups, provides variable-importance measures, controls complexity through regularization, and is directly coupled to the LLM's inference pipeline. Our contributions are further summarized as follows.

(1) Feature-selection tailored to serialized LLM prompts. We propose an embedding-level framework for group-wise feature selection in serialized prompts. Variable groups are embedded into block-structured prompt representations, where each group is either retained or replaced by a padding-token embedding, yielding a fixed-dimensional masked embedding matrix suitable for downstream optimization. We minimize a loss function with relaxed ℓ_0 penalty using the ReinMax gradient method (Liu et al., 2023), derive variable-importance scores from the learned logits, and select sparse subsets. The framework is model-agnostic and applies to both pretrained and fine-tuned LLMs.

(2) Generalization bounds for LLM variable selection. We show that the Bernoulli relaxation is lossless with respect to the original discrete selection problem in the population, justifying its practical application. We further derive a PAC-Bayesian risk bound. This bound scales linearly with the number of variable groups in general, but only logarithmically under sparse selection; consequently, sparser selections provide tighter generalization guarantees.

(3) Empirical insights in healthcare LLMs. In the Alzheimer's disease application, our framework improves amyloid-β and tau prediction while reducing the number of variable groups by nearly 45% and the token budget by 57%, identifying sparse, clinically meaningful subsets involving genetics, family history, vitals, and MRI-derived structural measures. In a large-scale HIV electronic medical record study from the Tanzanian national CTC3 repository, the framework achieves competitive or strongest performance on viral-load, ART non-adherence, and loss-to-follow-up prediction while prioritizing longitudinal treatment and care-engagement variables. Complementary Monte Carlo studies calibrated from the NIDDK diabetes dataset further show that optimal subsets for fine-tuned LLM inference are low-dimensional but only partially align with the underlying tabular data-generating process, motivating LLM-aware feature-selection strategies and opening new research directions at the interface of LLM inference and statistical variable selection.

1.1. *Related literature* A substantial body of work has examined interpretability and variable importance in neural networks. Post-hoc attribution methods such as Integrated Gradients (Sundararajan, Taly and Yan, 2017), LIME (Ribeiro, Singh and Guestrin, 2016), SHAP (Lundberg and Lee, 2017), and TokenSHAP (Goldshmidt and Horovicz, 2024) estimate feature- or token-level relevance via gradient-based attribution or input perturbation. While useful for descriptive analysis and model diagnostics, their goal is explanation rather than selection: they characterize the relevance of inputs already present in the prompt but do not identify which variable groups can be safely omitted, nor do they couple importance scores to a sparsity-inducing optimization objective.

Attention weights in Transformer models (Vaswani, 2017) are sometimes treated as a proxy for feature importance. However, raw attention weights are known to be influenced by positional bias, normalization, and architectural details (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). In semi-structured clinical prompts, attention can concentrate on punctuation, separators, or boundary tokens rather than on clinically meaningful content (Xiao et al., 2023; Guo, Kamigaito and Watanabe, 2024), making it difficult to infer variables that are actually necessary for prediction. As we show empirically, attention-based feature rankings do not reliably identify high-performing subsets of serialized variable groups for fine-tuned LLMs.

This paper also connects to the statistics literature on sparsity-inducing procedures, such as the LASSO (Tibshirani, 1996), Group Lasso (Yuan and Lin, 2006), Elastic Net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), supervised homogeneity fusion (Wang et al., 2024), and MCP (Zhang, 2010), all of which provide rigorous frameworks for variable selection in fixed-design tabular models. More recent works that account for structured dependence or individualized sparsity (Zhou and Qu, 2012; Tang, Xue and Qu, 2021) further highlight the importance of respecting predictor correlations and heterogeneity in selection. These approaches, however, assume a fixed-dimensional predictor matrix and hence do not easily extend to serialized prompts. Recent sparsity-inducing mechanisms for neural networks (Louizos, Welling and Kingma, 2017; Yamada et al., 2020) introduce learnable sparsity over fixed-dimensional vectorized inputs but do not perform discrete removal of semantically coherent token groups. The specific challenge of selecting subsets of serialized variable groups in the LLM input space, tractably and in alignment with the model's inference pipeline, falls outside the scope of these approaches, and motivates the framework developed in this paper.

2. LLM inference: Formal setup This section describes the LLM inference process and formulates the motivation of our proposed methods. To begin with, each individual's raw data is represented by K clinically or semantically separated variable groups $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)$, such that each component $\boldsymbol{z}_k$ collects a set of variables that may be structured (e.g., demographics) or unstructured (e.g., clinical notes); see Examples 1 and 2 below. These variable groups are first serialized into text segments x, then tokenized and embedded into an initial matrix $\boldsymbol{X}^{(0)}$ that the LLM transforms autoregressively into an output distribution over predictions or generated text $\widehat{\boldsymbol{y}}$; that is, $\boldsymbol{z} \to \mathsf{x} \to \boldsymbol{X}^{(0)} \to \widehat{\boldsymbol{y}}$.

2.1. *Data serialization and motivating examples* Raw healthcare data are often serialized into textual prompts before being supplied to a language model. Let $\mathsf{x} =$ "x_1; x_2; ..." denote the serialized prompt, where each x_k corresponds to a semantically meaningful variable group.

EXAMPLE 1 (Serialized Tabular Data; Wei et al. 2026; Wang et al. 2025). Consider patient-level data grouped into demographics, vitals, and laboratory measurements:

$$\boldsymbol{z}_1 = \{\text{Age}, \text{Sex}\}, \quad \boldsymbol{z}_2 = \{\text{BMI}, \text{BloodPressure}\}, \quad \boldsymbol{z}_3 = \{\text{Cholesterol}, \text{HbA1c}\}.$$

These variables can be serialized into prompt segments such as $\mathsf{x}_1 =$ "Age: 62; Sex: Female", $\mathsf{x}_2 =$ "BMI: 27.3; BP: 120/80", and $\mathsf{x}_3 =$ "Cholesterol: 212; HbA1c: 6.1%", which are concatenated into a single prompt for downstream prediction.

While serialized prompts preserve the original group structure, incorporating all groups may be suboptimal: some variables may contribute little predictive value while increasing context length, computational cost, and attention dilution (Croxford et al., 2025). This motivates group-wise feature selection for LLM inference.

EXAMPLE 2 (Segmented Electronic Medical Records; Davis et al. 2026; Zhou and Miller 2024). Electronic medical records (EMR) often combine structured variables with unstructured clinical notes. For example, a free-text note describing symptoms, review of systems, and assessment may be segmented by an LLM-based parser into groups such as HPI, ROS, and Assessment/Plan: $\mathsf{x}_4 =$ "HPI: intermittent chest discomfort", $\mathsf{x}_5 =$ "ROS: denies fever or dyspnea", $\mathsf{x}_6 =$ "Assessment: concern for angina". Combined with demographics, vitals, and medication history, these segments form the final serialized prompt used for prediction.

As this example illustrates, longitudinal EMR prompts can quickly become extremely long and noisy, especially when large amounts of historical free text are included. Since only a subset of variable groups may be relevant for a prediction task, indiscriminately including all information can increase inference cost and degrade predictive performance. This motivates a principled group-level feature-selection framework for LLM-based prediction.

2.2. *Tokenization and one-hot transformation* The next step of LLM inference converts the prompt x into a sequence of discrete tokens using a tokenizer (e.g., byte-pair encoding or SentencePiece), $\texttt{Tok}(\mathsf{x}_k) = (\boldsymbol{t}_{k,1}, \boldsymbol{t}_{k,2}, \dots, \boldsymbol{t}_{k,J_k})$, where J_k is the number of tokens in the kth segment in the prompt, and $\boldsymbol{t}_{k,j}$ is a token ID. Each token ID corresponds to a one-hot vector in a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$ (fixed after pre-training), $\boldsymbol{t}_{k,j} \to \boldsymbol{e}_{\boldsymbol{t}_{k,j}} \in \{0,1\}^{|\mathcal{V}|}$. Concatenating across all groups yields the full one-hot matrix

$$(\boldsymbol{e}_{\boldsymbol{t}_{1,1}}, \dots, \boldsymbol{e}_{\boldsymbol{t}_{1,J_1}},\ \boldsymbol{e}_{\boldsymbol{t}_{2,1}}, \dots, \boldsymbol{e}_{\boldsymbol{t}_{2,J_2}},\ \dots,\ \boldsymbol{e}_{\boldsymbol{t}_{K,1}}, \dots, \boldsymbol{e}_{\boldsymbol{t}_{K,J_K}}).$$

We note that tokenization and its implicit one-hot expansion occurs at the token level rather than the variable level, meaning that even a single structured field (e.g., "Age: 62") may map to multiple tokens.

2.3. *Transform tokenized input into high-dimensional embeddings* Given the serialized and tokenized input, Figure 1 illustrates the schematic of our inference pipeline. Following one-hot encoding, each vector is projected into a high-dimensional dense space via the model's learned embedding matrix. The resulting matrix of token representations is formed by horizontally stacking these vectors blockwise according to their respective prompt segments: $[\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \dots,\ \boldsymbol{X}_K]$. In this formulation, each block $\boldsymbol{X}_k$ comprises the token embeddings for segment x_k, meaning the number of columns in $\boldsymbol{X}_k$ corresponds to J_k (the total number of tokens in x_k). Ultimately, this initial embedding matrix, denoted as $\boldsymbol{X}^{(0)} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_K]$, is fed into the decoder-only LLM to perform autoregressive inference.

2.4. *Autoregressive output generation* Given the serialized prompt embedding $\boldsymbol{X}^{(0)}$ and previously generated tokens $(\widehat{y}_1, \dots, \widehat{y}_{t-1})$, let $\boldsymbol{X}^{(t-1)}$ denote the concatenated embeddings up to step $t-1$. These embeddings are processed by a transformer decoder: $\mathbf{h}^{(t-1)} = \texttt{Decoder}_\theta(\boldsymbol{X}^{(t-1)}) \in \mathbb{R}^d$, where d is the embedding dimension, θ denotes the LLM parameters , and the decoder consists of stacked L self-attention transformer blocks (Vaswani, 2017). The hidden state is then projected into the vocabulary space via an "unembedding"

$$\underbrace{\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_K)}_{\text{raw data in } K \text{ groups}} \xrightarrow{\text{serialization}} \underbrace{\mathsf{x}_{\text{prompt}} = \text{``}\mathsf{x}_1; \mathsf{x}_2; \ldots\text{''}}_{\text{serialized prompt}} \xrightarrow{\text{tokenization}} \underbrace{(\boldsymbol{t}_{1,1}, \cdots, \boldsymbol{t}_{1,J_1}, \ldots, \boldsymbol{t}_{K,J_K})}_{\text{tokenized prompt}}$$

$$\xrightarrow{\text{one-hot transformation}} \underbrace{(\boldsymbol{e}_{\boldsymbol{t}_{1,1}}, \ldots, \boldsymbol{e}_{\boldsymbol{t}_{1,J_1}}, \ldots, \boldsymbol{e}_{\boldsymbol{t}_{K,J_K}})}_{\text{one-hot matrix}} \xrightarrow{\text{embedding}} \underbrace{\boldsymbol{X}^{(0)} = [\boldsymbol{X}_1,\, \boldsymbol{X}_2,\, \ldots,\, \boldsymbol{X}_K]}_{\text{prompt initial embeddings}}$$

Fig 1: Illustration of input prompt embedding procedure in a decoder-only LLM.

matrix $= \mathbf{W}_{L+1}$ and a bias vector $\mathbf{b}_{L+1}$: $\mathbf{m}^{(t)} = \mathbf{W}_{L+1}\mathbf{h}^{(t-1)} + \mathbf{b}_{L+1} \in \mathbb{R}^{|\mathcal{V}|}$, yielding logits over the vocabulary $\mathcal{V}$. Applying a softmax gives the next-token distribution:

$$p_\theta(y_t = \texttt{Detokenization}(\boldsymbol{t}_v) \mid \boldsymbol{X}^{(t-1)}) = \frac{\exp(\mathbf{m}_v^{(t)}/\tau)}{\sum_{j=1}^{|\mathcal{V}|} \exp(\mathbf{m}_j^{(t)}/\tau)}, v = 1, ..., |\mathcal{V}|,$$

where τ is the temperature parameter. The next token $\widehat{y}_t$ is sampled from this distribution, and the process repeats autoregressively to generate the final output sequence $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \ldots, \widehat{y}_T)$.

3. Methodology Building upon the LLM inference pipeline above, we next discuss why feature selection for LLMs differs fundamentally from classical statistical learning (say, high-dimensional linear models). These challenges motivate our padding-based smoothed feature-selection framework, which operates directly in embedding space while preserving fixed input dimensions. The full algorithm and optimization details are provided in Subsection 3.2.

3.1. *Why feature selection in LLM input prompts is challenging?* As noted above, language models do not operate directly on variable groups, but rather on serialized text derived from them. Specifically, recall that an individual's raw data are partitioned into clinically or semantically meaningful groups $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)$, which are further transformed into the initial embedding matrix $\boldsymbol{X}^{(0)} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K]$. This inference pipeline makes feature selection fundamentally nontrivial, as removing a variable group $\boldsymbol{z}_k$ corresponds to deleting a variable-length block of tokens and altering the dimension of $\boldsymbol{X}^{(0)}$, while the self-attention mechanism further entangles information across groups. Consequently, naive adaptations of classical feature-selection methods to LLM prompts become combinatorially complex and numerically unstable.

To fix idea, let the binary vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_K) \in \{0, 1\}^K$ denote feature selection: a variable group k is retained if $s_k = 1$. Then, the naive variable group selection procedure can be framed as

$$\text{Naive selection:} \qquad \min_{\boldsymbol{s}} \mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \{\boldsymbol{X}_k : s_k = 1\})\Big) + \lambda \sum_{k=1}^{K} s_k,$$

where $\mathcal{L}_{\texttt{LLM}}$ is some loss function, which we discuss in detail further below. The naive selection problem is clearly computationally infeasible due to its combinatorial nature.

A natural attempt is to *replace the discrete variable group indicators by a set of real-valued parameters*, which we term as *relaxation* or *parameterization*. Specifically, we model each binary selection variable $s_k \in \{0, 1\}$ as a sample from a latent Bernoulli generation process:

$$s_k \sim \texttt{Bernoulli}(\texttt{Logit}(\beta_k)) \quad \text{with } \texttt{Logit}(\beta_k) = e^{\beta_k}/(1 + e^{\beta_k}).$$

Taking $\beta = (\beta_1, \beta_2, \ldots, \beta_K) \in \mathbb{R}^K$, the loss function is replaced with the expected value, and optimization is with respect to the Bernoulli parameters, leading to

$$\text{Naive selection:} \quad \min_{\beta} \mathcal{L}(\beta) \; = \; \min_{\beta} \mathbb{E}_{\boldsymbol{s}} \Big[\mathcal{L}_{\texttt{LLM}} \Big(\boldsymbol{y}, p_\theta(\cdot \mid \{ \boldsymbol{X}_k : s_k = 1 \}) \Big) + \lambda \sum_{k=1}^{K} s_k \Big].$$

Unfortunately, this parameterization step alone is inadequate to address the computational challenges unique to the LLM inference pipeline. To illustrate, consider $K = 2$ and assume we are only interested in selecting the first variable group. The loss function becomes a simple convex combination of $\mathcal{L}_{\texttt{LLM}}(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}_1, \boldsymbol{X}_2))$ and $\mathcal{L}_{\texttt{LLM}}(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}_2))$ with weights $\texttt{Logit}(\beta_1)$ and $1 - \texttt{Logit}(\beta_1)$. While the final objective function is smooth (in fact, linear in the logits), constructing it still requires evaluating the loss twice: once with the first variable group included, and once without. With K variable groups, this process requires 2^K evaluations. Thus, although the parameterization step yields a smooth objective, it fails to resolve the underlying combinatorial complexity, rendering gradient-based optimization methods inapplicable.

The core issue is that, through serialization, tokenization, one-hot transformation and embedding, the processed embedding matrix $\boldsymbol{X}$ depends on the variable group selection $\boldsymbol{s}$ in a highly nonsmooth manner through its dimension: the column dimension of $\boldsymbol{X}$ with selection is $\sum_{k:s_k=1} J_k$. Such *fundamental lack of smoothness* not only makes the naive optimization problem computationally prohibitive even with the parameterization introduced earlier, but also renders the loss function numerically unstable due to the changing dimension of its input embedding matrix.

To address this, our primary methodological contribution is to *introduce a smooth interpolation among the embedding matrices with different selection vectors by incorporating padding tokens* (which are fixed placeholder tokens whose embeddings carry no semantic meaning). Consequently, in conjunction with the Bernoulli relaxation introduced earlier, variable group selection becomes a smooth operation within the embedding space. Fortunately, the downstream components of the LLM (i.e., self-attention and feed-forward networks) are differentiable operations, thereby facilitating gradient-based optimization. For this reason, the embedding matrix serves as the natural and mathematically principled location to impose feature selection. Formally, we define the blockwise Hadamard product as $\boldsymbol{s} \odot \boldsymbol{X} = [s_1 \boldsymbol{X}_1, s_2 \boldsymbol{X}_2, \ldots, s_K \boldsymbol{X}_K]$. The resulting embedding matrix corresponding to the selection vector $\boldsymbol{s}$ is then given by:

$$\boldsymbol{X}(\boldsymbol{s}) = \boldsymbol{s} \odot \boldsymbol{X} + (1 - \boldsymbol{s}) \odot \boldsymbol{H},$$

where $\boldsymbol{H}$ is the embedding matrix of padding tokens, which has the same dimension as $\boldsymbol{X}$. We illustrate the difference between our method and naive selection on the embedding matrix in Figure 2.

3.2. *Proposed method: Smoothed feature selection in LLM* We now formally introduce our smoothed feature selection method in LLM. The unit-level objective function is

$$\text{(1)} \quad \text{Unit-level:} \; \min_{\beta} \mathcal{L}(\boldsymbol{y}, \boldsymbol{X}, \beta) \; = \; \min_{\beta} \mathbb{E}_{\boldsymbol{s}} \Big[\mathcal{L}_{\texttt{LLM}} \Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s})) \Big) \Big] + \lambda \sum_{k=1}^{K} \texttt{Logit}(\beta_k).$$

As discussed in the previous subsection, the expectation is taken over the random mask $\boldsymbol{s} \sim \texttt{Bernoulli}(\texttt{Logit}(\beta))$, which enters the objective function through the penalty term as well as the masked embedding matrix $\boldsymbol{X}(\boldsymbol{s})$. We remark that although sampling is discrete and non-differentiable, the gradient of $\mathcal{L}(\beta)$ with respect to β can be approximated through differentiable surrogate estimators. To be specific, we adopt the ReinMax (Liu et al., 2023),

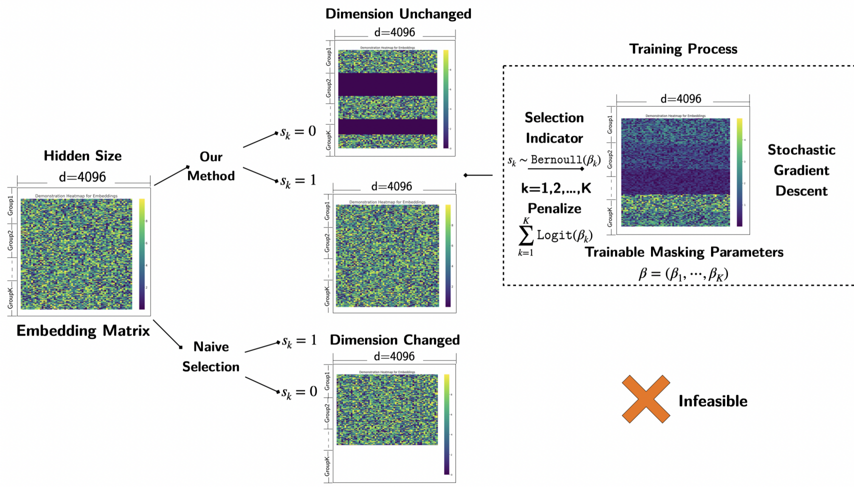


Fig 2: Visual comparison between our method and naive selection. The left column and the middle of the center column show the embedding matrix without any selection. The naive approach, demonstrated in the bottom of the center column, simply drops tokens and leads to embedding matrices with variable dimensions. Our method, shown at the top of the center, keeps the dimension unchanged through the incorporation of padding tokens.

which we discuss further below. Also see Supplementary Material (Lyu, Ma and Wang, 2026) for additional details.

Other ingredients of the objective function are (i) $\mathcal{L}_{\texttt{LLM}}$, which denotes the next-token prediction loss, (ii) $\boldsymbol{y}$, the desired output from the data, and (iii) λ, a non-negative regularization parameter that controls the trade-off between predictive performance and input sparsity. When the output is formatted as a single categorical variable, it is common to adopt the negative log-likelihood loss:

$$\mathcal{L}_{\texttt{LLM}}\Big(y, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) = -\log p_\theta(y \mid \boldsymbol{X}(\boldsymbol{s})).$$

For outputs that are formatted to contain multiple tokens, it is possible to employ a sequential loss, as characterized by Hu et al. (2021):

$$\mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) = -\sum_j \log p_\theta\left(y_j \mid \boldsymbol{X}(\boldsymbol{s}), y_1, \ldots, y_{j-1}\right).$$

As a companion to our methodological development, we first show that the relaxation or parameterization step, which replaces the discrete and combinatorial optimization with a search on a continuum in (1), does not result in any excess loss.

LEMMA 1 (No loss from stochastic relaxation). *Consider the setup in this section. Then*

$$\min_\beta \mathcal{L}(\boldsymbol{y}, \boldsymbol{X}, \beta) = \min_{\boldsymbol{s}} \left\{ \mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) + \lambda \sum_{k=1}^{K} s_k \right\},$$

where $\mathcal{L}(\boldsymbol{y}, \boldsymbol{X}, \beta)$ is defined in (1).

We now define the population objective function as

$$\text{(2)}\quad \text{Population: } \min_\beta \mathcal{L}(\beta) = \min_\beta \mathbb{E}_{(\boldsymbol{y},\boldsymbol{X})} \mathbb{E}_{\boldsymbol{s}} \Big[\mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) \Big] + \lambda \sum_{k=1}^{K} \texttt{Logit}(\beta_k).$$

Then the following is an immediate corollary, showing that relaxation/parameterization also does not lead to additional loss at the population level.

COROLLARY 1. *Assume the loss function is integrable with respect to $\boldsymbol{y}$ and $\boldsymbol{X}$. Then*

$$\min_{\beta} \mathcal{L}(\beta) = \min_{\boldsymbol{s}} \left\{ \mathbb{E}_{(\boldsymbol{y},\boldsymbol{X})} \left[\mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) \right] + \lambda \sum_{k=1}^{K} s_k \right\},$$

where $\mathcal{L}(\beta)$ is defined in (2).

3.3. *Sample-level implementation* We now discuss implementation at the sample level. Following standard practice, we partition the data into three disjoint subsets: a training set $\mathcal{D}_{\texttt{train}}$, a validation set $\mathcal{D}_{\texttt{val}}$, and a test set $\mathcal{D}_{\texttt{test}}$. The pair $(\mathcal{D}_{\texttt{train}}, \mathcal{D}_{\texttt{val}})$ is used exclusively for obtaining the LLM parameters θ and tuning any hyperparameters, either via pre-training or domain-specific fine-tuning; the test set $\mathcal{D}_{\texttt{test}} = \{\mathbf{y}_i, \mathbf{z}_i\}_{i=1}^n$ is held out and plays no role in this stage. Once θ is fixed (i.e., the LLM is "frozen"), our proposed procedure is applied to $\mathcal{D}_{\texttt{test}}$ to select important variable groups for downstream output generation, and to assign group-wise importance scores. Employing $\mathcal{D}_{\texttt{test}}$, the objective function in Equation (1) becomes

$$\min_{\beta} \mathcal{L}(\mathcal{D}_{\texttt{test}}, \beta) \;=\; \min_{\beta} \mathbb{E}_{\boldsymbol{s}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}_i, p_\theta(\cdot \mid \boldsymbol{X}_i(\boldsymbol{s}))\Big) \right] + \lambda \sum_{k=1}^{K} \texttt{Logit}(\beta_k). \tag{3}$$

We denote a minimizer by $\widehat{\beta}$. Thus, we define: (i) a feature importance score for each group k, and (ii) a set of selected features obtained by thresholding these scores

$$w_k = \texttt{Logit}(\widehat{\beta}_k), \quad \widehat{\mathcal{M}} = \{k : w_k > 0.5\}.$$

This yields a group-wise, sample-level characterization of which input components of the prompt most strongly drive the LLM's predictions.

A key step in optimizing the sample-level objective in (3) is to compute the gradient with respect to the Bernoulli logit parameter β. Let $f_i(\boldsymbol{s}) = \mathcal{L}_{\texttt{LLM}}(\boldsymbol{y}_i, p_\theta(\cdot \mid \boldsymbol{X}_i(\boldsymbol{s})))$. The exact calculation, however, requires averaging discrete toggle differences $f_i(s_k = 1, \boldsymbol{s}_{-k}) - f_i(s_k = 0, \boldsymbol{s}_{-k})$ over all configurations $\boldsymbol{s}_{-k} \in \{0,1\}^{K-1}$ of the remaining $K-1$ groups, which scales exponentially with 2^{K-1} and becomes computationally infeasible for moderate or large K. To avoid this exhaustive computation, we use the ReinMax method (Liu et al., 2023) that replaces the discrete toggle difference with a continuous surrogate, thereby enabling efficient optimization with gradient-based methods. Specifically, ReinMax employs a predictor-corrector (Heun-style) construction that approximates the true gradient to second-order accuracy. More technical details and implementations are given in Supplementary Material Section 4.5 (Lyu, Ma and Wang, 2026), and the complete training process is given in Algorithm 1.

3.4. *Tuning parameter selection with GBIC* Determining the optimal sparsity level for feature selection requires a grid search over a range of tuning parameters $\lambda \in \Lambda$, which control the sparsity penalty in (1). For each candidate λ, the selection parameters β are optimized on the test set $\mathcal{D}_{\texttt{test}}$, and to optimally balance predictive accuracy and model parsimony, we utilize a Generalized Bayesian Information Criterion (GBIC), computed as:

$$\texttt{GBIC}(\lambda) = 2 \cdot \sum_{i=1}^{n} \mathcal{L}_{\texttt{LLM}}(\boldsymbol{y}_i, p_\theta(\cdot | \boldsymbol{X}_i(\boldsymbol{s}_{\widehat{\mathcal{M}}(\lambda)}))) + \log(n) \cdot \texttt{DoF}(\lambda).$$

Here, $\boldsymbol{s}_{\widehat{\mathcal{M}}(\lambda)}$ denotes variable group selection for a specific penalty parameter λ, and $\texttt{DoF}(\lambda)$ is the degrees of freedom adjustment. We approximate $\texttt{DoF}(\lambda)$ using $|\widehat{\mathcal{M}}(\lambda)|$, the number of variable groups retained under λ.

Algorithm 1 Variable Group Selection

Require: Frozen LLM weights θ, number of variable groups K, test set $\mathcal{D}_{\texttt{test}}$, training steps T, sparsity tuning parameter λ, learning rate η, and batch size b.

Initialization: Initialize the Bernoulli logit as $\beta_k = 2$, corresponding to an initial retention probability of 0.88. Set the initial binary mask $s_k^{(0)} = 1$ for all k.

for $\mathsf{t} = 1, 2, \ldots, \mathsf{T}$ **do**

Sample minibatch $\{\boldsymbol{z}_i, \boldsymbol{y}_i\}_{i=1}^{b} \sim \mathcal{D}_{\texttt{test}}$.

Pad all samples in the batch to a uniform length across all variable groups.

Serialize $\boldsymbol{z}_i$ and compute its embedding matrix $\boldsymbol{X}_i$. Generate the masked input $\boldsymbol{X}_i(\boldsymbol{s}^{(\mathsf{t}-1)})$

Compute the LLM prediction loss:

$$\ell_{\texttt{pred}} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}_i, p_\theta(\cdot \mid \boldsymbol{X}_i(\boldsymbol{s}^{(\mathsf{t}-1)}))\Big).$$

Compute the expected sparsity penalty:

$$\ell_{\texttt{sparsity}} \leftarrow \lambda \sum_{k=1}^{K} \texttt{Logit}(\widehat{\beta}_k^{(\mathsf{t}-1)})$$

Update β using the ReinMax gradient estimator:

$$\widehat{\beta}^{(\mathsf{t})} \leftarrow \widehat{\beta}^{(\mathsf{t}-1)} - \eta\, \widehat{\nabla}_{\texttt{ReinMax}}\big(\ell_{\texttt{pred}} + \ell_{\texttt{sparsity}}\big)$$

Sample $\boldsymbol{s}^{(\mathsf{t})} \sim \texttt{Bernoulli}(\texttt{Logit}(\widehat{\beta}^{(\mathsf{t})}))$.

end for

Variable group importance: report $\texttt{Logit}(\widehat{\beta})$ as $\texttt{Logit}(\widehat{\beta}^{(\mathsf{T})})$.

Variable group selection: retain the kth variable group if $\texttt{Logit}(\widehat{\beta}_k^{(\mathsf{T})}) > 0.5$ and generate the selected variable groups: $\widehat{\mathcal{M}} = \{k : \texttt{Logit}(\widehat{\beta}_k) > 0.5\}$.

3.5. *Application of the proposed method on fine-tuned LLMs* Given the model-agnostic nature of our formulation, the LLM objective can readily incorporate domain-specific weight adjustments, denoted by $\Delta\theta$, leading to $\mathcal{L}_{\texttt{LLM}}(\boldsymbol{y}, p_{\theta+\Delta\theta}(\cdot \mid \boldsymbol{X}(\boldsymbol{s})))$. For instance, Low-Rank Adaptation (LoRA, Hu et al. 2021) fine-tunes a pretrained model by injecting trainable low-rank matrices into the attention and feed-forward layers, such that $\Delta\theta = BA$ for $A \in \mathbb{R}^{r\times h}$ and $B \in \mathbb{R}^{h\times r}$ $(r \ll h)$. Because both the base parameters θ and the adapter weights $\Delta\theta$ remain frozen during our optimization phase, this feature-selection procedure naturally accommodates fine-tuned language models. We illustrate this capability in Section 5 using case studies on AD pathology and people living with HIV (PLHIV) viral load, lost-to-follow-up, and Antiretroviral Therapy (ART) non-adherence predictions.

4. Theoretical investigation Having established that the parameterization step incurs no loss (Lemma 1 and Corollary 1), we now derive a PAC-Bayesian generalization bound for the relaxed estimator that connects the empirical risk to the population risk. For concreteness, we focus on the negative log-likelihood loss introduced earlier for one-token prediction, or its sequential extension whenever the output involves a sequence of tokens:

$$f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{s}) = \mathcal{L}_{\texttt{LLM}}\Big(\boldsymbol{y}, p_\theta(\cdot \mid \boldsymbol{X}(\boldsymbol{s}))\Big) = -\sum_{j} \log p_\theta\left(y_j \mid \boldsymbol{X}(\boldsymbol{s}), y_1, \ldots, y_{j-1}\right),$$

where we introduce the shorthand $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{s})$ to streamline notation. Let

$$\mathcal{R}(\mathbb{P}) = \mathbb{E}_{(\boldsymbol{y},\boldsymbol{X})} \mathbb{E}_{\boldsymbol{s}\sim\mathbb{P}}\big[f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{s}) + \lambda\|\boldsymbol{s}\|_0\big], \quad \widehat{\mathcal{R}}_n(\mathbb{P}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{s}\sim\mathbb{P}}\big[f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{s}) + \lambda\|\boldsymbol{s}\|_0\big].$$

This leads to the following result.

THEOREM 1 (PAC-Bayesian bound for relaxed LLM gating). *Let $\mathbb{P}$ be any posterior over $\{0,1\}^K$ and $\mathbb{P}_0$ be any prior independent of the data. Then with probability at least $1-\delta$,*

$$\mathcal{R}(\mathbb{P}) \le \widehat{\mathcal{R}}_n(\mathbb{P}) + \mathfrak{c}\sqrt{\frac{\mathtt{KL}(\mathbb{P}\|\mathbb{P}_0) + \log(\mathfrak{c}'/\delta)}{n}},$$

where $\mathtt{KL}(\cdot)$ is the Kullback-Leibler divergence, and $\mathfrak{c}$ and $\mathfrak{c}'$ are constants.

Theorem 1 establishes that the population risk of the relaxed selection mechanism is bounded by its empirical risk plus a data-dependent complexity penalty. To build intuition, consider the KL divergence between the prior and the posterior: $\mathtt{KL}(\mathbb{P} \parallel \mathbb{P}_0)$. As demonstrated by Lemma 1 and Corollary 1, the stochastic relaxation preserves the exact global minimum, given that it simply forms a convex combination of the discrete configuration losses. Consequently, the global optimum $\min_\beta \mathcal{L}(\mathcal{D}_{\mathtt{test}}, \beta)$ corresponds to a deterministic configuration in $\{0,1\}^K$. Assuming a uniform prior $\mathbb{P}_0$, the KL divergence takes a much simpler form $\mathtt{KL}(\text{point mass} \parallel \text{uniform prior}) = \log(2^K) \le K$, which yields the following bound:

$$\mathcal{R}(\mathbb{P}) \le \widehat{\mathcal{R}}_n(\mathbb{P}) + \mathfrak{c}\sqrt{\frac{K + \log(\mathfrak{c}'/\delta)}{n}}.$$

This theoretical bound is quite practical, given that the total number of variable groups is typically much smaller than the overall sample size.

We make two further remarks. First, the generalization bound established using a point-mass posterior matches that in the classical PAC-ERM (empirical risk minimization) framework. However, directly applying empirical risk minimization is computationally intractable, as discussed in the previous section. Our proposed stochastic relaxation effectively circumvents this computational bottleneck. Second, we recognize that our theoretical guarantees implicitly assume the numerical optimization successfully converges to the global minimum. In practice, this convergence is inherently constrained by the total number of training steps and therefore by the available computational resources.

While our numerical algorithm employs hard-thresholding to deliver a fixed set of selected variable groups, we provide another result below to complement our theoretical analysis in Theorem 1, which accommodates randomized selection under sparsity.

COROLLARY 2. *Let $\mathbb{P}_0$ factorize as i.i.d.* $\mathtt{Bernoulli}(s_0/K)$*, and assume the posterior satisfies $\mathbb{E}_{\boldsymbol{s}\sim\mathbb{P}}\|\boldsymbol{s}\|_0 \le s_0 \le K/2$ almost surely. Then*

$$\mathtt{KL}(\mathbb{P}\|\mathbb{P}_0) \le s_0 \log\left(\frac{eK}{s_0}\right).$$

This new generalization bound scales only logarithmically with the total number of variable groups K, provided that one is willing to impose a sparsity level on the posterior randomized variable group selection.

5. Case study I: What are important features for Alzheimer's Disease pathology prediction? Amyloid and tau are hallmarks of Alzheimer's disease (AD) (Nordberg, 2004; Okamura et al., 2014), yet their measurement relies on PET imaging, which remains invasive, costly, and often inaccessible. This has motivated growing interest in predicting PET-based AD pathology from noninvasive, routinely collected data such as demographics, cognitive assessments, MRI-derived measures, and, more recently, plasma biomarkers. Different cohorts, however, collect different subsets of these modalities (e.g., plasma biomarkers primarily in newer studies, MRI more frequently in resource-rich settings), leading to heterogeneous and partially overlapping feature spaces across studies.

In this setting, LLM-based prediction models operating on serialized multimodal inputs across multiple studies offer a natural framework for leveraging heterogeneous data to predict AD pathology endpoints, including amyloid and tau burden measured through PET imaging and future cognitive decline. In response to this need, we fine-tuned LLaMA-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) as backbone models using LoRA (Hu et al., 2021) on serialized data from four large-scale AD cohorts: (1) ADNI (Mueller et al., 2005), a longitudinal multi-center study of late-onset AD; (2) HABS-HD (Petersen et al., 2025), which follows racially and ethnically diverse communities; (3) the Anti-Amyloid Treatment in Asymptomatic Alzheimer's (A4) study (Sperling et al., 2014); and (4) baseline data from the U.S. POINTER imaging study (Harrison et al., 2025), a randomized lifestyle-intervention trial for healthy older adults at risk for cognitive decline. Following standard practice, we use 72% of the data as $\mathcal{D}_{\texttt{train}}$, 8% as $\mathcal{D}_{\texttt{val}}$, and the remaining 20% as $\mathcal{D}_{\texttt{test}}$. We fine-tune the LLaMA-3.1-8B base model on serialized cohort-specific prompts using LoRA (rank = 8, alpha = 16, dropout = 0.0) applied to projection modules. Training was conducted for 3 epochs (10,548 global steps) with per-device batch size 1 and gradient accumulation 4, using AdamW with cosine learning-rate scheduling and initial learning rate 10^{-4}. The full training process took approximately 15.2 hours and processed 67.1M input tokens, achieving a final training loss of approximately 0.0557. Additional implementation details for Qwen3-8B are provided in Supplementary Material (Lyu, Ma and Wang, 2026). The resulting **ADLLM** serves as a flexible generative predictor of PET-based amyloid and tau categories as well as cognitive decline trajectories.

Realizing the full potential of ADLLM in real-world clinical practice, however, requires addressing two practical considerations. First, many of the variables included in ADLLM are not routinely available at deployment, particularly at clinical sites with limited resources. It is therefore important to identify which inputs are most influential for accurate ADLLM inference and to select a parsimonious subset of variables. Our exploratory analyses (Supplementary Material, Lyu, Ma and Wang 2026) further show that including certain variable groups can even degrade predictive performance on held-out data. Second, the full serialized input for ADLLM can saturate the context window of its LLaMA-3.1-8B and Qwen3-8B backbones, making it desirable to reduce input token length while preserving predictive accuracy.

Our proposed feature-selection method naturally addresses both considerations. Using the held-out POINTER cohort, we first show that **ADLLM benefits from sparse prompt input**. In both backbones, a parsimonious subset of six or seven variable groups is sufficient to improve ADLLM performance on Aβ and meta-tau prediction compared to using all eleven available variable groups, while reducing the number of variable groups by nearly 45% and the token budget by nearly 57% (Table 1).

Our framework also delivers **clinically meaningful feature identification with strong predictive performance**. Relative to baseline approaches including LLM-Rank (Jeong, Lipton and Ravikumar, 2024), attention-based selection (Wiegreffe and Pinter, 2019), CatBoost feature importance (Dorogush, Ershov and Gulin, 2018), and deterministic or stochastic gating (Bengio, Léonard and Courville, 2013; Louizos, Welling and Kingma, 2017), our method is the only one that simultaneously achieves competitive or strongest discriminative performance (AUC, accuracy, and specificity) and selects a feature subset aligned with established AD biology (Section 5.3).

Finally, we also observe **divergent feature importance between ADLLM and classical machine learning methods**. For example, CatBoost and attention-based methods assign high importance to cognitive diagnosis, even though all POINTER participants are cognitively normal at baseline, whereas ADLLM-based selection prioritizes genetics and MRI-derived measures that more directly reflect AD pathology (Table 1).

5.1. *Clinically meaningful variable groups* ADLLM operates on thirteen clinically meaningful variable groups, capturing complementary aspects of AD biology and clinical presentation, including genetic susceptibility, vascular and metabolic risk factors, cognitive status, neurodegeneration measured by MRI, and molecular or metabolic signatures from blood biomarkers and FDG PET (x_1: demographics, x_2: genetics, x_3: family history of AD, x_4: vital signs, x_5: cognitive tests, x_6: physical measures, x_7: brain volume (MRI), x_8: brain thickness (MRI), x_9: AD biomarker in blood plasma, x_{10}: fluorodeoxyglucose (FDG) PET imaging, x_{11}: comorbidities, x_{12}: medications, and x_{13}: cognitive diagnosis). They are used to predict three endpoints of AD pathology and clinical progression: amyloid-β (Aβ) PET burden (Negative, Intermediate negative, Positive, Clearly positive), tau PET burden in meta-temporal regions (Very low, Low, High, Very high), and MMSE change rate (Stable, Slow, Moderate, Fast).

Not all cohorts cover all thirteen variable groups. For example, HABS-HD, POINTER, and A4 do not include FDG PET imaging, and the availability of blood-based biomarkers and cognitive assessments also varies across studies. This heterogeneity, where different cohorts contribute overlapping but non-identical subsets of variables, makes an LLM-based prediction framework particularly attractive: ADLLM can flexibly serialize the variables observed in each cohort into a natural-language format, without requiring a fully aligned tabular feature space or ad hoc imputation of missing modalities. Further details on the training datasets are provided in Supplementary Material (Lyu, Ma and Wang, 2026).

Through a structured prompt design that (i) encodes each variable group as a set of short, standardized natural-language statements and (ii) casts each prediction target as a constrained multiple-choice generation over pre-specified category tokens, ADLLM produces structured outputs directly compatible with standard multi-class evaluation metrics such as AUC, accuracy, and specificity.

5.2. *Benchmarking methods in comparison* We compare the proposed method against several feature-selection baselines, including LLM-based attribution methods, gradient-boosting feature importance, and differentiable gating approaches. The first baseline is LLM-Rank (Jeong, Lipton and Ravikumar, 2024), a zero-shot feature-selection framework that prompts the LLM to rank variables according to their expected predictive relevance. The second baseline is attention-based feature selection, where we compute accumulated attention scores for all tokens within each variable group across transformer layers and normalize by token length (Wiegreffe and Pinter, 2019). The third baseline uses CatBoost feature importance scores (Dorogush, Ershov and Gulin, 2018), representing a strong non-LLM tabular feature-selection approach. We also added three gating-based baselines that are more directly comparable to the proposed embedding-level selection framework. First, we added a deterministic sigmoid-gate baseline based on continuous optimization on the embedding matrix representation:

$$\boldsymbol{X}(\sigma(\gamma)) = \sigma(\gamma) \odot \boldsymbol{X} + (1 - \sigma(\gamma)) \odot \boldsymbol{H},$$

where $\sigma(\cdot)$ is the sigmoid function, γ are learnable parameters, $\boldsymbol{X}$ is the original embedding representation, and $\boldsymbol{H}$ is the replacement embedding for the padding token. This approach performs continuous optimization without discrete sampling. We further included two stochastic-gating baselines commonly used in differentiable sparsity learning: the hard concrete estimator (Louizos, Welling and Kingma, 2017) and the straight-through Bernoulli estimator (Bengio, Léonard and Courville, 2013). These methods provide direct comparisons against our adopted differentiable discrete-selection strategy. For the sigmoid-gate, hard concrete, and straight-through Bernoulli baselines, the number of selected variable groups is determined using the same GBIC criterion as the proposed method. For the gating-based feature selection (including the proposed), we trained the gating parameters using the Adam optimizer with an initial learning rate of 1×10^{-2} and trained for 25 epochs by default. The learning rate was decayed with a cosine-annealing schedule to a minimum of 1×10^{-5} implemented with PyTorch's CosineAnnealingLR. Training used small mini-batches with default batch size 2. Additional implementation details of these baseline are provided in Supplementary Material Sections 3 (Lyu, Ma and Wang, 2026).

5.3. *Results* We focus on the held-out POINTER imaging subset, consisting of cognitively normal older adults at elevated risk for cognitive decline, making it a suitable setting for evaluating preclinical Aβ and tau pathology prediction. POINTER contains $K = 11$ clinically meaningful variable groups, including genetics, family history, demographics, vital signs, cognitive tests, physical measures, MRI-derived brain volume and thickness, comorbidities, medication use, and cognitive diagnosis. Since only baseline data are available, MMSE change-rate prediction is not considered. To adapt LLM-Rank to POINTER, we query the LLaMA 3.2-3B model using the customized prompt in Supplementary Material Section 6.4 (Lyu, Ma and Wang, 2026).

As shown in Table 1, with $\lambda = 0.057$ selected by GBIC for LLaMA-3.1-8B ($\lambda = 0.01$ for Qwen3-8B), our method selects sparse subsets of clinically meaningful variable groups while masking the remainder. For LLaMA-3.1-8B, the selected groups are *demographics, genetics, family history, vitals, MRI thickness*, and *MRI volume*; for Qwen3-8B, the selected groups additionally include *cognitive tests* and *comorbidities*. This reduces the number of variable groups by nearly 45% and the prompt length from roughly 1,000 to 430 tokens (approximately 57% reduction).

Across both backbones, the proposed framework achieved competitive or strongest overall predictive performance while using substantially fewer input variables. Under LLaMA-3.1-8B, the proposed method achieved the best Aβ prediction performance (AUC = 0.675, 95% CI: 0.590–0.750; specificity = 0.435, 95% CI: 0.345–0.526), improving specificity relative to competing approaches (approximately 0.27–0.28). For meta-tau prediction, the proposed framework again achieved the strongest AUC and specificity (AUC = 0.578, 95% CI: 0.527–0.630; specificity = 0.749, 95% CI: 0.725–0.773). Similar trends were observed under the newer Qwen3-8B backbone, where the proposed method achieved the highest Aβ AUC (0.679, 95% CI: 0.598–0.752) and substantially improved meta-tau accuracy (0.372, 95% CI: 0.298–0.450) relative to competing methods (0.239–0.272). Together, these results suggest that the proposed framework generalizes across different LLM families while reducing prompt complexity.

The selected variable groups were also clinically meaningful and broadly stable across backbone architectures for Aβ and meta tau prediction. Across both backbones, genetics and MRI-derived structural measures received high importance scores, aligning with established AD pathology literature linking APOE-related susceptibility and neurodegeneration markers to amyloid and tau progression (Safieh, Korczyn and Michaelson, 2019; Blumenfeld et al., 2024; Frisoni et al., 2010; Vemuri and Jack Jr, 2010). The variable groups selected by certain benchmarking methods, however, appear less consistent with clinical expectations. Both LLM-Rank (Jeong, Lipton and Ravikumar, 2024) and attention-based selection (Wiegreffe and Pinter, 2019) frequently emphasized cognitive diagnosis, despite the POINTER cohort enrolling only cognitively normal participants. These discrepancies may reflect inherent differences in what each method captures: attention-based scores can be sensitive to the attention-sink effect (Xiao et al., 2023), in which positional or boundary tokens accumulate disproportionate attention regardless of predictive relevance, while gradient-boosted importance scores (e.g., CatBoost Dorogush, Ershov and Gulin 2018) are derived from a representation space distinct from the LLM and may not translate directly to informative variable subsets at the embedding level.

TABLE 1

Comparison of variable selection methods and variable-group importance scores across LLM backbones on AD prediction tasks. Bold entries indicate the best result within each backbone and metric; confidence intervals are at 95% level and are constructed with 2,000 bootstrap iterations.

Backbone	Outcome	Method	AUC	Accuracy	Specificity
LLaMA-3.1-8B	Aβ	Attention	0.625 (0.538–0.701)	0.723 (0.702–0.739)	0.284 (0.253–0.329)
		Gradient Boosting	0.586 (0.504–0.667)	0.718 (0.691–0.739)	0.283 (0.252–0.328)
		LLM Rank	0.622 (0.535–0.702)	0.713 (0.676–0.745)	0.283 (0.251–0.329)
		Hard Concrete	0.658 (0.576–0.735)	0.723 (0.697–0.745)	0.269 (0.252–0.300)
		Sigmoid Gate	0.663 (0.583–0.738)	0.729 (0.702–0.750)	0.269 (0.252–0.301)
		ST-Bernoulli	0.651 (0.572–0.730)	0.723 (0.697–0.745)	0.269 (0.252–0.300)
		All Variable Groups	0.659 (0.579–0.736)	0.729 (0.707–0.750)	0.269 (0.251–0.301)
		Proposed	**0.675 (0.590–0.750)**	**0.734 (0.697–0.766)**	**0.435 (0.345–0.526)**
	Meta-tau	Attention	0.502 (0.451–0.557)	0.267 (0.239–0.294)	0.730 (0.719–0.740)
		Gradient Boosting	0.563 (0.512–0.615)	0.283 (0.233–0.333)	0.703 (0.683–0.722)
		LLM Rank	0.516 (0.464–0.570)	0.289 (0.228–0.350)	0.727 (0.702–0.751)
		Hard Concrete	0.566 (0.514–0.619)	0.311 (0.256–0.367)	0.724 (0.702–0.746)
		Sigmoid Gate	0.549 (0.498–0.599)	0.278 (0.228–0.328)	0.707 (0.686–0.728)
		ST-Bernoulli	0.569 (0.515–0.629)	**0.339 (0.283–0.394)**	0.746 (0.723–0.768)
		All Variable Groups	0.555 (0.501–0.609)	0.300 (0.239–0.361)	0.731 (0.707–0.756)
		Proposed	**0.578 (0.527–0.630)**	0.322 (0.267–0.383)	**0.749 (0.725–0.773)**
	Selected Variable Groups		Demographics (0.843), Genetics (0.920), Family History (0.832), Vitals (0.769), Brain Volume MRI (0.794), Brain Thickness MRI (0.758)		
Qwen3-8B	Aβ	Attention	0.491 (0.400–0.589)	**0.745 (0.745–0.745)**	0.255 (0.255–0.255)
		Gradient Boosting	0.600 (0.520–0.678)	0.713 (0.681–0.745)	0.283 (0.251–0.329)
		LLM Rank	0.617 (0.537–0.697)	0.622 (0.564–0.678)	**0.546 (0.452–0.641)**
		Hard Concrete	0.644 (0.564–0.722)	0.702 (0.665–0.739)	0.267 (0.217–0.322)
		Sigmoid Gate	0.656 (0.581–0.726)	0.686 (0.649–0.723)	0.356 (0.293–0.433)
		ST-Bernoulli	**0.679 (0.598–0.752)**	0.707 (0.676–0.739)	0.343 (0.281–0.419)
		All Variable Groups	0.646 (0.564–0.726)	0.713 (0.686–0.734)	0.268 (0.251–0.299)
		Proposed	**0.679 (0.598–0.752)**	0.707 (0.676–0.739)	0.343 (0.281–0.419)
	Meta-tau	Attention	0.477 (0.426–0.525)	0.272 (0.250–0.294)	0.728 (0.720–0.736)
		Gradient Boosting	0.519 (0.470–0.570)	0.256 (0.206–0.311)	0.734 (0.716–0.754)
		LLM Rank	0.480 (0.442–0.523)	0.256 (0.211–0.300)	**0.768 (0.753–0.784)**
		Hard Concrete	0.542 (0.493–0.590)	0.239 (0.183–0.300)	0.738 (0.717–0.758)
		Sigmoid Gate	0.535 (0.486–0.581)	0.261 (0.211–0.317)	0.740 (0.722–0.760)
		ST-Bernoulli	0.542 (0.493–0.590)	0.267 (0.217–0.322)	0.740 (0.722–0.760)
		All Variable Groups	0.533 (0.481–0.584)	0.272 (0.228–0.317)	0.741 (0.726–0.757)
		Proposed	**0.556 (0.505–0.606)**	**0.372 (0.298–0.450)**	0.740 (0.722–0.760)
	Selected Variable Groups		Demographics (0.844), Genetics (0.840), Vitals (0.953), Cognitive Tests (0.953), Brain Volume MRI (0.941), Brain Thickness MRI (0.890), Comorbidities (0.555)		

6. Case study II: What are important features for predicting HIV treatment outcomes and care disengagement? To evaluate the proposed framework beyond AD and neuroimaging settings, we conducted a second case study using a subset of the large-scale HIV electronic medical record (EMR) data from the Tanzanian national Care and Treatment Clinic (CTC3) repository. The HIV setting provides a clinically meaningful testbed for LLM-based feature selection: HIV care systems in sub-Saharan Africa frequently operate under severe resource constraints, including limited staffing, laboratory testing capacity, and outreach resources for longitudinal follow-up. Sustained retention in care and adherence to antiretroviral therapy (ART) remain critical for achieving viral suppression and reducing onward transmission, yet proactively identifying patients at high risk of ART non-adherence, viral non-suppression, or loss-to-follow-up (LTFU) remains challenging in large-scale public-health systems (Wei et al., 2026). Understanding which variable groups most inform LLM-based predictions is therefore clinically valuable, as it can guide data prioritization, reduce operational burden, and help allocate limited intervention resources toward the highest-risk individuals.

The full Tanzanian CTC3 repository spans January 2018 to June 2023 and contains over 4.8 million longitudinal visit records from more than 261,000 people living with HIV (PLHIV). Unlike the AD case study, which relies heavily on multimodal biomarker and neuroimaging data, the HIV cohort consists primarily of longitudinal EMR data collected during routine clinical care, including demographics, visit histories, ART adherence records, laboratory biomarkers, and facility-level information. This setting therefore allows us to evaluate whether the proposed framework remains effective in a substantially different clinical environment, characterized by high-dimensional longitudinal records in a large-scale public-health context.

Following the participatory predictor-selection procedure of (Wei et al., 2026), we identify 21 clinically meaningful variables suitable for real-world deployment, covering demographics, clinical visits, medical history, therapy, physical condition, clinical stage, immunological and virological status, health facility information, and static clinical information. Additional details on predictors and serialization are provided in Supplementary Material (Lyu, Ma and Wang, 2026).

6.1. *Models finetuning and feature selection* For computational feasibility, we fine-tuned LLaMA-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) on an adult subgroup from the Tanzanian CTC3 repository, using longitudinal EMRs spanning 2018–2021. The cohort was restricted to adults aged 26–42 years from Kyerwa DC in the Kagera region with sufficient longitudinal information, including viral-load measurements, visit records, nutritional status, and demographic variables, resulting in 1,727 PLHIV and 20,970 visit records. We focused on this subgroup because these individuals remain actively engaged in HIV care while facing substantial risks of future disengagement due to treatment fatigue, work and family responsibilities, and comorbidities.

Using these EMRs, we predicted three future outcomes during 2022–2023: ART non-adherence risk, viral non-suppression (viral load >1000 copies/mL), and loss to follow-up (LTFU; missing a scheduled visit by at least 28 days, following the PEPFAR definition (Wei et al., 2026)). The resulting fine-tuned model is referred to as **HIV-LLM**. The prediction task was formulated as a structured question-answering problem using serialized longitudinal EMRs and fixed-format prompt outputs: "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."

To evaluate feature importance under distributional shift and within a clinically distinct population, we additionally conducted an exploratory pilot study using records from 500 pregnant women in Nyang'hwale DC in the Geita region, which is geographically distinct from the primary training cohort in Kagera. This setting was chosen because pregnancy represents a high-priority period in HIV care: maternal viral suppression and sustained ART adherence are essential not only for maternal health, but also for preventing vertical HIV transmission during pregnancy, delivery, and breastfeeding (Newell, 2001; Abuogi et al., 2018). The pregnancy subgroup also differs from the general adult HIV treatment population. Pregnant women often experience different visit schedules, changing adherence, and retention challenges during the antenatal and postpartum periods (Psaros et al., 2015; Harris and Yudin, 2020). These differences make the subgroup a clinically meaningful setting for evaluating whether feature-selection patterns learned by HIV-LLM remain interpretable under realistic deployment shifts.

6.2. *Results* Table 2 summarizes the HIV real-data experiments under both the LLaMA-3.1-8B and Qwen3-8B backbones. Overall, the HIV prediction tasks were more challenging than the AD tasks. This is expected because HIV-LLM was fine-tuned on a broad adult treatment population from Kagera, whereas feature selection was evaluated on pregnant women from Nyang'hwale DC in Geita. Moreover, since adherence, retention, and monitoring needs during pregnancy and postpartum differ from those of the general adult HIV population (Psaros et al., 2015; Harris and Yudin, 2020), this pilot study is best viewed as an evaluation of clinically meaningful variable selection under distributional shift, rather than state-of-the-art HIV prediction performance.

Despite this challenging setting, the proposed framework achieved competitive or strongest overall performance across viral load, ART non-adherence, and LTFU prediction while selecting sparse variable subsets rather than using the full serialized prompt. Under LLaMA-3.1-8B, the proposed method achieved the highest AUC for all three outcomes: viral-load prediction (AUC 0.529, 95% CI: 0.491–0.566), LTFU prediction (AUC 0.543, 95% CI: 0.491–0.593), and ART non-adherence prediction (AUC 0.507, 95% CI: 0.492–0.522). Although these AUC values are modest in absolute terms, they are strongest among the compared methods in this shifted pregnancy cohort. This suggests that the proposed feature-selection procedure can preserve or improve predictive signal even when the target population differs from the training population. Under Qwen3-8B, the proposed framework achieved the strongest viral-load prediction performance (AUC 0.575, 95% CI: 0.535–0.612) and the highest ART non-adherence AUC (0.526, 95% CI: 0.501–0.552), while remaining competitive for LTFU prediction. These findings indicate that sparse prompt selection can be beneficial not only for reducing input complexity, but also for improving robustness when applying healthcare LLMs to clinically specialized subgroups.

The selected variables provide clinically interpretable evidence for what HIV-LLM relied on in the pregnant women subgroup. Under LLaMA-3.1-8B, the proposed framework selected variables related to pregnancy status, scheduled appointments, client category, ARV pill prescription history, ART status, days not on ART, days since HIV diagnosis, TB TPT status, facility type, HIV testing or referral source, weight, nutritional status, and viral load. These variables are clinically coherent because they capture treatment continuity, adherence burden, longitudinal care engagement, maternal health status, and virological monitoring. In particular, ART status, days not on ART, and ARV prescription history directly reflect treatment interruption and adherence patterns, which are closely related to viral suppression and future disengagement from care (Adams et al., 2015; Young et al., 2024).

Qwen3-8B selected a related but slightly broader set of variables, including demographics, marital status, pregnancy status, visit dates, scheduled appointments, visit type, client category, TB TPT status, ART status, weight, WHO stage, CD4 status, and HIV testing or referral source. Compared with LLaMA-3.1-8B, Qwen3-8B placed more emphasis on disease severity and social-context variables, such as WHO stage, CD4 status, demographics, and marital status. This pattern is clinically plausible because immunological status and WHO stage reflect HIV disease progression, while marital status and demographic factors may proxy social support, care access, and retention risk (Teshale et al., 2025; Ojukwu et al., 2023).

Importantly, several variables were selected across both backbones, including pregnancy status, scheduled appointments, client category, TB TPT status, ART status, weight, and HIV testing or referral source. This overlap suggests that the proposed framework identifies stable and clinically coherent predictors rather than architecture-specific artifacts. These shared variables align with PMTCT (prevention of mother-to-child transmission) and HIV-care priorities, including intensified maternal monitoring, ART continuity, comorbidity burden, care-entry pathways, and long-term engagement with the treatment system (Woldesenbet et al., 2020; Teshale et al., 2025; Adams et al., 2015; Young et al., 2024).

7. Simulation studies In this section, we conduct simulation studies using synthetic tabular data generated from parameters calibrated to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) diabetes dataset. The original dataset contains 768 individuals with eight predictors: age, number of pregnancies, diastolic blood pressure, triceps skin-fold thickness, 2-hour plasma glucose concentration, 2-hour serum insulin, body mass index (BMI), and diabetes pedigree function (DPF), together with a binary diabetes diagnosis

TABLE 2

Comparison of variable selection methods and variable-group importance scores across LLM backbones on HIV-related prediction tasks. Bold entries indicate the best result within each backbone and metric; confidence intervals are at 95% level and are constructed with 2,000 bootstrap iterations.

Backbone	Outcome	Method	AUC	Accuracy	Specificity
LLaMA-3.1-8B	Viral Load	Attention	0.519 (0.481–0.555)	0.438 (0.410–0.464)	0.669 (0.654–0.683)
		Gradient Boosting	0.489 (0.450–0.526)	0.392 (0.362–0.422)	0.646 (0.629–0.663)
		LLM Rank	0.497 (0.459–0.533)	0.394 (0.362–0.428)	0.648 (0.630–0.666)
		Hard Concrete	0.513 (0.473–0.550)	0.440 (0.416–0.462)	0.668 (0.655–0.681)
		Sigmoid Gate	0.527 (0.489–0.564)	0.430 (0.400–0.456)	0.665 (0.648–0.680)
		ST-Bernoulli	0.513 (0.473–0.550)	**0.444 (0.414–0.474)**	**0.675 (0.658–0.691)**
		All Variable Groups	0.526 (0.489–0.562)	0.420 (0.386–0.450)	0.661 (0.643–0.678)
		Proposed	**0.529 (0.491–0.566)**	**0.444 (0.414–0.474)**	**0.675 (0.658–0.691)**
	LTFU	Attention	0.520 (0.470–0.572)	0.400 (0.374–0.426)	0.505 (0.476–0.532)
		Gradient Boosting	0.492 (0.439–0.544)	0.402 (0.374–0.432)	0.500 (0.469–0.530)
		LLM Rank	0.536 (0.484–0.587)	0.402 (0.376–0.428)	0.506 (0.479–0.533)
		Hard Concrete	0.538 (0.488–0.593)	**0.418 (0.388–0.448)**	0.512 (0.480–0.543)
		Sigmoid Gate	0.513 (0.460–0.565)	**0.418 (0.386–0.454)**	0.501 (0.465–0.537)
		ST-Bernoulli	0.539 (0.488–0.593)	**0.418 (0.388–0.448)**	0.512 (0.480–0.543)
		All Variable Groups	0.538 (0.487–0.589)	0.414 (0.382–0.444)	0.508 (0.476–0.539)
		Proposed	**0.543 (0.491–0.593)**	0.410 (0.386–0.434)	**0.522 (0.499–0.544)**
	ART Non-adherence	Attention	0.498 (0.481–0.514)	0.340 (0.322–0.356)	0.797 (0.793–0.802)
		Gradient Boosting	0.496 (0.477–0.516)	0.334 (0.314–0.354)	0.797 (0.791–0.802)
		LLM Rank	0.499 (0.481–0.516)	0.338 (0.320–0.354)	0.797 (0.792–0.801)
		Hard Concrete	0.492 (0.471–0.512)	0.344 (0.322–0.366)	0.800 (0.794–0.806)
		Sigmoid Gate	0.498 (0.476–0.519)	0.326 (0.300–0.350)	0.797 (0.790–0.804)
		ST-Bernoulli	0.492 (0.471–0.512)	0.344 (0.322–0.366)	0.800 (0.794–0.806)
		All Variable Groups	0.506 (0.486–0.524)	0.330 (0.308–0.350)	0.797 (0.791–0.802)
		Proposed	**0.507 (0.492–0.522)**	**0.364 (0.348–0.380)**	**0.803 (0.799–0.808)**
	Selected Variables	Pregnancy status, Scheduled appointments, Client category, Number of ARV pills prescribed on each visit date, ART status, Number of days not on ART, The number of days since HIV diagnosis, TB TPT status, Health facility type, HIV testing or referral source, Weight, Nutrition status, Viral load			
Qwen3-8B	Viral Load	Attention	0.500 (0.462–0.539)	0.442 (0.432–0.450)	0.665 (0.660–0.669)
		Gradient Boosting	0.549 (0.512–0.589)	0.446 (0.434–0.458)	0.668 (0.662–0.675)
		LLM Rank	0.556 (0.517–0.594)	0.428 (0.404–0.454)	0.663 (0.650–0.677)
		Hard Concrete	0.458 (0.422–0.494)	0.430 (0.416–0.444)	0.662 (0.655–0.668)
		Sigmoid Gate	0.482 (0.445–0.519)	0.372 (0.340–0.404)	0.643 (0.626–0.660)
		ST-Bernoulli	0.492 (0.454–0.529)	0.438 (0.426–0.450)	0.664 (0.659–0.670)
		All Variable Groups	0.549 (0.513–0.585)	0.432 (0.404–0.458)	0.666 (0.651–0.681)
		Proposed	**0.575 (0.535–0.612)**	**0.460 (0.446–0.474)**	**0.674 (0.665–0.683)**
	LTFU	Attention	0.500 (0.450–0.556)	0.370 (0.352–0.390)	0.488 (0.467–0.508)
		Gradient Boosting	0.520 (0.468–0.575)	**0.406 (0.376–0.438)**	**0.502 (0.470–0.534)**
		LLM Rank	0.477 (0.423–0.530)	0.364 (0.358–0.368)	0.495 (0.486–0.500)
		Hard Concrete	**0.532 (0.477–0.584)**	0.374 (0.364–0.384)	**0.502 (0.492–0.513)**
		Sigmoid Gate	0.516 (0.467–0.570)	0.384 (0.364–0.406)	0.498 (0.475–0.521)
		ST-Bernoulli	0.504 (0.451–0.555)	0.368 (0.362–0.374)	0.499 (0.492–0.505)
		All Variable Groups	0.504 (0.452–0.556)	0.384 (0.362–0.408)	0.494 (0.469–0.520)
		Proposed	0.512 (0.460–0.567)	0.374 (0.362–0.388)	**0.502 (0.487–0.514)**
	ART Non-adherence	Attention	0.480 (0.458–0.503)	0.354 (0.338–0.368)	0.798 (0.794–0.802)
		Gradient Boosting	0.511 (0.489–0.535)	0.352 (0.330–0.378)	**0.802 (0.796–0.809)**
		LLM Rank	0.491 (0.466–0.519)	0.364 (0.358–0.368)	0.799 (0.797–0.800)
		Hard Concrete	0.509 (0.495–0.524)	**0.370 (0.362–0.378)**	0.801 (0.799–0.804)
		Sigmoid Gate	0.515 (0.493–0.539)	0.356 (0.338–0.372)	0.800 (0.795–0.805)
		ST-Bernoulli	0.508 (0.485–0.530)	0.366 (0.362–0.368)	0.800 (0.798–0.800)
		All Variable Groups	0.504 (0.481–0.528)	0.360 (0.340–0.380)	0.801 (0.796–0.807)
		Proposed	**0.526 (0.501–0.552)**	**0.370 (0.358–0.378)**	**0.802 (0.798–0.804)**
	Selected Variables	Demographics, Marital status, Pregnancy status, Visit dates, Scheduled appointments, Visit type, Client category, TB TPT status, ART status, Weight, WHO stage, CD4 status, HIV testing or referral source			

outcome. The low-dimensional eight-variable setting allows an exhaustive 2^8 combinatorial search over all possible variable subsets, which we use as an oracle benchmark for studying feature selection in fine-tuned LLMs.

Our goal is to investigate whether the feature selection behavior observed in the real-data analyses (ADLLM and HIV-LLM) also arises in a controlled setting where the data-generating process (DGP) and oracle sparsity structure are known. The exhaustive subset search further allows us to directly compare selected subsets against the globally optimal subsets for the fine-tuned LLM. In particular, the simulation studies demonstrate three main findings. First, **LLMs benefit from sparse prompt input.** When the tabular DGP is sparse, the fine-tuned LLM achieves its best predictive performance on low-dimensional subsets of variables. Second, **LLM feature importance is related to, but distinct from, tabular DGP sparsity.** Variables that are crucial under the true tabular DGP do not fully align with the subsets most effective for the fine-tuned LLM, suggesting that prompt serialization and pretrained language representations reshape feature importance while still preserving some truly informative variables.

Third, **recovery of high-performing LLM subsets is feasible.** Compared with heuristic and classical baseline approaches, the proposed framework reliably recovers the oracle-optimal subsets and achieves the strongest predictive performance in this controlled setting.

7.1. *Synthetic NIDDK study in Monte Carlo simulations* To construct realistic yet fully controlled simulation data, we model the joint distribution of the eight NIDDK predictors using a generative adversarial network (GAN) (Goodfellow et al., 2020), whose generator and discriminator each comprise multilayer perceptrons with hidden dimensions up to 1024. The model is trained using the Adam optimizer with binary cross-entropy loss for 1,000 epochs and includes a correlation-penalization term to discourage spurious feature correlations. This calibration step allows us to generate arbitrarily many synthetic individuals preserving the marginal distributions and dependence structure of the original cohort, while using clinically meaningful variable names that can be serialized into natural-language prompts and leverage the LLM's pretrained medical knowledge. Details of GAN training are provided in Supplementary Material (Lyu, Ma and Wang, 2026).

We then generate 102 Monte Carlo (MC) samples, each consisting of 768 synthetic individuals with the same eight covariates. The binary outcome in each synthetic dataset is generated from a sparse logistic regression model:

$$\texttt{Logit}\left(\widehat{\mathbb{P}}(Y=1)\right) = 0.117\,\text{Pregnancies} + 0.028\,\text{Glucose} + 0.060\,\text{BMI} + 0.678\,\text{DPF} - 5.893,$$

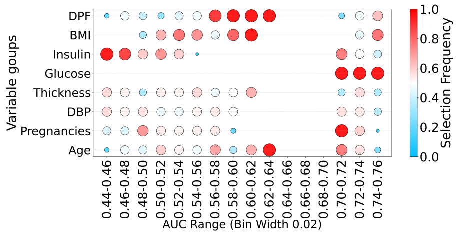
where the coefficients are estimated by fitting logistic regression to the original NIDDK data and are treated as the true data-generating process (DGP).

After serializing the tabular data across all MC samples, we designate one MC sample as $\mathcal{D}_{\texttt{train}}$ to fine-tune LLaMA-3.1-8B on serialized prompts for diabetes status prediction, yielding the "diabetes-LLM," one MC sample as $\mathcal{D}_{\texttt{oracle}}$ to investigate sparsity patterns in the diabetes-LLM input space via exhaustive combinatorial search and to benchmark selected variable subsets, and the remaining 100 MC samples as test sets $\{\mathcal{D}^{(m)}_{\texttt{test}}\}_{m=1}^{100}$ for feature-selection comparisons. A serialized prompt example for this domain-specific LLM is provided in Supplementary Material (Lyu, Ma and Wang, 2026).

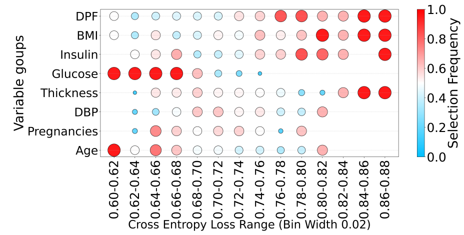
7.2. *Does sparsity exist in fine-tuned LLM?* Before presenting the main results, we first investigate how sparsity in the tabular DGP changes after covariates are serialized into prompts for a fine-tuned LLM. In particular, we ask whether the diabetes-LLM achieves its best predictive performance on the true DGP support or on a different subset induced by the serialized prompt representation.

To study this, we exhaustively evaluate all 2^8 subsets of variables in $\mathcal{D}_{\texttt{oracle}}$, replacing excluded variables with padding tokens in the prompt. For each subset, we record the next-token prediction loss and AUC. Figure 3 summarizes how often each variable appears among subsets achieving different performance ranges. Variables such as *Glucose* and *DPF* appear most frequently in high-performing subsets, while variables such as *Skin Thickness* and *Pregnancies* contribute less consistently. The subset (*Age*, *Glucose*) achieves the lowest cross-entropy, whereas (*Glucose*, *BMI*, *DPF*) achieves the highest AUC.

Interestingly, the best-performing subsets are low-dimensional, involving only two or three variables, and do not exactly match the true DGP support (*Pregnancies*, *Glucose*, *BMI*, *DPF*). At the same time, there remains overlap between the oracle DGP variables and the important variables selected by the LLM. These results suggest that serialization and pretrained language representations can reshape feature importance in prompt space while still retaining some predictive signals from the original tabular DGP.



(a) AUC-based selection frequency.



(b) Cross-entropy-based selection frequency.

Fig 3: Variable selection frequency across fine-grained AUC (Left) and cross-entropy (Right) performance ranges on the validation dataset. Each circle shows the proportion of variable subsets falling in a performance range that include the given variable, computed over all 2^8 combinations.

7.3. *Can Existing Feature-Selection Methods Recover Oracle LLM Subsets?* To evaluate whether existing feature-selection strategies can recover the subsets that are truly effective for LLM inference, we compare our framework against both LLM-based heuristics and classical statistical feature-selection methods.

We first consider two commonly used LLM-oriented baselines: (i) LLM-Rank (Jeong, Lipton and Ravikumar, 2024), which queries an off-the-shelf LLaMA-3.2-3B model to rank variables by perceived predictive importance, and (ii) an attention-based approach, which ranks variables using accumulated transformer attention weights. For each Monte Carlo replication, we evaluate top-k subsets ($k = 2, \ldots, 6$) selected by each method on the oracle-selection dataset $\mathcal{D}_{\texttt{oracle}}$.

Figure 4 shows that our method achieves the highest and most stable AUC across most subset sizes. Even with only two or three variables, our framework consistently attains AUCs near 0.74, whereas LLM-Rank exhibits larger variability and the attention baseline performs substantially worse. These results suggest that neither prompting-based ranking nor raw attention weights reliably recover the subsets most useful for fine-tuned LLM inference. We next

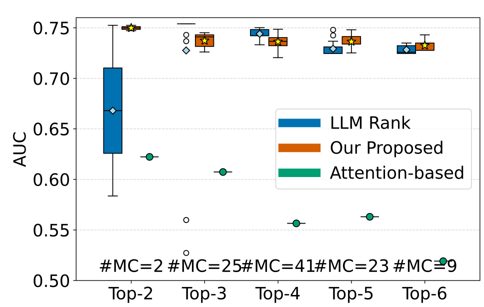
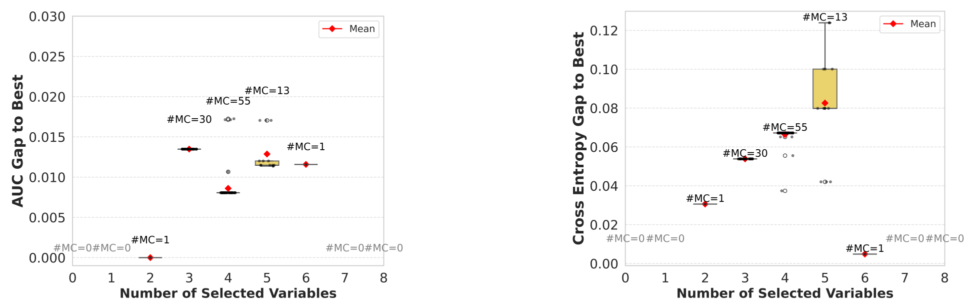


Fig 4: Comparison of AUC across variable subsets selected by different methods over 100 Monte Carlo replications. Each boxplot shows the AUC distribution on $\mathcal{D}_{\texttt{oracle}}$ for the top-k variables ($k = 2, \ldots, 6$) selected by each method. The proposed framework (orange) selects k adaptively via GBIC; LLM-Rank (blue; 100 queries) and the attention-based baseline (green) use matching subset sizes.

compare against a classical tabular-data approach by fitting logistic regression models directly on the original covariates and selecting variables with statistically significant coefficients

($p < 0.05$). The selected subsets are then serialized back into prompts and evaluated using the diabetes-LLM. Figure 5 compares these subsets against the optimal subsets of the same size identified through exhaustive search. Although logistic regression is correctly specified for the



(a) AUC gap.

(b) Cross-entropy gap.

Fig 5: LLM performance gap between variables selected via logistic regression p-values and the best-performing subset of the same size from exhaustive search. Red diamonds indicate the mean gap per subset size; subset sizes with no Monte Carlo samples are marked #MC $= 0$.

tabular DGP, they underperform the oracle LLM subsets in both AUC and cross-entropy. This gap indicates that sparsity in the original tabular DGP does not translate directly into the sparsity structure most effective for LLM inference after serialization. One possible explanation is that classical parametric models operate directly in tabular feature space, whereas fine-tuned LLMs additionally rely on pretrained language representations and prompt-level interactions.

Overall, these experiments suggest that both heuristic LLM interpretations and classical statistical feature-selection methods fail to fully recover the subsets most effective for serialized LLM inference, motivating the need for LLM-aware feature-selection strategies.

8. Significance statement Large language models (LLMs) are increasingly applied as predictive engines across diverse domains, taking as input serialized prompts that combine heterogeneous variable groups. It remains unclear which groups are necessary for accurate prediction: a question with direct implications for data collection, computational cost, and interpretability. This paper fills this gap by introducing a variable-group selection framework tailored to the LLM inference pipeline. Specifically, it performs selection directly in the LLM embedding space, where unselected groups are replaced by padding-token embeddings. Combined with a Bernoulli relaxation of the discrete selection variables and an approximate gradient estimator, this enables efficient gradient-based optimization. The framework is model-agnostic, provides variable-importance scores, and comes with theoretical generalization guarantees.

We demonstrate the method on Alzheimer's disease pathology prediction using multimodal clinical data and HIV treatment outcome prediction using large-scale longitudinal electronic medical records. The framework identifies sparse, clinically meaningful variable subsets that maintain or improve predictive performance while substantially reducing prompt length. More broadly, our results show that feature importance for LLM-based prediction can differ from classical tabular importance owing to prompt serialization and pretrained language representations, motivating LLM-aware feature-selection strategies across applications where input efficiency and interpretability matter.

SUPPLEMENTARY MATERIAL

Supplementary Material includes all additional details and proofs.

REFERENCES

ABUOGI, L. L., HUMPHREY, J. M., MPODY, C., YOTEBIENG, M., MURNANE, P. M., CLOUSE, K., OTIENO, L., COHEN, C. R. and WOOLS-KALOUSTIAN, K. (2018). Achieving UNAIDS 90-90-90 targets for pregnant and postpartum women in sub-Saharan Africa: progress, gaps and research needs. *Journal of virus eradication* **4** 33–39.

ADAMS, J. W., BRADY, K. A., MICHAEL, Y. L., YEHIA, B. R. and MOMPLAISIR, F. M. (2015). Postpartum engagement in HIV care: an important predictor of long-term retention in care and viral suppression. *Clinical Infectious Diseases* **61** 1880–1887.

BENGIO, Y., LÉONARD, N. and COURVILLE, A. (2013). Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. arXiv:1308.3432.

BLUMENFELD, J., YIP, O., KIM, M. J. and HUANG, Y. (2024). Cell Type-Specific Roles of APOE4 in Alzheimer Disease. *Nature Reviews Neuroscience* **25** 91–110.

CROXFORD, E., GAO, Y., FIRST, E., PELLEGRINO, N., SCHNIER, M., CASKEY, J. et al. (2025). Evaluating Clinical AI Summaries With Large Language Models as Judges. *npj Digital Medicine* **8** 640.

DAVIS, J., SOUNACK, T., SCIACCA, K., BRAIN, J. M., DURIEUX, B. N., AGARONNIK, N. D. et al. (2026). MedSlice: Fine-Tuned Large Language Models for Secure Clinical Note Sectioning. *JAMIA Open* **9** ooaf179.

DOROGUSH, A. V., ERSHOV, V. and GULIN, A. (2018). CatBoost: Gradient Boosting With Categorical Features Support. arXiv:1810.11363.

FAN, J. and LI, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96** 1348–1360.

FRISONI, G. B., FOX, N. C., JACK JR, C. R., SCHELTENS, P. and THOMPSON, P. M. (2010). The Clinical Use of Structural MRI in Alzheimer Disease. *Nature Reviews Neurology* **6** 67–77.

GANJDANESH, A., SHIRKAVAND, R., GAO, S. and HUANG, H. (2024). Not All Prompts Are Made Equal: Prompt-Based Pruning of Text-to-Image Diffusion Models. arXiv:2406.12042.

GOLDSHMIDT, R. and HOROVICZ, M. (2024). TokenSHAP: Interpreting Large Language Models With Monte Carlo Shapley Value Estimation. arXiv:2407.10114.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S. et al. (2020). Generative Adversarial Networks. *Communications of the ACM* **63** 139–144.

GRATTAFIORI, A., DUBEY, A., JAUHRI, A., PANDEY, A., KADIAN, A., AL-DAHLE, A. et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783.

GUO, Z., KAMIGAITO, H. and WATANABE, T. (2024). Attention Score Is Not All You Need for Token Importance Indicator in KV Cache Reduction: Value Also Matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 21158–21166.

HARRIS, K. and YUDIN, M. H. (2020). HIV infection in pregnant women: a 2020 update. *Prenatal Diagnosis* **40** 1715–1721.

HARRISON, T. M., WARD, T., TAGGETT, J., MAILLARD, P., LOCKHART, S. N., JUNG, Y. et al. (2025). The POINTER Imaging Baseline Cohort: Associations Between Multimodal Neuroimaging Biomarkers, Cardiovascular Health, and Cognition. *Alzheimer's & Dementia* **21** e14399.

HEGSELMANN, S., BUENIDA, A., LANG, H., AGRAWAL, M., JIANG, X. and SONTAG, D. (2023). TabLLM: Few-Shot Classification of Tabular Data With Large Language Models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics* 5549–5581.

HU, E. J., SHEN, Y., WALLIS, P., ALLEN-ZHU, Z., LI, Y., WANG, S. et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

JAIN, S. and WALLACE, B. C. (2019). Attention Is Not Explanation. In *Proceedings of NAACL-HLT 2019* 3543–3556.

JEONG, D. P., LIPTON, Z. C. and RAVIKUMAR, P. (2024). LLM-Select: Feature Selection With Large Language Models. arXiv:2407.02694.

JIN, M., YU, Q., ZHANG, C., SHU, D., ZHU, S., DU, M. et al. (2024). Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model. arXiv:2402.00746.

LEE, S. A., WU, A. and CHIANG, J. N. (2025). Clinical ModernBERT: an Efficient and Long Context Encoder for Biomedical Text. arXiv:2504.03964.

LIU, L., DONG, C., LIU, X., YU, B. and GAO, J. (2023). Bridging Discrete and Backpropagation: Straight-Through and Beyond. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* 12291–12311.

LOUIZOS, C., WELLING, M. and KINGMA, D. P. (2017). Learning Sparse Neural Networks Through L_0 Regularization. arXiv:1712.01312.

LUNDBERG, S. M. and LEE, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing System* 4768–4777.

LYU, Q., MA, X. and WANG, J. (2026). Supplement to "Feature Importance and Selection in Domain Fine-Tuned Language Models With Semi-Structured Data".

MAITY, S. and SAIKIA, M. J. (2025). Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **12** 631.

MIENYE, I. D., OBAIDO, G., JERE, N., MIENYE, E., ARULEBA, K., EMMANUEL, I. D. et al. (2024). A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges. *Informatics in Medicine Unlocked* **51** 101587.

MUELLER, S. G., WEINER, M. W., THAL, L. J., PETERSEN, R. C., JACK, C., JAGUST, W. et al. (2005). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics* **15** 869–877.

NAVED, B. A., RAVISHANKAR, S., COLBERT, G. E., JOHNSTON, A., SLOTT, Q. M. and LUO, Y. (2025). LLM Enabled Classification of Patient Self-Reported Symptoms and Needs in Health Systems Across the USA. *npj Digital Medicine* **8** 390.

NEWELL, M.-L. (2001). Prevention of mother-to-child transmission of HIV: challenges for the current decade. *Bulletin of the World Health Organization* **79** 1138–1144.

NORDBERG, A. (2004). PET Imaging of Amyloid in Alzheimer's Disease. *The Lancet Neurology* **3** 519–527.

OJUKWU, E. N., CIANELLI, R., RODRIGUEZ, N. V., GATTAMORTA, K., DE OLIVEIRA, G. and DUTHELY, L. (2023). Predictors and social determinants of HIV treatment engagement among post-partum Black women living with HIV in southeastern United States. *Journal of advanced nursing* **79** 4365–4380.

OKAMURA, N., HARADA, R., FURUMOTO, S., ARAI, H., YANAI, K. and KUDO, Y. (2014). Tau PET Imaging in Alzheimer's Disease. *Current Neurology and Neuroscience Reports* **14** 500.

PETERSEN, M. E., ZHOU, Z., HALL, J. R., PHILLIPS, N., MEEKER, K. L., BORZAGE, M. T. et al. (2025). Health and Aging Brain Study–Health Disparities (HABS-HD) Methods and Partner Characteristics. *Alzheimer's & Dementia* **11** e70140.

PSAROS, C., REMMERT, J. E., BANGSBERG, D. R., SAFREN, S. A. and SMIT, J. A. (2015). Adherence to HIV care after pregnancy among women in sub-Saharan Africa: falling off the cliff of the treatment cascade. *Current HIV/AIDS Reports* **12** 1–5.

RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144.

SAFIEH, M., KORCZYN, A. D. and MICHAELSON, D. M. (2019). ApoE4: an Emerging Therapeutic Target for Alzheimer's Disease. *BMC Medicine* **17** 64.

SPERLING, R. A., RENTZ, D. M., JOHNSON, K. A., KARLAWISH, J., DONOHUE, M., SALMON, D. P. et al. (2014). The A4 Study: Stopping AD Before Symptoms Begin? *Science Translational Medicine* **6** 228fs13.

SUNDARARAJAN, M., TALY, A. and YAN, Q. (2017). Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning* 3319–3328.

TANG, X., XUE, F. and QU, A. (2021). Individualized Multidirectional Variable Selection. *Journal of the American Statistical Association* **116** 1280–1296.

TESHALE, M. Y., SPIGT, M., HUSSEN, S., MELES, G. G., AJEMA, D., CRUTZEN, R. and STUTTERHEIM, S. E. (2025). Determinants of loss to follow-up among women on option B+ in Southern Ethiopia, 2014–2022: a cohort study. *Scientific Reports* **15** 24980.

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58** 267–288.

VASWANI, A. (2017). Attention Is All You Need. 6000–6010.

VEMURI, P. and JACK JR, C. R. (2010). Role of Structural MRI in Alzheimer's Disease. *Alzheimer's Research & Therapy* **2** 23.

WANG, W., WU, S., ZHU, Z., ZHOU, L. and SONG, P. X. (2024). Supervised Homogeneity Fusion: A Combinatorial Approach. *Annals of Statistics* **52** 285.

WANG, J., AHN, S., DALAL, T., ZHANG, X., PAN, W., ZHANG, Q. et al. (2025). Augmented Risk Prediction for the Onset of Alzheimer's Disease From Electronic Health Records With Large Language Models. *Innovation in Aging* **9** 357.

WEI, W., SHAO, J., LYU, R. Q., HEMONO, R., MA, X., GIORGIO, J. et al. (2026). Enhanced Language Models for Predicting and Understanding HIV Care Disengagement: A Case Study in Tanzania. *npj Digital Medicine* **9** 165.

WIEGREFFE, S. and PINTER, Y. (2019). Attention Is Not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* 11–20.

WOLDESENBET, S. A., KUFA, T., BARRON, P., CHIROMBO, B. C., CHEYIP, M., AYALEW, K., LOMBARD, C., MANDA, S., DIALLO, K., PILLAY, Y. et al. (2020). Viral suppression and factors associated with failure to achieve viral suppression among pregnant women in South Africa. *Aids* **34** 589–597.

WYSOCKI, O., WYSOCKA, M., CARVALHO, D., BOGATU, A. T., GUSICUMA, D. M., DELMAS, M. et al. (2024). An LLM-Based Knowledge Synthesis and Scientific Reasoning Framework for Biomedical Discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 355–364.

XIAO, G., TIAN, Y., CHEN, B., HAN, S. and LEWIS, M. (2023). Efficient Streaming Language Models With Attention Sinks. arXiv:2309.17453.

YAMADA, Y., LINDENBAUM, O., NEGAHBAN, S. and KLUGER, Y. (2020). Feature Selection Using Stochastic Gates. In *Proceedings of the 37th International Conference on Machine Learning* 10648–10659.

YANG, A., LI, A., YANG, B., ZHANG, B., HUI, B., ZHENG, B. et al. (2025). Qwen3 Technical Report. arXiv:2505.09388.

YOUNG, C. M., CHANG, C. A., SAGAY, A. S., IMADE, G., OGUNSOLA, O. O., OKONKWO, P. and KANKI, P. J. (2024). Antiretroviral therapy retention, adherence, and clinical outcomes among postpartum women with HIV in Nigeria. *Plos one* **19** e0302920.

YUAN, M. and LIN, Y. (2006). Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68** 49–67.

ZHANG, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics* **38** 894–942.

ZHOU, W. and MILLER, T. A. (2024). Generalizable Clinical Note Section Identification With Large Language Models. *JAMIA Open* **7** ooae075.

ZHOU, J. and QU, A. (2012). Informative Estimation and Selection of Correlation Structure for Longitudinal Data. *Journal of the American Statistical Association* **107** 701–710.

ZHU, Y., WANG, Z., GAO, J., TONG, Y., AN, J., LIAO, W. et al. (2024). Prompting Large Language Models for Zero-Shot Clinical Prediction with Structured Longitudinal Electronic Health Record Data. arXiv:2402.01713.

ZOU, H. and HASTIE, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67** 301–320.