

SUPPLEMENTARY MATERIAL FOR “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, ^aqiuran_lyu@berkeley.edu; ^bjingshenwang@berkeley.edu*

²*Department of Economics, University of California San Diego, ^cx1ma@ucsd.edu*

1. Brief introduction to LLM inference This section provides a concise overview of the language model inference pipeline, serving as a complement to the discussion presented in the main text.

1.1. Input processing and embedding representation Modern large language models (LLMs) operate on tokenized text rather than raw variables. Given a serialized input prompt $\mathbf{x}$ (constructed as described in Section 2 of the main manuscript), the tokenizer decomposes it into a discrete sequence of token IDs,

$$(\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_J), \quad \mathbf{t}_j \in \{1, \dots, |\mathcal{V}|\},$$

where J is the total number of tokens and $|\mathcal{V}|$ is the vocabulary size. To process these discrete symbols algebraically, the model maps them into a continuous vector space where geometric relationships capture semantic meaning. First, each token ID $\mathbf{t}_j$ is mapped to a one-hot vector $\mathbf{e}_{\mathbf{t}_j} \in \{0, 1\}^{|\mathcal{V}|}$. Subsequently, these vectors are projected into a dense, continuous space via the embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ learned during pre-training. The resulting continuous embedding of token $\mathbf{t}_j$ is therefore $\mathbf{E}^\top \mathbf{e}_{\mathbf{t}_j} \in \mathbb{R}^d$. Stacking all individual token embeddings column-wise yields the complete initial embedding matrix,

$$\mathbf{X}^{(0)} = [\mathbf{E}^\top \mathbf{e}_{\mathbf{t}_1}, \mathbf{E}^\top \mathbf{e}_{\mathbf{t}_2}, \dots, \mathbf{E}^\top \mathbf{e}_{\mathbf{t}_J}] \in \mathbb{R}^{d \times J}.$$

This dense matrix representation forms the continuous input that enters the decoder-only Transformer, replacing the original discrete text.

1.2. Transformer decoder layers A decoder-only LLM consists of L stacked Transformer blocks. Each block sequentially updates the token representations using two main components: a multi-head self-attention mechanism, and a position-wise feed-forward network. The self-attention mechanism is the core driver of the LLM, allowing the model to dynamically weigh the importance of all previous context when processing a given token. For this component, the embeddings $\mathbf{X}_{(\ell-1)}^{(0)}$ entering layer ℓ are linearly projected into query, key, and value matrices:

$$\mathbf{Q}_{(\ell)} = \mathbf{W}_{(\ell),Q} \mathbf{X}_{(\ell-1)}^{(0)}, \quad \mathbf{K}_{(\ell)} = \mathbf{W}_{(\ell),K} \mathbf{X}_{(\ell-1)}^{(0)}, \quad \mathbf{V}_{(\ell)} = \mathbf{W}_{(\ell),V} \mathbf{X}_{(\ell-1)}^{(0)},$$

where $\mathbf{W}_{(\ell),Q}, \mathbf{W}_{(\ell),K}, \mathbf{W}_{(\ell),V} \in \mathbb{R}^{d \times d}$ are learned weight matrices (and hence are part of the LLM parameter space), and d is the embedding dimension. Note that the subscript $(\ell - 1)$ denotes the $(\ell - 1)$ th Transformer layer, not the variable index. Furthermore, the superscript (0) in $\mathbf{X}_{(\ell-1)}^{(0)}$ indicates that this is the initial embedding matrix, formed prior to any autoregressive prediction.

^{*}[Corresponding author.]

Keywords and phrases: First keyword, second keyword.

Intuitively, these matrices act as a retrieval system: the query represents what information a token is looking for, the key represents what information it holds, and the value is the actual semantic content. The mechanism computes the relevance score between a query token i and a key token j using a scaled dot-product, which is further normalized via a row-wise softmax function to ensure these relevance scores form a valid probability distribution over the context:

$$\alpha_{ij,(\ell)} = \frac{\mathbf{q}_{i,(\ell)}^\top \mathbf{k}_{j,(\ell)}}{\sqrt{d}}, \quad A_{ij,(\ell)} = \frac{\exp(\alpha_{ij,(\ell)})}{\sum_{j'=1}^J \exp(\alpha_{ij',(\ell)})}.$$

Using these normalized attention weights, the model aggregates the values to form a context-aware representation for each token:

$$\mathbf{z}_{j,(\ell)} = \sum_{j'=1}^J A_{jj',(\ell)} \mathbf{v}_{j',(\ell)}.$$

The resulting aggregated matrix $\mathbf{Z}_{(\ell-1)}$ then passes through a position-wise feed-forward network, adding non-linearity, and serves as the input to the next layer:

$$\mathbf{X}_{(\ell)}^{(0)} = \text{FFN}(\text{SelfAttn}(\mathbf{X}_{(\ell-1)}^{(0)})) = \text{FFN}(\mathbf{Z}_{(\ell-1)}).$$

After passing through all L layers, the model produces a highly contextualized hidden state $\mathbf{h}^{(0)}$. Concretely, the process introduced above can be summarized as

$$\mathbf{h}^{(0)} = \text{Decoder}_\theta(\mathbf{X}^{(0)}).$$

At this point, the hidden state exists in the continuous latent space and needs to be translated back into the discrete vocabulary space to predict a word. This is achieved via the unembedding matrix $\mathbf{W}_{L+1}$ and bias $\mathbf{b}_{L+1}$, which map the hidden state into raw logits:

$$\mathbf{m}^{(1)} = \mathbf{W}_{L+1} \mathbf{h}^{(0)} + \mathbf{b}_{L+1} \in \mathbb{R}^{|\mathcal{V}|}.$$

These raw logits are then softmaxed into a final next-token probability distribution:

$$p_\theta(y_1 = \text{Detokenization}(\mathbf{t}_v) | \mathbf{X}^{(0)}) = \frac{\exp(\mathbf{m}_v^{(1)} / \tau)}{\sum_{j=1}^{|\mathcal{V}|} \exp(\mathbf{m}_j^{(1)} / \tau)}, \quad v = 1, \dots, |\mathcal{V}|,$$

where $\mathbf{t}_v$ is the v th token ID in the vocabulary, $\text{Detokenization}(\cdot)$ maps a token ID to a readable text token, τ is the temperature parameter controlling generation randomness, and $|\mathcal{V}|$ is the vocabulary size. Finally, a decoding strategy (e.g., greedy search or temperature sampling) converts this distribution into the actual prediction $\hat{y}_1$, whose embedding is appended to form $\mathbf{X}^{(1)}$ for the next step.

1.3. Autoregressive next-token prediction During text generation, the model predicts subsequent tokens autoregressively, meaning it uses its own past outputs as inputs for the next step. At decoding step t , let

$$\mathbf{X}^{(t-1)} = [\mathbf{X}^{(0)}, \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(t-1)}]$$

denote the concatenated embeddings of the initial prompt and the previously generated tokens. Passing this accumulated context $\mathbf{X}^{(t-1)}$ through the entire decoder stack yields the final contextual representation:

$$\mathbf{h}^{(t-1)} = \text{Decoder}_\theta(\mathbf{X}^{(t-1)}).$$

Following the discussion earlier, the information captured by $\mathbf{h}^{(t-1)}$ is then processed through unembedding and the softmax operation to form a probability distribution over the next token, from which $\hat{y}_t$ is generated.

1.4. *Relevance to variable-group selection* A review of this pipeline reveals a critical bottleneck for feature selection: LLM computation becomes differentiable only after the token embeddings $\mathbf{X}^{(0)}$ are formed. Because all prior discrete operations, including serialization, tokenization, and one-hot mapping, are inherently non-differentiable, any learnable, gradient-based variable-group selection must operate at the embedding level rather than the raw text level. On the other hand, the downstream transformer computation is completely differentiable with respect to $\mathbf{X}^{(0)}$. This architectural constraint directly motivates our masking formulation:

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

where $\mathbf{H}$ represents the embedding matrix corresponding to padding tokens, and s is a binary selection vector. By performing the selection dynamically in the continuous space, this construction preserves the differentiability of the entire pipeline, enabling standard gradient-based optimization of the selection variables.

2. Additional motivation examples We list one more additional motivation example for structured data serialization:

EXAMPLE 1 (Outcome-aware summary retrieval; Cui et al. 2025). Longitudinal medical records are typically structured such that each row represents an observational event (e.g., a visit, lab test, or procedure), and each column corresponds to the variables recorded during that event. Although separated into discrete rows, these observations naturally fall into clinically coherent groups: $z_1 = \{\text{Outpatient visits}\}$, $z_2 = \{\text{Laboratory events}\}$, $z_3 = \{\text{Imaging reports}\}$, $z_4 = \{\text{Procedures}\}$. $\square$

3. Numerical implementation

3.1. *Training procedure for gating baselines* We trained the proposed variable-selection framework by optimizing only the sigmoid/hard concrete/straight-through Bernoulli /ReinMax gating module while keeping the entire LLaMA-3.1-8B or Qwen3-8B backbone frozen. All experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU (49 GB VRAM) using bfloat16 precision. Because the serialized POINTER prompts vary in length, we adopted dynamic batching in which all sequences were right-padded using the tokenizer’s reserved special token (ID 128004 for LLaMA3.1-8B, ID 151654). For stochastic gating methods (ReinMax, straight-through Bernoulli, hard concrete) The same token was also used to mask any variable group that the gating elected to prune. Concretely, each group corresponds to a known span of token indices in the serialized prompt; when a group is pruned, its entire token span is replaced at the embedding level by the embedding vector associated with token 128004 in LLaMA-3.1-8B and Qwen3-8B. This strategy preserves the autoregressive structure of the LLM and avoids introducing artificial constraints into the causal attention mask.

During training, the gating module assigns each variable group a pair of logits that parameterize a temperature-scaled softmax distribution ($\tau = 1.5$). A hard keep-or-prune decision is sampled by drawing once from this categorical distribution, producing a discrete mask that overwrites the corresponding token embeddings. ReinMax then applies a second-order Heun-style correction in the backward pass, allowing a stable and unbiased gradient approximation despite the discrete sampling step. Only the gating logits are updated; the LLaMA backbone, positional encodings, and embedding table remain fixed throughout. Optimization uses AdamW with learning rate 10^{-4} , cosine annealing schedule, weight decay 10^{-5} , gradient clipping at 1.0, and batch size 2. At inference time, gating becomes deterministic: a group is retained if its learned inclusion probability exceeds 0.5, and pruned otherwise. This training design ensures numerical stability, prevents information leakage from pruned groups, and allows interpretable group-wise importance to be inferred directly from the discrete gating behavior.

4. Synthetic Data Generating Details To evaluate the stability of our feature-selection method under controlled conditions, we generate a semi-synthetic dataset that mirrors the statistical characteristics of the eight-dimensional diabetes feature space. The data are produced using a fully connected generative adversarial network (GAN) designed to approximate the joint distribution of normalized clinical covariates.

4.1. GAN Architecture The generator $G : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^p$ maps latent noise vectors to synthetic covariate vectors. Each latent variable is sampled from a standard multivariate Gaussian,

$$z \sim \mathcal{N}(0, I_{40}),$$

and passed through a sequence of six hidden layers with dimensions

$$40 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024 \rightarrow 1024 \rightarrow 8.$$

All hidden layers use ReLU activation functions, and the output layer uses a Tanh activation to match the $[-1, 1]$ -scaled real features, which are obtained via min-max normalization.

The discriminator $D : \mathbb{R}^p \rightarrow [0, 1]$ contains four hidden layers,

$$8 \rightarrow 512 \rightarrow 512 \rightarrow 256 \rightarrow 1,$$

each followed by ReLU activation and dropout with rate 0.4 for regularization. The final layer uses a sigmoid activation to return the probability that the input sample is real.

4.2. Training Objective The discriminator is trained using the standard GAN binary cross-entropy objective,

$$\mathcal{L}_D = \mathbb{E}_{x \sim \mathbb{P}_{\text{real}}} [\log D(x)] + \mathbb{E}_{z \sim \mathcal{N}(0, I)} [\log(1 - D(G(z)))],$$

while the generator is trained to maximize the discriminator’s likelihood of classifying generated samples as real,

$$\mathcal{L}_G = \mathbb{E}_{z \sim \mathcal{N}(0, I)} [\log D(G(z))].$$

To improve training stability, we apply label smoothing (real labels set to 0.9, fake labels to 0.1) and introduce random label flipping with probability 0.05, which helps prevent discriminator overconfidence.

4.3. Correlation Regularization To prevent the generator from learning spurious dependencies and to encourage disentangled representations, we add a correlation regularization term. For each batch of generated samples $\mathbf{X} = G(z) \in \mathbb{R}^{n \times p}$, we first standardize each feature to zero mean and unit variance and compute the Pearson correlation matrix $\mathbf{C} \in \mathbb{R}^{p \times p}$. The penalty is defined as the mean absolute off-diagonal correlation:

$$\mathcal{L}_{\text{corr}} = \frac{1}{p(p-1)} \sum_{i \neq j} |\mathbf{C}_{ij}|.$$

The total generator loss becomes

$$\mathcal{L}_G^{\text{total}} = \mathcal{L}_G + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}}, \quad \lambda_{\text{corr}} = 0.01,$$

balancing sample fidelity with reduced within-batch feature correlation.

4.4. Optimization Details Both networks are trained for 1000 epochs with a batch size of 32. We use the Adam optimizer with learning rate 2×10^{-4} for the generator and 5×10^{-5} for the discriminator, applying a weight decay of 10^{-5} to both. These settings provide stable convergence and yield synthetic samples that preserve the marginal distributions and broad dependence structure of the original diabetes covariates while supporting controlled evaluations of feature-selection performance.

4.5. *ReinMax approximation for discrete gating* A central computational challenge in optimizing the relaxed objective arises from the discrete gating variables. Although the Bernoulli parameterization replaces the original combinatorial optimization over $\mathbf{s} \in \{0, 1\}^K$ with continuous parameters β , the exact gradient of the expected loss still depends on pairwise comparisons across exponentially many gate configurations.

To illustrate, define

$$f(\mathbf{s}) = \mathcal{L}_{\text{LLM}}(\mathbf{y}, p_{\theta}(\cdot | \mathbf{X}(\mathbf{s}))),$$

where $\mathbf{X}(\mathbf{s})$ denotes the masked embedding matrix induced by gate vector $\mathbf{s}$. For a fixed coordinate k , let $\mathbf{s}_{k=1}$ and $\mathbf{s}_{k=0}$ denote two configurations that differ only in whether the k th variable group is retained or masked.

Then the exact gradient with respect to β_k can be written as

$$\frac{\partial}{\partial \beta_k} \mathbb{E}_{\mathbf{s}}[f(\mathbf{s})] = \sum_{\mathbf{s}' \in \mathcal{S}_{-k}} \mathbb{P}(\mathbf{s}'; \beta) \left(f(\mathbf{s}'_{k=1}) - f(\mathbf{s}'_{k=0}) \right),$$

where $\mathcal{S}_{-k}$ denotes all configurations of the remaining $K - 1$ gates.

This expression reveals the core computational bottleneck: evaluating the exact gradient requires computing the discrete difference

$$f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0})$$

across exponentially many neighboring configurations.

ReinMax (Liu et al., 2023) addresses this problem by replacing the discrete pairwise difference with a continuous second-order approximation that only requires first-order gradient information.

Specifically, define the interpolation path

$$g(x) = f(x\mathbf{s}_{k=1} + (1-x)\mathbf{s}_{k=0}), \quad x \in [0, 1].$$

Then

$$g(1) - g(0) = f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0}).$$

Rather than directly evaluating this discrete difference, ReinMax approximates it using Heun’s predictor–corrector rule:

$$g(1) - g(0) \approx \frac{1}{2} \left(g'(1) + g'(0) \right).$$

Applying the chain rule yields

$$g'(0) = \nabla f(\mathbf{s}_{k=0})^{\top} (\mathbf{s}_{k=1} - \mathbf{s}_{k=0}),$$

and similarly

$$g'(1) = \nabla f(\mathbf{s}_{k=1})^{\top} (\mathbf{s}_{k=1} - \mathbf{s}_{k=0}).$$

Substituting these expressions gives

$$f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0}) \approx \frac{1}{2} \left(\nabla f(\mathbf{s}_{k=1}) + \nabla f(\mathbf{s}_{k=0}) \right)^{\top} (\mathbf{s}_{k=1} - \mathbf{s}_{k=0}).$$

Importantly, this approximation only requires evaluating first-order gradients of the LLM loss at two nearby gate configurations, rather than explicitly computing all discrete differences over 2^{K-1} subsets. At the same time, because Heun’s rule is a second-order numerical integration method, the resulting estimator achieves second-order approximation accuracy while remaining compatible with standard backpropagation and stochastic gradient descent.

Consequently, ReinMax provides a computationally tractable surrogate for optimizing discrete variable-selection masks in high-dimensional LLM prompt space.

5. Technical Details of ReinMax This appendix provides a complete derivation of the ReinMax estimator (Liu et al., 2023) as used in our binary variable–selection setting. In the main manuscript, we illustrate the gating mechanism using independent Bernoulli variables

$$s_k \sim \text{Bernoulli}(\text{Logit}(\beta_k)),$$

where $\text{Logit}(\beta_k)$ denotes the inclusion probability. This Bernoulli formulation provides intuition, but the *implementation* follows the multinomial (two-class) formulation used in the original ReinMax paper (Liu et al., 2023), in which each gating decision is represented by a pair of logits $(\beta_{k,1}, \beta_{k,0})$ and sampled via a temperature-scaled softmax. The binary Bernoulli model is mathematically equivalent to the multinomial two-class model, so the derivations below remain valid and make explicit the pairwise-difference structure that motivates the ReinMax estimator.

5.1. Exact pairwise-difference gradient in the binary setting In this subsection, we will first derive why the gradient of the loss function involves the exact pairwise-difference. Recall our objective is

$$\mathcal{L}(\beta) = \mathbb{E}_{\mathbf{s} \sim \mathbb{P}(\beta)}[f(\mathbf{s})] + \lambda \sum_{k=1}^K \text{Logit}(\beta_k), \quad f(\mathbf{s}) = \mathcal{L}_{\text{LLM}}(\mathbf{y}, p_\theta(\cdot | \mathbf{X}(\mathbf{s}))).$$

Because $\mathbf{s}$ contains discrete coordinates, the gradient of $\mathbb{E}[f(\mathbf{s})]$ cannot be obtained via automatic differentiation. In the following discussions, we show that the exact gradient takes a pairwise-difference form:

$$(1) \quad \frac{\partial}{\partial \beta_k} \mathbb{E}_{\mathbf{s}}[f(\mathbf{s})] = \sum_{\mathbf{s}' \in \mathcal{S}_{-k}} \mathbb{P}(\mathbf{s}'; \beta) (f(\mathbf{s}'_{k=1}) - f(\mathbf{s}'_{k=0})),$$

where $\mathcal{S}_{-k}$ are the possible configurations of the other $K - 1$ indicators.

Step 1: Start from the definition.

$$\frac{\partial}{\partial \beta_k} \mathbb{E}_{\mathbf{s}}[f(\mathbf{s})] = \sum_{\mathbf{s}} f(\mathbf{s}) \frac{\partial \mathbb{P}(\mathbf{s}; \beta)}{\partial \beta_k}.$$

Step 2: Bernoulli factorization. Because s_k is independent of $\mathbf{s}_{-k}$, the joint distribution can be decomposed.

$$\mathbb{P}(\mathbf{s}; \beta) = \mathbb{P}(s_k; \beta_k) \mathbb{P}(\mathbf{s}_{-k}; \beta_{-k}),$$

where β_{-k} is the parameters related to $\mathbf{s}_{-k}$.

Then the differentiation of the β_k only involves the first probability, which is Bernoulli distribution with logit probability,

$$(2) \quad \frac{\partial \mathbb{P}(\mathbf{s}; \beta)}{\partial \beta_k} = \mathbb{P}(\mathbf{s}; \beta) (s_k - \text{Logit}(\beta_k)).$$

Step 3: Substitute into (1), (2) and partition by s_k .

$$\begin{aligned} \frac{\partial}{\partial \beta_k} \mathbb{E}[f(\mathbf{s})] &= \sum_{\mathbf{s}: s_k=1} f(\mathbf{s}) \mathbb{P}(\mathbf{s}) - \text{Logit}(\beta_k) \sum_{\mathbf{s}: s_k=1} \mathbb{P}(\mathbf{s}) \\ &\quad - \sum_{\mathbf{s}: s_k=0} f(\mathbf{s}) \mathbb{P}(\mathbf{s}) + \text{Logit}(\beta_k) \sum_{\mathbf{s}: s_k=0} \mathbb{P}(\mathbf{s}). \end{aligned}$$

Using the logit formulation of the probability:

$$\sum_{\mathbf{s}: s_k=1} \mathbb{P}(\mathbf{s}) = \text{Logit}(\beta_k), \quad \sum_{\mathbf{s}: s_k=0} \mathbb{P}(\mathbf{s}) = 1 - \text{Logit}(\beta_k),$$

we obtain the conditional-expectation form

$$(3) \quad \frac{\partial}{\partial \beta_k} \mathbb{E}[f(\mathbf{s})] = \text{Logit}(\beta_k) \mathbb{E}[f(\mathbf{s}) \mid s_k = 1] - (1 - \text{Logit}(\beta_k)) \mathbb{E}[f(\mathbf{s}) \mid s_k = 0].$$

Step 4: Expand conditional expectations. For any $\mathbf{s}' \in \mathcal{S}_{-k}$,

$$\mathbb{P}(\mathbf{s}' \mid s_k = 1) = \frac{\mathbb{P}(\mathbf{s}'_{k=1}; \beta)}{\text{Logit}(\beta_k)}, \quad \mathbb{P}(\mathbf{s}' \mid s_k = 0) = \frac{\mathbb{P}(\mathbf{s}'_{k=0}; \beta)}{1 - \text{Logit}(\beta_k)}.$$

Thus,

$$\begin{aligned} \mathbb{E}[f(\mathbf{s}) \mid s_k = 1] &= \sum_{\mathbf{s}'} f(\mathbf{s}'_{k=1}) \frac{\mathbb{P}(\mathbf{s}'_{k=1}; \beta)}{\text{Logit}(\beta_k)}, \\ \mathbb{E}[f(\mathbf{s}) \mid s_k = 0] &= \sum_{\mathbf{s}'} f(\mathbf{s}'_{k=0}) \frac{\mathbb{P}(\mathbf{s}'_{k=0}; \beta)}{1 - \text{Logit}(\beta_k)}. \end{aligned}$$

Step 5: Substitute into (3). We can get

$$\frac{\partial}{\partial \beta_k} \mathbb{E}[f(\mathbf{s})] = \sum_{\mathbf{s}'} f(\mathbf{s}'_{k=1}) \mathbb{P}(\mathbf{s}'_{k=1}) - \sum_{\mathbf{s}'} f(\mathbf{s}'_{k=0}) \mathbb{P}(\mathbf{s}'_{k=0}).$$

Step 6: Factor out marginals. Using

$$\mathbb{P}(\mathbf{s}'_{k=1}) = \text{Logit}(\beta_k) \mathbb{P}(\mathbf{s}'), \quad \mathbb{P}(\mathbf{s}'_{k=0}) = (1 - \text{Logit}(\beta_k)) \mathbb{P}(\mathbf{s}'),$$

we obtain the pairwise-difference gradient

$$\frac{\partial}{\partial \beta_k} \mathbb{E}_s[f(\mathbf{s})] = \sum_{\mathbf{s}' \in \mathcal{S}_{-k}} \mathbb{P}(\mathbf{s}'; \beta) (f(\mathbf{s}'_{k=1}) - f(\mathbf{s}'_{k=0})),$$

which completes the derivation of (1).

5.2. Binary ReinMax via second-order difference approximation As we can see above, the gradient of the loss function involves computing the exact pairwise difference of 2^{K-1} configurations, which is intractable. In this subsection, we use binary ReinMax as the case to show how it uses the first order information to approximate the pairwise difference at the second-order approximation accuracy.

We start by defining the interpolation path

$$g(x) = f(x\mathbf{s}_{k=1} + (1-x)\mathbf{s}_{k=0}), \quad x \in [0, 1].$$

Then, we can use $g(1)$ and $g(0)$ to get the pairwise difference:

$$g(1) - g(0) = f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0}).$$

Heun's rule gives the second-order surrogate of the exact pairwise difference using the first order information. Please see more in the original paper (Liu et al., 2023):

$$f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0}) \approx \frac{1}{2} (g'(1) + g'(0)),$$

with

$$g'(0) = \partial f(\mathbf{s}_{k=0})(\mathbf{s}_{k=1} - \mathbf{s}_{k=0}), \quad g'(1) = \partial f(\mathbf{s}_{k=1})(\mathbf{s}_{k=1} - \mathbf{s}_{k=0}).$$

Thus, the pairwise difference could be written as:

$$f(\mathbf{s}_{k=1}) - f(\mathbf{s}_{k=0}) \approx \frac{1}{2} (\partial f(\mathbf{s}_{k=1}) + \partial f(\mathbf{s}_{k=0})) (\mathbf{s}_{k=1} - \mathbf{s}_{k=0}).$$

Replacing (1), giving the binary ReinMax estimator

$$\hat{\nabla}_{\text{ReinMax},k} = \sum_{\mathbf{s}' \in \mathcal{S}_{-k}} \mathbb{P}(\mathbf{s}') \frac{1}{2} (\partial f(\mathbf{s}'_{k=1}) + \partial f(\mathbf{s}'_{k=0})) (\mathbf{s}'_{k=1} - \mathbf{s}'_{k=0}).$$

5.3. *Expected ℓ_0 penalty* We finally consider the ℓ_0 penalty gradient. Since

$$\mathbb{E}_{\mathbf{s}}[\|\mathbf{s}\|_0] = \sum_{k=1}^K \text{Logit}(\beta_k),$$

its gradient is deterministic and adds no variance.

5.4. *Multinomial ReinMax (implementation form)* For completeness, we describe the general multinomial form of ReinMax (Liu et al., 2023), of which our binary gating setup is a special case. Let each discrete decision be

$$D_k \in \{I_1, \dots, I_m\},$$

where I_i is the i -th standard basis vector in $\mathbb{R}^m$. Given logits $\beta_k \in \mathbb{R}^m$ and temperature τ , the sampling distribution is

$$\mathbb{P}(D_k = I_i) = \text{softmax}\left(\frac{\beta_k}{\tau}\right)_i, \quad i = 1, \dots, m.$$

Let $D = (D_1, \dots, D_K)$ denote the joint discrete mask. The exact gradient of the expected loss is

$$\nabla_{\beta} \mathbb{E}[f(D)] = \sum_{i,j=1}^m \mathbb{P}(D = I_j) (f(I_i) - f(I_j)) \frac{d\mathbb{P}(D = I_i)}{d\beta},$$

which matches Equation (7) of Liu et al. (2023). ReinMax replaces the discrete difference with the Heun midpoint surrogate:

$$f(I_i) - f(I_j) \approx \frac{1}{2} \left(\partial f(I_i) + \partial f(I_j) \right) (I_i - I_j),$$

yielding the second-order estimator

$$\mathbb{E}[\widehat{\nabla}_{\text{ReinMax}}] = \sum_{i,j=1}^m \mathbb{P}(D = I_j) \frac{1}{2} (\partial f(I_i) + \partial f(I_j)) (I_i - I_j) \frac{d\mathbb{P}(D = I_i)}{d\beta}.$$

REMARK 1. Binary gating as a special case. In our variable-selection setup, each group has only two choices (keep or prune), so $m = 2$ and

$$D_k \in \{I_1, I_0\},$$

reducing the multinomial formulation to the binary pairwise-difference expression derived earlier.

5.5. *ReinMax forward and backward propagation* We now describe the practical forward-backward mechanics of ReinMax in our binary variable-selection setting. Although the multinomial estimator of Section 5.4 applies to any number of discrete choices, our gates are binary ($m = 2$), corresponding to

$$D_k \in \{I_1, I_0\},$$

which represent “keep” and “prune” for variable group k . Each gate is parameterized by a pair of logits

$$\beta_k = (\theta_{k,1}, \theta_{k,0}) \in \mathbb{R}^2,$$

and sampling is performed through a temperature-controlled softmax with $\tau = 1.5$.

Forward pass. For each group k :

1. Compute soft probabilities

$$\mathbb{P}(D_k = I_i) = \text{softmax}\left(\frac{\beta_k}{\tau}\right)_i, \quad i \in \{1, 0\}.$$

2. Sample a discrete hard gate

$$y_k^{\text{hard}} \sim \text{Categorical}(\mathbb{P}(D_k)),$$

producing a one-hot vector in $\{I_1, I_0\}$.

3. Save both the hard sample y_k^{hard} and the soft probabilities $y_k^{\text{soft}} = \mathbb{P}(D_k)$.

Backward pass. To obtain a second-order accurate gradient while keeping the forward pass strictly discrete, ReinMax forms a midpoint surrogate

$$\tilde{y}_k = \frac{1}{2} \left(y_k^{\text{soft}} + y_k^{\text{hard}} \right),$$

mirroring the Heun midpoint term used in the surrogate $f(s_{k=1}) - f(s_{k=0})$. The gradient with respect to the logits then receives contributions from both

$$\nabla_{y_k^{\text{hard}}} \mathcal{L} \quad \text{and} \quad \nabla_{y_k^{\text{soft}}} \mathcal{L},$$

propagated through $\tilde{y}_k$ following standard automatic differentiation.

To stabilize optimization, especially when one logit has extremely small probability, we add a small constant ε to the “prune” logit before computing the softmax:

$$\theta_{k,0} \leftarrow \theta_{k,0} + \varepsilon.$$

This forward–backward mechanism preserves the hard discrete sampling used to construct the masked LLM input, while providing a smooth second-order gradient path that avoids the high variance of score-function estimators and the bias of straight-through relaxations. Additional examples of ReinMax in discrete optimization problems can be found in [Gao et al. \(2024\)](#); [Liu et al. \(2023\)](#).

6. Details of case study in Alzheimer’s Disease

6.1. *Fine-tuning dataset structure and example* One example “instruction-input-output” pair in fine-tuning dataset from ADNI is:

Instruction: You are a specialist in diagnosing and forecasting the disease progression for Alzheimer’s disease participants. Your task is to predict the following for a patient based on their medical records: the current amyloid beta (abeta) burden in the brain PET scan, the current tau burden of the temporal meta regions of interest (ROIs) in the brain PET scan, Mini-Mental State Examination (MMSE) score change rate, and whether this participant is the ideal candidate for anti-amyloid treatment. Please provide the prediction in this structured format: Abeta: Unknown or This patient abeta burden is [Negative/Weakly positive/Positive/Clearly positive]. Meta tau: Unknown or This patient Meta tau burden is [Very low/Low/High/Very high]. MMSE rate: Unknown or The MMSE is [Slow decline/Moderate decline/Fast decline/Stable].

Input: Patient Baseline information: **Demographic information:** This patient is born in 1938, ..., a self-identified race White, ethnicity Hispanic or Latino. **Genetic information:** This is patient is 0 APOE4 allele carrier, and is not APOE2 allele carrier. This visit is the 1 time visit. In total, the patient has 1 visits on 2005-08-18. The trajectory of features during these visits (from the very first to the latest) are: **Vitals:** Age: 67.6, ..., Temperature (in Celsius): 36.1, **Cognitive test:** Mini-Mental State Examination score: 27, ..., PACC score: -8.42, **AD biomarker in blood plasma:** Amyloid beta 42 in Cerebrospinal Fluid (pg/mL): 1501.0, ..., **Brain Volume in MRI:** Left Hippocampus: 3504, ..., Total intracranial: 1679440, **Brain Thickness in MRI:** Left Entorhinal (10E-3): 3080, ..., Right Meta ROI (10E-3): 2538, **Cognitive Diagnosis:** Diagnosis: Mildly cognitive impairment.

Output: Abeta: This patient abeta burden is Negative. Meta tau: Unknown. MMSE rate: The MMSE is Fast decline.

6.2. *Fine-tuning datasets* In $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, we have 1878 individuals with 9906 records from ADNI, 1872 individuals with 3129 records from HABS-HD, 752 individuals with 752 records from POINTER, and 354 individuals with 354 records from A4. The ADNI dataset does not contain Physical Measure, while HABS-HD does not have Family History, FDG-PET, and Medication, and A4 does not have FDG-PET and Comorbidities. Although all the variables used for fine-tuning can be categorized into thirteen groups, in fact, the specific variables in each variable group differ drastically from study to study. We list the raw variable names and the corresponding variable group in Table 1 to show which variable is included into which variable in four studies.

Table 1: Overview of input variables across cohorts (ADNI, HABS-HD, POINTER, A4).

Cohort	Variable Group	Variables
ADNI (83 vars)	Family History	MOTHDEM, MOTHAD, MOTHSXAGE, FATHDEM, FATHAD, FATHSXAGE, SIBDEMENT, SIBAD, SIBSXAGE
	Genetics	APOE4_NUM, APOE2_BIN
	Demographics	PTDOBY, PTHAND, PTNOTRT, PTMARRY, PTEDUCAT, PTRACCAT, PTETHCAT, AGE_AD_ON_SET, PTGENDER
	Vitals	Age, VSWEIGHT, VSHEIGHT, VSBPSYS, VSBPDIA, VSPULSE, VSRESP, VSTEMP
	Plasma Biomarkers	ABETA40, ABETA42, PTAU217, TAU
	Brain Volume in MRI	Left/Right Hippocampus, Left/Right Lateral Ventricle, ICV
	Brain Thickness in MRI	Thck_L/R_entorhinal, fusiform, inferiorparietal, inferiortemporal, middletemporal, precuneus, insula; Left_meta; Right_meta
	Medication	Transformed_KEYMED
	FDG-PET	FDG
	Cognitive Diagnosis	DIAGNOSIS
HABS-HD (75 vars)	Comorbidities	DXDEP, DXPARK
	Cognitive Tests	MMSCORE, GDTOTAL, COPYSCOR, ANARTERR, NPISCORE, CDRSB, ADAS11, ADAS13, ADASQ4, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting, LDEL-TOTAL, DIGITSCOR, TRABSCOR, FAQ, MOCA, mPACCdigit, EcogSP, EcogPt
	Genetics	APOE4, APOE2_carrier
	Demographics	ID_Retire, ID_MaritalStatus, Edu, Race, ID_Hispanic, IMH_AlzheimersAge, Smoke_Ever, ID_Gender
	Vitals	Age, BW_Glucose
	Brain Volume in MRI	Left/Right Hippocampus, Left/Right Lateral Ventricle, TIV_VOLUME
	Brain Thickness in MRI	THCK:L/R_entorhinal, fusiform, inferiorparietal, inferiortemporal, middletemporal, precuneus, insula; Left_meta; Right_meta
	Plasma Biomarkers	r5_QTX_Plasma_Abeta40, Abeta42, pTau181, Total_Tau

Table 1 (continued)

Cohort	Variable Group	Variables
	Physical Measure	HI_Score
	Cognitive Diagnosis	CDX_Cog
	Comorbidities	IMH_Cancer, IMH_Diabetes, IMH_HighBP, IMH_HighCholesterol, IMH_HeartAttack, IMH_ThyroidDisease, IMH_KidneyDisease, IMH_Stroke, IMH_Anxiety, IMH_Osteoporosis, IMH_Arthritis, IMH_TBI
	Cognitive Tests	MMSE_Total, GDS_Total, AMNART_Er, DS_Total, LM2_A/B_Total, CDR_Sum, Digit_Symbol_Substitution, AUDIT_Total, SMC_Total, PED_Total, SEVLT_T1_Total, SEVLT_T2_Total, SEVLT_T3_Total, SEVLT_T5_Total, SEVLT_T1235_Total, FAS_Total, WAT_Correct, Animal_Total, TOMM_Total, PACC_only_cog
POINTER (93 vars)	Family History	memoryloss_family, memoryloss_mother, memory- loss_father, memoryloss_sibling
	Genetics	APOE2_BIN, APOE4_NUM
	Demographics	bl_retired, smokerChar, bl_relationship, Edu_ysr, bl_occupation, PersonOfColor, race_simple, raceEth, gender
	Vitals	age, bpavgsys, bpavgdia, weightavg_kg, heightavg_cm, hba1c, glucose
	Brain Volume in MRI	Left/Right Lateral Ventricle, Left/Right Hippocampus, tiv
	Brain Thickness in MRI	Thck_L/R_entorhinal, fusiform, inferiorparietal, in- feriortemporal, middletemporal, precuneus, insula; Left_meta; Right_meta
	Medication	medName
	Physical Measure	PSQI_Global_Score, freq_cog, freq_social, phys_func_comp, phys_hlth_comp, emot_prob_comp, ener_fati_comp, emot_well_comp, soci_func_comp, pain_comp, gen_hlth_comp, hlth_chg_comp
	Cognitive Diagnosis	CDX_Cog
	Comobidities	diabetes_type2, diabetes_status, CVD_bl
	Cognitive Tests	mindtotal, ISI_score, TICS_adj, animalstotcor- rect, mmsetotscore, delayrecalltot, cdr_scr_sum, FCSRT_Free_Cued_96, episodic_score_z, execu- tive_score_z, processing_score_z, global_composite_Z, FINGER_NTB_z, PACC_5_z, gds_total, UCB_ADNIPOINTER_PACC, CFI_ppt, CFI_sp, DCTScore, ecog_ppt_total, ecog_ppt_memory, ecog_ppt_language, ecog_ppt_visuosp, ecog_ppt_planning, ecog_ppt_organize, ecog_ppt_attention, ecog_sp_total, ecog_sp_memory, ecog_sp_language, ecog_sp_visuosp, ecog_sp_planning, ecog_sp_organize, ecog_sp_attention
A4 (85 vars)	Family History	MOTHER, MOTHDIAG, MOTHAGE, FATHER, FATH- DIAG, FATHAGE, SIBDEMENT, SIBDIAG, SIBAGE
	Genetics	APOE4

Table 1 (continued)

Cohort	Variable Group	Variables
	Demographics	PTNOTRT, PTMARRY, PTEDUCAT, PTRACE, PTETHNIC
	Vitals	PTAGE, VSWEIGHT, VSHEIGHT, VSBPSYS, VSBPDIA, VSPULSE, VSRESP, VSTEMP
	Plasma Biomarkers	Plasma_Abeta40, Abeta42, pTau217, Total_Tau
	Brain Volume in MRI	Left/Right Lateral Ventricle, Left/Right Hippocampus, tiv
	Brain Thickness in MRI	Thck:L/R_entorhinal, fusiform, inferiorparietal, inferior temporal, middle temporal, precuneus, insula; Left_meta; Right_meta
	Habits	SMOKE, ALCOHOL, CAFFEINE, SUBUSE, AEROBIC, WALKING, SLEEP, SLEEPDAY
	Physical Measure	PXHEADEY, PXCARD, PXPULM, PXABDOM, PXMUSCUL, PXEDEMA, PXSKIN, PXOTHER, NXGAIT, NXMOTOR, NXSENSOR, NXTREMOR, NXFINGER, NXHEEL, NXNERVE, NXOTHER
	Cognitive Diagnosis	CDX_Cog
	Cognitive Tests	PACCRN, MMSCORE, DIGITTOTAL, FCFREET1-3, FCCUEDT1-3, LIMMTOTAL, LDELTOTAL, CDSOB

6.3. *POINTER cohort* The U.S. POINTER Study is a two-year, multi-site randomized lifestyle-intervention trial designed to evaluate whether a combination of physical activity, nutritional guidance, cognitive and social engagement, and cardiovascular risk management can preserve cognitive function in older adults at elevated risk of dementia. The cohort is currently enrolling and following approximately 2,100 participants aged 60–79 years across five clinical sites in the United States. At baseline, the mean age was 68.2 years; 68.9% were female; 30.8% identified as members of racially or ethnically minoritized groups; and 30% were APOE ϵ 4 carriers. Participants reported an average of 15.9 years of education, and 78% had a first-degree family history of memory loss. We have 752 individuals and records from POINTER as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$. The left 188 individuals and records from POINTER are $\mathcal{D}_{\text{test}}$.

6.4. *LLM-Rank customized prompt about feature importance*

“Given a list of features, rank them according to their importance in predicting the current amyloid beta (abeta) burden in the brain PET scan, the current tau burden of the temporal meta regions of interest (ROIs) in the brain PET scan, and the Mini-Mental State Examination (MMSE) score change rate of a patient. The ranking should be in descending order, starting with the most important feature. Your response should be a numbered list with each item on a new line. For example: 1. foo 2. bar 3. baz. Only output the ranking. Do not output dialogue or explanations for the ranking. Do not exclude any features in the ranking. Rank all 11 features in the following list: demographics (including education), genetics, family history, vital signs (including age), cognitive test scores, physical measure score, medications, brain volume in MRI, brain thickness in MRI, cognitive diagnosis, and comorbidities.”

6.5. *Accumulated attention scores* Let the serialized input contain N tokens partitioned into K clinical variable groups, with group k containing the token index set $\mathcal{G}_k \subset \{1, \dots, N\}$ and cardinality $N_k = |\mathcal{G}_k|$. Let $\mathbf{D} \in \mathbb{R}^{N \times d}$ denote the dense embedding matrix and $\mathbf{P} \in \mathbb{R}^{N \times d}$ the positional encoding. The combined representation $(\mathbf{D} + \mathbf{P})$ is processed by L Transformer layers, each with H attention heads.

For head h in layer l , token embeddings are projected into queries, keys, and values via

$$\mathbf{Q}_{l,h} = (\mathbf{D} + \mathbf{P})W_{l,h}^Q, \quad \mathbf{K}_{l,h} = (\mathbf{D} + \mathbf{P})W_{l,h}^K, \quad \mathbf{V}_{l,h} = (\mathbf{D} + \mathbf{P})W_{l,h}^V,$$

where $W_{l,h}^Q, W_{l,h}^K, W_{l,h}^V \in \mathbb{R}^{d \times d_k}$ and $d_k = d/H$ is the subspace dimension for each head.

To enforce autoregressive decoding, a causal mask $\mathbf{M} \in \mathbb{R}^{N \times N}$ is applied, where $\mathbf{M}_{ij} = 1$ if token j may be attended to by token i (i.e., $j \leq i$) and 0 otherwise. The scaled dot-product attention for head h in layer l is therefore

$$\mathbf{S}_{l,h} = \text{softmax}\left(\frac{\mathbf{Q}_{l,h}\mathbf{K}_{l,h}^\top \odot \mathbf{M}}{\sqrt{d_k}}\right), \quad \mathbf{S}_{l,h} \in \mathbb{R}^{N \times N}.$$

Group-aggregated attention. To relate token-level attention to higher-level clinical groups, we aggregate attention weights across the token indices in each group. Let v_0 denote the index of the first generated token in the autoregressive output. For subject i , we define the attention from group k to the generated token v_0 in head h and layer l as the block average:

$$\mathbf{S}_{l,h}^*(i)[k] = \frac{1}{N_k} \sum_{u \in \mathcal{G}_k} \mathbf{S}_{l,h}(u, v_0).$$

This summarizes how strongly tokens in clinical group k attend to the model’s first prediction step, which aligns with the autoregressive interpretation of attention influence.

Aggregating across layers and heads. The final attention-derived importance score for group k in subject i is the average over all layers and attention heads:

$$\mathbf{A}_{i,k} = \frac{1}{LH} \sum_{l=1}^L \sum_{h=1}^H \mathbf{S}_{l,h}^*(i)[k].$$

The vector $(\mathbf{A}_{i,1}, \dots, \mathbf{A}_{i,K})$ measures how much predictive attention the LLM allocates to each variable group for subject i .

6.6. Catboost training details To ensure comparability across outcomes, we trained three separate CatBoost models corresponding to the prediction of $A\beta$ burden, meta-tau burden, and MMSE change rate. For each model, we computed the average feature importance across all features within each variable category. We then combined these results by averaging the category-level importances across the three outcomes, yielding an overall ranking of the most influential variable groups. This aggregation provides a unified measure of predictive contribution for three outcomes.

6.7. Embedding-level stochastic gate baselines In response to the concern that LLM-Rank, attention-based selection, and CatBoost operate under substantially different mechanisms from our proposed method, we further compare against several embedding-level gating baselines. These baselines use the same serialized POINTER prompts, the same frozen ADLLM, the same group-wise padding replacement mechanism, and the same GBIC tuning procedure as the proposed method.

They differ only in how the binary group-selection gates are relaxed and optimized. This design isolates the contribution of the ReinMax gradient estimator from the broader embedding-level masking framework.

Deterministic sigmoid gate First, we consider a deterministic sigmoid gate. For each variable group k , a learnable logit β_k is transformed into a continuous retention weight $p_k = \sigma(\beta_k)$, and the masked embedding block is defined as

$$\mathbf{X}_k(p_k) = p_k \mathbf{X}_k + (1 - p_k) \mathbf{H}_k,$$

where $\mathbf{H}_k$ denotes the special-token replacement embedding block as our main paper. The sparsity penalty is $\lambda \sum_k p_k$, and the final selected set is obtained by thresholding $p_k > 0.5$. This baseline evaluates whether a non-stochastic continuous relaxation alone is sufficient for group selection.

Straight-through Bernoulli gate Second, we include a straight-through (Liu et al., 2023; Bengio, Léonard and Courville, 2013) Bernoulli gate. In this baseline, each group indicator is sampled as

$$s_k \sim \text{Bernoulli}(\sigma(\beta_k)),$$

and the hard sample is used in the forward pass. Gradients are propagated through the corresponding sigmoid probability using the straight-through estimator. This provides a simple stochastic binary-gate comparator that is close to the discrete feature-selection objective but uses a lower-order gradient approximation with more variance than ReinMax. More details of approximating accuracy can be referred to (Liu et al., 2023).

Hard-concrete L_0 gate Third, we compare with a hard-concrete L_0 gate, following the neural sparsification literature (Louizos, Welling and Kingma, 2017). The hard-concrete relaxation samples a stretched and clipped concrete random variable, allowing the relaxed gate to take exact values at zero or one while remaining differentiable almost everywhere. Its sparsity penalty is based on the expected number of nonzero gates, yielding an approximation to an L_0 -regularized objective. This is a natural baseline because it is specifically designed for sparse neural feature selection and model pruning.

For all embedding-level gate baselines, the ADLLM parameters remain frozen and only the group-gate logits are optimized. After optimization, we fix the selected subset by thresholding the learned retention probabilities and recompute the deterministic prediction loss on the POINTER cohort. The tuning parameter λ is selected by the same GBIC criterion used for the proposed ReinMax method:

$$\text{GBIC}(\lambda) = 2 \sum_{i=1}^n \mathcal{L}_{\text{LLM}}\{y_i, p_\theta(\cdot \mid \mathbf{X}_i(\hat{\mathbf{s}}_\lambda))\} + \log(n) |\widehat{\mathcal{M}}(\lambda)|.$$

Thus, comparisons among the proposed method and these gate baselines are made under the same prompt representation, masking operation, frozen LLM, sparsity grid, and model-selection criterion.

6.8. LLaMA-3.1-8B: Deterministic sigmoid gate results We ran the POINTER embedding gate selection pipeline using the deterministic sigmoid gate. The sweep tested $\lambda \in \{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 3 \times 10^{-2}\}$. All runs used seed=42, batch size=2, the same 25 epochs as the main experiments, and a selection threshold of 0.5 on learned keep probabilities. The minimum GBIC is when $\lambda = 3 \times 10^{-4}$, yielding 6 selected variable groups: Demographic information, Generic information, Family history, Vitals, Cognitive test, Physical measure score. All the variable group and their keep probabilities are listed in Table 2.

A practical point worth emphasizing is that the deterministic sigmoid gate operates continuously at the embedding level rather than performing discrete token- or sentence-level masking. Specifically, the learned keep probability interpolates between the original group embeddings and a fixed replacement embedding used in the experiment. Consequently, a high keep probability indicates that the original group embeddings are largely preserved, whereas a low keep probability indicates that the embeddings are moved closer to the replacement embedding.

Importantly, this mechanism does not correspond directly to human-interpretable deletion or insertion of words, sentences, or variable groups. Instead, the perturbation occurs in the

latent embedding space, and its downstream effect on the language model may be substantially subtler than completely removing textual content from the prompt. Therefore, the final selected groups should not be interpreted as exact semantic inclusion or exclusion decisions. This continuous embedding interpolation may also explain the relatively weaker selection performance observed for the deterministic sigmoid gate, since the model may still retain partial information from groups even when their keep probabilities are low.

In addition, because the gate is continuous, the selection threshold (0.5 in our experiments) is an operational choice rather than a strict theoretical boundary. Variable groups with intermediate keep probabilities may therefore be considered borderline selections and can be sensitive to the regularization parameter λ and random initialization.

TABLE 2
Group-wise selection results of deterministic sigmoid gate under minimum GBIC.

Group name	Keep probability	Selected
Demographic information	0.522	Yes
Genetic information	0.671	Yes
Family history	0.699	Yes
Vitals	0.886	Yes
Cognitive test	0.864	Yes
Physical measure score	0.746	Yes
Medication	0.450	No
Brain Volume in MRI	0.448	No
Brain Thickness in MRI	0.426	No
Cognitive Diagnosis	0.212	No
Comorbidities	0.338	No

6.9. *LLaMA-3.1-8B: Hard-concrete estimator results* We also evaluated the hard-concrete estimator on the POINTER cohort. The grid search considered $\lambda \in \{0.001, 0.005, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0\}$.

All runs used seed = 42, batch size = 2, 25 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.1$, yielding 7 selected variable groups: *demographic information*, *genetic information*, *family history*, *vitals*, *cognitive test*, *physical measure score*, and *brain volume in MRI*.

The hard-concrete estimator produced near-binary keep probabilities for most variable groups, with selected variables assigned probabilities close to 1 and several unselected variables assigned probabilities close to 0. This behavior is consistent with its design as a differentiable approximation to ℓ_0 -regularized hard selection.

6.10. *LLaMA-3.1-8B: Straight-through Bernoulli estimator results* We additionally evaluated the straight-through (ST) Bernoulli estimator for group-wise feature selection on the POINTER cohort. The grid search considered $\lambda \in \{10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 4 \times 10^{-3}, 5 \times 10^{-3}, 6 \times 10^{-3}, 7 \times 10^{-3}, 8 \times 10^{-3}, 9 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}\}$. All experiments used seed = 42, batch size = 2, 25 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 4 \times 10^{-3}$, yielding 8 selected variable groups: *demographics*, *genetics*, *family history*, *vitals*, *cognitive test*, *physical measure score*, *brain volume (MRI)*, and *brain thickness (MRI)*.

Compared with the deterministic sigmoid gate, the straight-through Bernoulli estimator produced substantially sharper selection probabilities, with several groups pushed close to

TABLE 3
Group-wise selection results of the hard-concrete estimator under minimum GBIC.

Group name	Keep probability	Selected
Demographic information	0.997	Yes
Genetic information	0.998	Yes
Family history	0.992	Yes
Vitals	0.999	Yes
Cognitive test	0.998	Yes
Physical measure score	0.999	Yes
Medication	0.013	No
Brain Volume in MRI	0.667	Yes
Brain Thickness in MRI	0.034	No
Cognitive diagnosis	0.012	No
Comorbidities	0.062	No

either 0 or 1. This behavior is expected because the ST estimator performs discrete forward-pass sampling while using a continuous surrogate gradient during backpropagation, thereby more closely approximating hard binary selection decisions. As a result, the selected groups are generally easier to interpret than those produced by the deterministic sigmoid gate, whose embedding-level interpolation may retain partial information even for groups with relatively small keep probabilities.

At the same time, several variable groups exhibited intermediate probabilities, suggesting that the inferred importance structure remains sensitive to the regularization parameter λ and the stochastic optimization trajectory. Overall, the ST-Bernoulli estimator selected a broader subset of clinically meaningful AD-related domains, particularly MRI-derived neurodegeneration markers, while still inducing substantial sparsity in the serialized prompt.

TABLE 4
Group-wise selection results of the straight-through Bernoulli estimator under minimum GBIC.

Group name	Keep probability	Selected
Demographic information	0.881	Yes
Genetic information	0.906	Yes
Family history	0.709	Yes
Vitals	0.896	Yes
Cognitive test	0.810	Yes
Physical measure score	0.641	Yes
Medication	0.338	No
Brain Volume in MRI	0.890	Yes
Brain Thickness in MRI	0.814	No
Cognitive diagnosis	0.029	No
Comorbidities	0.265	No

6.11. *Qwen3-8B: Deterministic sigmoid gate results* We additionally evaluated the deterministic sigmoid gate under the Qwen3-8B backbone. The grid search considered $\lambda \in \{10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. All experiments used seed = 42, batch size = 2, 25 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 10^{-4}$, yielding 7 selected variable groups: *genetic information*, *vitals*, *cognitive test*, *physical measure score*, *medication*, *brain volume (MRI)*, and *brain thickness (MRI)*.

Compared with the LLaMA-3.1-8B sigmoid-gate results, the Qwen3-8B backbone selected a somewhat broader set of clinically relevant domains, particularly MRI-related and medication-related variables. Similar to the main experiments, several groups exhibited intermediate keep probabilities, reflecting the continuous embedding-level nature of the deterministic sigmoid gate.

TABLE 5

Group-wise selection results of the deterministic sigmoid gate under Qwen3-8B at minimum GBIC.

Group name	Keep probability	Selected
Demographic information	0.969	Yes
Genetic information	0.309	No
Family history	0.777	Yes
Vitals	0.882	Yes
Cognitive test	0.878	Yes
Physical measure score	0.518	Yes
Medication	0.998	Yes
Brain Volume in MRI	0.663	Yes
Brain Thickness in MRI	0.349	No
Cognitive diagnosis	0.264	No
Comorbidities	0.953	Yes

6.12. *Qwen3-8B: Straight-through Bernoulli estimator results* We additionally evaluated the straight-through (ST) Bernoulli estimator under the Qwen3-8B backbone. The grid search considered $\lambda \in \{10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}, 6 \times 10^{-3}, 8 \times 10^{-3}, 10^{-2}\}$. All experiments used seed = 42, batch size = 2, 25 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 3 \times 10^{-3}$, yielding 7 selected variable groups: *demographic information*, *genetic information*, *vitals*, *cognitive test*, *brain volume (MRI)*, *brain thickness (MRI)*, and *cormorbidities*.

Compared with the deterministic sigmoid gate, the ST-Bernoulli estimator again produced sharper keep probabilities and more discrete selection behavior. The selected groups emphasize both MRI-derived neurodegeneration measures and broader clinical health-condition information under the Qwen3-8B backbone.

TABLE 6

Group-wise selection results of the ST-Bernoulli estimator under Qwen3-8B at minimum GBIC.

Group name	Keep probability	Selected
Demographic information	0.734	Yes
Genetic information	0.810	Yes
Family history	0.466	No
Vitals	0.768	Yes
Cognitive test	0.942	Yes
Physical measure score	0.498	No
Medication	0.072	No
Brain Volume in MRI	0.837	Yes
Brain Thickness in MRI	0.917	Yes
Cognitive diagnosis	0.036	No
Comorbidities	0.721	Yes

6.13. *Qwen3-8B: Hard-concrete estimator results* We further evaluated the hard-concrete estimator under the Qwen3-8B backbone. The grid search considered $\lambda \in \{0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.2, 0.25\}$. All experiments used seed = 42, batch size = 2, 25 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.14$, yielding 7 selected variable groups: *demographic information*, *genetic information*, *vitals*, *cognitive test*, *physical measure score*, *brain volume (MRI)*, and *brain thickness (MRI)*.

Similar to the LLaMA-3.1-8B experiments, the hard-concrete estimator produced sparse near-binary selections under the Qwen3-8B backbone. The selected variables consistently emphasized MRI-derived neurodegeneration markers and cognitive-related domains, while excluding several broader clinical-history groups.

TABLE 7
Selected variable groups of the hard-concrete estimator under Qwen3-8B at minimum GBIC.

Group name	Selected
Demographic information	Yes
Genetic information	Yes
Family history	No
Vitals	Yes
Cognitive test	Yes
Physical measure score	Yes
Medication	No
Brain Volume in MRI	Yes
Brain Thickness in MRI	Yes
Cognitive diagnosis	No
Comorbidities	No

6.14. *Real data marginal masking results: LLaMA-3.1-8B* To investigate whether fewer variable groups may improve prediction performance, we compare ADLLM performance on POINTER using all variable groups against performance obtained after marginally masking one variable group at a time. The results are shown in Figure 1 and Table 8. The dashed horizontal line denotes the performance of ADLLM using all variable groups, achieving $A\beta$ AUC = 0.659, accuracy = 0.729, and specificity = 0.269, and meta-tau AUC = 0.528, accuracy = 0.300, and specificity = 0.731. The x-axis indicates the variable group removed during the marginal masking procedure.

Figure 1 should be interpreted as an exploratory local perturbation analysis rather than evidence that most variables are globally unimportant. Since each variable group is removed individually while all remaining groups are retained, marginal masking does not measure the joint contribution of variables. In settings with partially redundant, noisy, or highly correlated information, removing one variable group may improve held-out performance even when the full set of variables remains jointly informative.

The results suggest that the deployed LLM predictor does not rely equally on all serialized variable groups, motivating a framework to conduct feature selection over LLM globally instead of marginally.

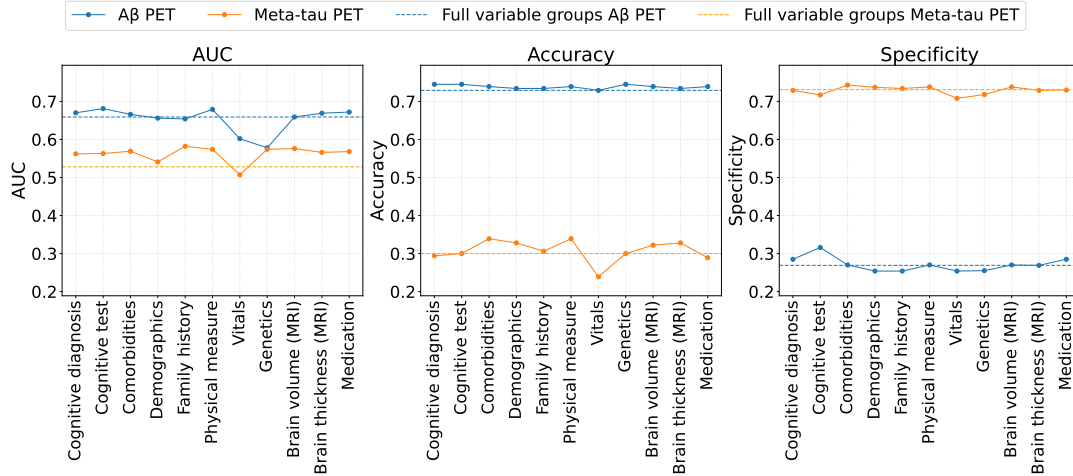


Fig 1: Marginal masking results for ADLLM evaluated on the POINTER cohort. Each point represents predictive performance after marginally masking one variable group from the serialized input while retaining all remaining groups. The dashed horizontal line denotes performance using all available variable groups ($A\beta$: AUC = 0.659, accuracy = 0.729, specificity = 0.269; meta-tau: AUC = 0.528, accuracy = 0.300, specificity = 0.731). The figure should be interpreted as an exploratory local perturbation analysis illustrating that different variable groups contribute unequally to the deployed LLM predictor.

We list the numeric number of marginal masking analysis on the POINTER cohort in Table 8.

TABLE 8

Marginal masking analysis on the POINTER cohort: each variable group is sequentially masked to evaluate its contribution to predictive performance. Excluding (or masking) certain features from LLM inference enhances the model’s predictive performance.

Variable Group	$A\beta$ Prediction			Meta-tau Prediction		
	AUC	Accuracy	Specificity	AUC	Accuracy	Specificity
Masking Cognitive diagnosis	0.670	0.745	0.285	0.562	0.294	0.729
Masking Cognitive test	0.681	0.745	0.316	0.563	0.300	0.717
Masking Comorbidities	0.666	0.739	0.270	0.569	0.339	0.743
Masking Demographics	0.656	0.734	0.254	0.541	0.328	0.737
Masking Family history	0.654	0.734	0.254	0.582	0.306	0.734
Masking Physical measure	0.679	0.739	0.270	0.574	0.339	0.738
Masking Vitals	0.602	0.729	0.254	0.507	0.239	0.708
Masking Genetics	0.578	0.745	0.255	0.574	0.300	0.718
Masking Brain volume (MRI)	0.659	0.739	0.270	0.576	0.322	0.738
Masking Brain thickness (MRI)	0.669	0.734	0.269	0.566	0.328	0.729
Masking Medication	0.672	0.739	0.285	0.568	0.289	0.730
All variable group	0.659	0.729	0.269	0.555	0.300	0.731

6.15. *Read data marginal masking results: Qwen3-8B* To investigate whether the marginal analysis results are not restricted to only LLaMA-3.1-8B, we add marginal masking results on Qwen3-8B in Table 9. We can see the similar trends. There are variable groups that we remove can either drop or increase the performance compared to using all variable groups baseline.

These results show a similar trend as LLaMA 3.1-8B, motivating the need of principle global feature selection framework.

TABLE 9

Marginal masking analysis on the POINTER cohort in Qwen3-8B: each variable group is sequentially masked to evaluate its contribution to predictive performance. Excluding (or masking) certain features from LLM inference enhances the model’s predictive performance.

Variable Group	A β Prediction			Meta-tau Prediction		
	AUC	Accuracy	Specificity	AUC	Accuracy	Specificity
Masking Cognitive diagnosis	0.660	0.713	0.298	0.524	0.267	0.742
Masking Cognitive test	0.667	0.702	0.447	0.551	0.267	0.746
Masking Comorbidities	0.641	0.707	0.312	0.547	0.244	0.736
Masking Demographics	0.664	0.718	0.328	0.536	0.233	0.730
Masking Family history	0.655	0.723	0.269	0.522	0.250	0.730
Masking Physical measure	0.663	0.723	0.389	0.539	0.256	0.738
Masking Vitals	0.624	0.676	0.400	0.482	0.256	0.752
Masking Genetics	0.582	0.745	0.255	0.546	0.267	0.731
Masking Brain volume (MRI)	0.656	0.713	0.402	0.516	0.239	0.735
Masking Brain thickness (MRI)	0.631	0.707	0.297	0.551	0.250	0.734
Masking Medication	0.652	0.718	0.283	0.536	0.261	0.737
All variable group	0.646	0.713	0.268	0.533	0.272	0.741

7. Details of case study in HIV electronic medical records

7.1. Background We utilized data from two regions in Tanzania from the national electronic medical records (EMRs) repository, which includes digitized medical records manually entered after clinical visits and maintained by the Ministry of Health (MoH) in Tanzania and the National AIDS, STIs, and Hepatitis Control Programme (NASHCoP), which are referred to as CTC3 medical records. Our analysis adopted the EMRs from January 1, 2018, to June 30, 2023, resulting in 4,809,765 records for 261,192 PLHIV. To alleviate the computation burden, we use subgroup with council in Kyerwa DC, marital status not missing, visit type not missing, reference source not missing, and $26 < \text{age} \leq 42$, viral load not missing, nutritional status not missing, and client category not missing from Kagera region, resulting in 20,970 records from 1,727 PLHV to get the enhanced large language model (LLM) to predict individuals at risk of disengaging from HIV care in Tanzania following procedure from [Wei et al. \(2026\)](#). We also follow their variable selection and obtain 24 variables, falling into 9 variable groups: clinical visit, demographics, physical condition, clinical stage, therapy, medical history, immunological and virological stats, health facility information, and static clinical information.

Table 10 summarizes the electronic health record variables used to construct longitudinal patient histories for people living with HIV. Repeated clinic visits were grouped by patient identifier and ordered chronologically by visit date. Time-varying clinical, demographic, treatment, laboratory, and facility-level variables were serialized as feature-wise sequences, so that each bracketed list represents the patient’s observed history up to a given visit. Static variables, such as the HIV testing or referral source, were included once per patient history. The prediction targets were viral load status, loss to follow-up for more than 28 days, and ART non-adherence risk based on days not on ART during 2022–2023.

Table 10: Electronic health record variables used for longitudinal serialization of HIV care histories.

Time-varying	Variable	Definition and serialization example	Values or coding	Variable group
No	ID	Patient identifier used to group repeated clinic visits into a longitudinal record.	Unique patient identifier.	Identifier
Yes	VisitDate	Date of the patient's clinic visit. Serialized as historical visit dates, e.g., <i>Actual visit dates: [2020-03-02, 2020-04-03, 2020-05-05, 2020-06-04]</i> .	Calendar date.	Clinical visit
Yes	visitscheduled	Scheduled appointment history. Because no separate scheduled appointment date was available, scheduled appointments were approximated using VisitDate when VisitTypeCode = S; otherwise the value was recorded as missing.	Derived from VisitTypeCode = S and VisitDate; otherwise NA.	Clinical visit
Yes	Sex, age	Demographic profile serialized jointly, e.g., <i>The patient is a 37 year-old female.</i>	Sex and age at visit or recruitment.	Demographics
Yes	MaritalStatus	Marital status history, serialized as a visit-level sequence, e.g., <i>Marital status: [divorced, divorced, divorced, divorced]</i> .	C/HH = child or not applicable; M = married; CO = cohabitating; D/WT = divorced or separated; S = single; W = widowed; NW = polygamy.	Demographics
Yes	NowPregnant	Pregnancy status history, e.g., <i>Pregnancy status: [not pregnant, not pregnant, not pregnant, not pregnant]</i> .	0 = No; 1 = Yes.	Demographics

Continued on next page

Time-varying	Variable	Definition and serialization example	Values or coding	Variable group
Yes	VisitTypeCode	Type of clinic visit, serialized as a visit-level sequence, e.g., <i>Visit type: [Unscheduled visit at this clinic, Scheduled Visit at this clinic, Scheduled Visit at this clinic]</i> .	S = scheduled visit; US = unscheduled visit; TK = traced back after LTFU; TS = treatment supporter drug pick-up; O = visit other clinic; IP = in-patient consultation; RVo = facility-led outreach community refill; RVp = fast-track refill at health facility pharmacy; NA = missing.	Clinical visit
Yes	clientcategory_lr	Client category history, e.g., <i>Client category: [Unstable Client, NA, Stable Client, Unstable Client]</i> .	S = stable client; UN = unstable client; NA = missing.	Clinical visit
Yes	Weight	Weight measurements across visits, e.g., <i>Weight: [68, 66, 66, 65]</i> .	Continuous numeric weight.	Physical condition
Yes	NutritionalStatusID	Nutrition status history, e.g., <i>Nutrition status: [Not malnourished, Not malnourished, Not malnourished]</i> .	OK = normal or well nourished; MOD = moderate malnutrition; SEV = severe malnutrition; OB = obese; NULL = not recorded.	Physical condition
Yes	WHOStage	WHO clinical stage history, e.g., <i>WHO stage: [WHO stage 1, WHO stage 1, WHO stage 1]</i> .	1 = WHO Stage 1; 2 = WHO Stage 2; 3 = WHO Stage 3; 4 = WHO Stage 4.	Clinical stage
Yes	tb_tpt	Tuberculosis preventive therapy status, serialized across visits.	0 = not on TPT; 1 = start TPT; 2 = continue TPT; 3 = complete TPT; 4 = stop TPT; 5 = restart TPT.	Therapy

Continued on next page

Time-varying	Variable	Definition and serialization example	Values or coding	Variable group
Yes	ARVStatusCode_lr	Antiretroviral therapy status history, e.g., <i>ART status: [continue ART, continue ART, continue ART]</i> .	1 = not started ARV; 2 = start ARV; 3 = continue ARV; 4 = change/substitute/switch ARV regimen; 5 = stop; 6 = restart; NA = missing.	Therapy
Yes	NumDaysDispensed	Number of days of ARV medication dispensed at each visit.	Continuous or integer count of days dispensed.	Medical history
Yes	numDaysNotOnART	Number of days not on ART.	Continuous or integer count of days.	Medical history
Yes	Arttime	Number of days the patient has been on ART, e.g., <i>The patient has been on ART for [0, 32, 64, 94] days.</i>	Continuous or integer count of days since ART initiation.	Medical history
Yes	Art_mins_hivdiag	Time interval component used with <i>Arttime</i> to calculate days since HIV diagnosis.	Continuous or integer count of days.	Medical history
Yes	Art_mins_hivdiag + Arttime	Number of days since HIV diagnosis, serialized across visits, e.g., <i>The number of days since HIV diagnosis: [2131, 2163, 2195, 2225] days.</i>	Calculated as <i>Art_mins_hivdiag</i> + <i>Arttime</i> .	Medical history
Yes	VL	Viral load history, e.g., <i>Viral load: [suppressed, unknown, unknown, unknown]</i> .	Suppressed = viral load < 1000 copies/ml; unsuppressed/detectable = viral load $\geq$ 1000 copies/ml; no VL = unknown.	Immunological and virological status
Yes	CD4_lr	CD4 measurement history.	1 = no CD4 measured; 2 = CD4 350+; 3 = CD4 200–350; 4 = CD4 below 200.	Immunological and virological status

Continued on next page

Time-varying	Variable	Definition and serialization example	Values or coding	Variable group
Yes	hf_type_lr	Health facility type attended by the patient, e.g., <i>Health facility type: [Health Center]</i> .	1 = clinic; 2 = dispensary; 3 = health center; 4 = hospital; 5 = maternity home.	Health facility information
No	ref_final	HIV testing or referral source. This static feature records how the patient entered HIV testing or care.	1 = provider-initiated testing and counseling (PITC); 2 = voluntary counseling and testing (VCT); 3 = index testing or partner/family tracing; 4 = community-based testing or referral; 5 = other.	Static clinical information
No	ltfu_risk	Risk profile based on days not on ART in 2022–2023.	High, moderate, low, or no risk.	Outcome
No	ever_ltfu	Indicator of whether the patient was lost to follow-up for more than 28 days.	0 = unlikely; 1 = possibly.	Outcome
No	single_VL	Viral load outcome.	Detectable, suppressed, or unknown.	Outcome

7.2. Fine-tuning data set format

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

Input: This is the patient’s 2nd visit. Demographics: the patient is a 35 year-old female; Marital status: [married, married]; Pregnancy status: [not pregnant, not pregnant]. Clinical visit: Visit dates: [2018-08-22, 2018-10-25]; Scheduled appointments: [2018-08-22, 2018-10-25]; Visit type: [Scheduled Visit at this clinic, Scheduled Visit at this clinic]; Client category: [Stable Client, Stable Client]. Medical history: Number of ARV pills prescribed on each visit date: [60, 30]; Number of days not on ART: [33, 4]; The patient has been on ART for [2219, 2283] days; The number of days since HIV diagnosis: [4494, 4622] days. Therapy: TB TPT status: [Continue taking TPT, Continue taking TPT]; ART status: [CONTINUE ARV, CONTINUE ARV]. Physical condition: Weight: [58, 55]; Nutrition status: [Not malnourished, Not malnourished]. Clinical stage: WHO stage: [WHO stage is stage 4 (severe symptoms), WHO stage is stage 4 (severe symptoms)]. Immunological and virological status: CD4 status: [No CD4 measured, No CD4 measured]; Viral load: [suppressed, unknown]. Health facility information: Health facility type: [Dispensary]. Static clinical information: HIV testing or referral source: [VCT].

Output: This PLHIV will have [suppressed] viral load, is [unlikely] to be lost to follow-up for over 28 days, and is at [no] risk of ART non-adherence based on their days not on ART in the year 2022-2023.

7.3. LLM-Rank customized prompt about feature importance

“Given a list of features, rank them according to their importances in predicting whether a person living with HIV (PLHIV) will have suppressed, detectable, or unknown viral load; whether the PLHIV will be lost to follow-up for more than 28 days; and the PLHIV’s risk of ART non-adherence based on days not on ART in the year 2022-2023. The ranking should be in descending order, starting with the most important feature. Your response should be a numbered list with each item on a new line. For example: 1. foo 2. bar 3. baz. Only output the ranking. Do not output dialogue or explanations for the ranking. Do not exclude any features in the ranking. Rank all features in the following list: demographics, marital status, pregnancy status, visit dates, scheduled appointments, visit type, client category, number of ARV pills prescribed on each visit date, number of days not on ART, duration on ART, number of days since HIV diagnosis, TB TPT status, ART status, weight, nutrition status, WHO stage, CD4 status, viral load status, health facility type, and HIV testing or referral source.”

This returns the ranked list: 1. number of days not on ART, 2. duration on ART, 3. number of ARV pills prescribed on each visit date, 4. number of days since HIV diagnosis, 5. viral load status, 6. CD4 status, 7. WHO stage, 8. TB TPT status, 9. ART status, 10. weight, 11. nutrition status, 12. client category, 13. scheduled appointments, 14. visit dates, 15. visit type, 16. HIV testing or referral source, 17. health facility type, 18. demographics, 19. marital status, 20. pregnancy status.

7.4. LLaMA-3.1-8B: Deterministic sigmoid gate results We additionally evaluated the deterministic sigmoid gate on the HIV cohort. The grid search considered $\lambda \in \{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 1.2 \times 10^{-3}, 1.4 \times 10^{-3}, 1.6 \times 10^{-3}, 1.8 \times 10^{-3}, 10^{-2}\}$. All experiments used seed = 42, batch size = 1 (because of the long prompt and the GPU capacity, 8 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 10^{-3}$, yielding 15 selected variable groups: *demographics, marital status, pregnancy status, visit dates, scheduled appointments, visit type, client category, number of ARV pills prescribed, number of days not on ART, duration on ART, number of days since HIV diagnosis, ART status, weight, WHO stage, and health facility type*.

The selected variables emphasize longitudinal treatment adherence, healthcare engagement, and clinical follow-up patterns rather than static demographic characteristics alone. In particular, ART-related adherence trajectories, visit scheduling information, and duration since HIV diagnosis were consistently prioritized, aligning with the clinical objective of identifying future disengagement from HIV care and treatment instability.

7.5. LLaMA-3.1-8B: hard concrete results We further evaluated the hard-concrete estimator on the HIV cohort. The grid search considered $\lambda \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 0.85, 0.9, 0.95, 1.0, 1.2, 1.4\}$. All experiments used seed = 42, batch size = 1, 8 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.9$, yielding 13 selected variable groups: *demographics, marital status, pregnancy status, visit dates, scheduled appointments, visit type, number of ARV pills prescribed, number of days not on ART, duration on ART, number of days since HIV diagnosis, TB TPT status, WHO stage, and HIV testing or referral source*.

Compared with the deterministic sigmoid gate, the hard-concrete estimator again produced sharper near-binary selection behavior. The selected variables consistently emphasized longitudinal HIV-care engagement, ART adherence history, and treatment monitoring variables, while additionally prioritizing tuberculosis preventive treatment (TB TPT) status and HIV referral-source information. These results further suggest that the fine-tuned HIV-LLM relies strongly on longitudinal treatment trajectories and healthcare-engagement patterns when predicting future HIV treatment outcomes.

TABLE 11

Selected variable groups of the deterministic sigmoid gate on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Demographics	0.854	Yes
Marital status	0.736	Yes
Pregnancy status	0.727	Yes
Visit dates	0.927	Yes
Scheduled appointments	0.536	Yes
Visit type	0.840	Yes
Client category	0.989	Yes
Number of ARV pills prescribed on each visit date	0.525	Yes
Number of days not on ART	0.709	Yes
The patient has been on ART for	0.978	Yes
The number of days since HIV diagnosis	0.788	Yes
ART status	0.588	Yes
Weight	0.864	Yes
WHO stage	0.650	Yes
Health facility type	0.991	Yes

TABLE 12

Selected variable groups of the hard-concrete estimator on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Demographics	0.994	Yes
Marital status	0.997	Yes
Pregnancy status	0.973	Yes
Visit dates	0.993	Yes
Scheduled appointments	0.996	Yes
Visit type	0.980	Yes
Number of ARV pills prescribed on each visit date	0.998	Yes
Number of days not on ART	0.842	Yes
The patient has been on ART for	0.911	Yes
The number of days since HIV diagnosis	0.984	Yes
TB TPT status	0.991	Yes
WHO stage	0.962	Yes
HIV testing or referral source	0.944	Yes

7.6. LLaMA-3.1-8B: Straight-through Bernoulli estimator results We additionally evaluated the straight-through (ST) Bernoulli estimator on the HIV cohort. The grid search considered $\lambda \in \{0.001, 0.002, 0.003, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$. All experiments used seed = 42, batch size = 1, 8 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.002$, yielding 13 selected variable groups: *demographics, pregnancy status, client category, number of days not on ART, number of days since HIV diagnosis, TB TPT status, ART status, weight, nutrition status, CD4 status, viral load, health facility type, and HIV testing or referral source.*

Compared with the deterministic sigmoid and hard-concrete estimators, the ST-Bernoulli estimator selected a broader set of laboratory and treatment-monitoring variables, including CD4 status, viral load, and nutritional status. The selected variables collectively emphasize longitudinal disease progression, ART adherence trajectories, and healthcare engagement patterns, suggesting that the fine-tuned HIV-LLM relies heavily on dynamic treatment and monitoring information when predicting future HIV outcomes.

TABLE 13
Selected variable groups of the ST-Bernoulli estimator on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Demographics	0.525	Yes
Pregnancy status	0.859	Yes
Client category	0.740	Yes
Number of days not on ART	0.879	Yes
The number of days since HIV diagnosis	0.956	Yes
TB TPT status	0.887	Yes
ART status	0.956	Yes
Weight	0.973	Yes
Nutrition status	0.967	Yes
CD4 status	0.924	Yes
Viral load	0.972	Yes
Health facility type	0.967	Yes
HIV testing or referral source	0.983	Yes

7.7. *Qwen3-8B: Deterministic sigmoid gate results* We additionally evaluated the deterministic sigmoid gate on the HIV cohort. The grid search considered $\lambda \in \{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$. All experiments used seed = 42, batch size = 1, 8 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 5 \times 10^{-4}$, yielding 12 selected variable groups: *marital status, pregnancy status, visit dates, scheduled appointments, visit type, number of ARV pills prescribed, number of days not on ART, duration on ART, TB TPT status, weight, viral load, and HIV testing or referral source*.

The selected variables again emphasize longitudinal treatment adherence and healthcare-engagement patterns, particularly ART refill behavior, duration on treatment, visit scheduling, and viral-load monitoring. Compared with the AD experiments, the HIV cohort places substantially greater emphasis on longitudinal healthcare utilization and treatment-monitoring variables, reflecting the chronic and engagement-dependent nature of HIV care prediction.

TABLE 14
Selected variable groups of the deterministic sigmoid gate on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Marital status	0.561	Yes
Pregnancy status	0.962	Yes
Visit dates	0.561	Yes
Scheduled appointments	0.920	Yes
Visit type	0.659	Yes
Number of ARV pills prescribed on each visit date	0.956	Yes
Number of days not on ART	0.914	Yes
The patient has been on ART for	0.771	Yes
TB TPT status	0.897	Yes
Weight	0.649	Yes
Viral load	0.547	Yes
HIV testing or referral source	0.697	Yes

7.8. *Qwen3-8B: hard concrete results* We further evaluated the hard-concrete estimator on the HIV cohort. The grid search considered $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 0.9, 0.32, 0.34, 0.36, 0.38, 0.42, 0.44, 0.46, 0.48, 1.0\}$. All experiments used seed = 42, batch size = 1, 8 training

epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.34$, yielding 13 selected variable groups: *demographics*, *marital status*, *pregnancy status*, *visit dates*, *scheduled appointments*, *visit type*, *number of ARV pills prescribed*, *number of days not on ART*, *duration on ART*, *ART status*, *nutrition status*, *WHO stage*, and *HIV testing or referral source*.

The selected variables again emphasize longitudinal treatment adherence and healthcare engagement patterns, particularly ART refill behavior, scheduled visits, and duration on treatment. Compared with the ST-Bernoulli estimator, the hard-concrete estimator selected a slightly broader set of adherence-related variables while maintaining sparse near-binary selection behavior.

TABLE 15
Selected variable groups of the hard-concrete estimator on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Demographics	0.969	Yes
Marital status	0.694	Yes
Pregnancy status	0.990	Yes
Visit dates	0.997	Yes
Scheduled appointments	0.976	Yes
Visit type	0.975	Yes
Number of ARV pills prescribed on each visit date	0.995	Yes
Number of days not on ART	0.993	Yes
The patient has been on ART for	0.988	Yes
The number of days since HIV diagnosis	0.670	Yes
Nutrition status	0.770	Yes
WHO stage	0.935	Yes
HIV testing or referral source	0.996	Yes

7.9. Qwen3-8B: Straight-through Bernoulli estimator results We additionally evaluated the straight-through (ST) Bernoulli estimator on the HIV cohort with Qwen3-8B. The grid search considered $\lambda \in \{0.001, 0.002, 0.003, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$. All experiments used seed = 42, batch size = 1, 8 training epochs, and a selection threshold of 0.5 on the learned keep probabilities. The minimum GBIC was achieved at $\lambda = 0.005$, yielding 12 selected variable groups: *demographics*, *marital status*, *pregnancy status*, *visit dates*, *scheduled appointments*, *visit type*, *TB TPT status*, *ART status*, *weight*, *nutrition status*, *WHO stage*, and *CD4 status*.

The selected variables emphasize longitudinal healthcare engagement, ART monitoring, nutritional status, and immunological progression. Compared with the hard-concrete and deterministic sigmoid estimators, the ST-Bernoulli estimator placed stronger emphasis on clinical progression and monitoring variables such as WHO stage and CD4 status.

8. Selection Variability To assess the stability of the proposed ReinMax-based feature-selection procedure with respect to data splitting, and to separate feature-selection optimization from predictive evaluation, we conducted a repeated-split analysis on the POINTER cohort using LLaMA-3.1-8B. POINTER contains 188 participants. For each of 10 independent random splits, 150 participants (approximately 80%) were used exclusively for feature selection and hyperparameter tuning, while the remaining 38 participants (approximately 20%) were reserved as an independent held-out evaluation set. The gating parameters and selected variable subset were determined using only the feature-selection subset. The resulting subset was then fixed before evaluation on the corresponding held-out subset.

TABLE 16
Selected variable groups of the ST-Bernoulli estimator on the HIV cohort under minimum GBIC.

Group name	Keep probability	Selected
Demographics	0.525	Yes
Marital status	0.859	Yes
Pregnancy status	0.740	Yes
Visit dates	0.879	Yes
Scheduled appointments	0.956	Yes
Visit type	0.887	Yes
TB TPT status	0.956	Yes
ART status	0.973	Yes
Weight	0.967	Yes
Nutrition status	0.924	Yes
WHO stage	0.972	Yes
CD4 status	0.967	Yes

We applied this procedure to the proposed ReinMax method as well as three alternative embedding-level gating approaches: hard-concrete, sigmoid gating, and straight-through Bernoulli gating. These methods provide a direct comparison because each jointly estimates gating parameters and determines the selected subset from the data. For every split and method, we recorded both predictive performance on the held-out evaluation subset and the resulting selected variable groups.

This repeated-split design allows us to quantify two distinct sources of empirical variability: variability in out-of-sample predictive performance and variability in the selected feature subsets themselves. For each variable group k , we compute its selection frequency across the 10 replications as

$$\hat{\pi}_k = \frac{1}{10} \sum_{r=1}^{10} \mathbb{I}\{k \in \widehat{\mathcal{M}}^{(r)}\},$$

where $\widehat{\mathcal{M}}^{(r)}$ denotes the subset selected in replication r . A selection frequency close to one indicates that a variable group is consistently retained across different data splits, whereas lower selection frequencies indicate greater sensitivity to the particular split.

Table 17 summarizes held-out predictive performance across the 10 replications. For $A\beta$, the proposed ReinMax method achieved a mean weighted AUC of 0.7057 ± 0.0672 , compared with 0.6955 ± 0.0754 , 0.7056 ± 0.0633 , and 0.6592 ± 0.0431 for hard-concrete, sigmoid, and straight-through Bernoulli gating, respectively. For meta-tau, the corresponding mean weighted AUC for ReinMax was 0.5423 ± 0.0385 , compared with 0.5389 ± 0.0511 , 0.5323 ± 0.0405 , and 0.5258 ± 0.0760 for the alternative gating approaches. These repeated held-out evaluations are intended to characterize reproducibility and predictive variability rather than to establish statistically significant pairwise differences between methods.

Table 18 reports feature-wise selection frequencies across the same 10 repeated splits. For the proposed ReinMax method, several variable groups were selected consistently across replications, including Demographic information (10/10), Vitals (10/10), Genetic information (8/10), Family history (7/10), and Brain Volume in MRI (7/10). Other feature groups showed greater variability, including Physical Measure Score (6/10), Brain Thickness in MRI (5/10), Cognitive test (3/10), and Medication (2/10). This pattern suggests that the selected subsets contain a relatively stable core of repeatedly retained feature groups, while a smaller set of features is more sensitive to the particular data split.

The alternative gating approaches exhibit a similar combination of stable and variable selections. Demographic information is selected in all 10 splits by all four methods, while

Genetic information, Vitals, and Physical Measure Score are also selected at high frequencies across methods. In contrast, Family history and the MRI-derived feature groups exhibit greater method- and split-dependent variability, and Medication and Other health condition are selected only infrequently.

TABLE 17

Predictive performance across 10 repeated POINTER splits for LLaMA-3.1-8B. In each split, 150 participants are used for feature selection and tuning and 38 participants are reserved for independent evaluation. Values are reported as mean $\pm$ standard deviation across the 10 held-out evaluations.

Biomarker	Method	Weighted AUC	Accuracy	Weighted Specificity
$A\beta$	Hard Concrete	0.6955 ± 0.0754	0.7079 ± 0.0438	0.2959 ± 0.0956
	ReinMax (proposed)	0.7057 ± 0.0672	0.7158 ± 0.0508	0.3053 ± 0.0736
	Sigmoid	0.7056 ± 0.0633	0.7158 ± 0.0444	0.2859 ± 0.0680
	ST Bernoulli	0.6592 ± 0.0431	0.7158 ± 0.0508	0.3663 ± 0.0909
Meta-tau	Hard Concrete	0.5389 ± 0.0511	0.3266 ± 0.0802	0.7012 ± 0.0526
	ReinMax (proposed)	0.5423 ± 0.0385	0.3340 ± 0.0727	0.7117 ± 0.0280
	Sigmoid	0.5323 ± 0.0405	0.3073 ± 0.0723	0.6829 ± 0.0335
	ST Bernoulli	0.5258 ± 0.0760	0.2870 ± 0.1016	0.7152 ± 0.0341

TABLE 18

Feature-selection frequencies across 10 repeated POINTER splits for LLaMA-3.1-8B. Each entry reports the number and percentage of splits in which the corresponding feature group was selected.

Variable group	Hard Concrete	ReinMax (proposed)	Sigmoid	ST Bernoulli
Demographic information	10/10 (100%)	10/10 (100%)	10/10 (100%)	10/10 (100%)
Genetic information	10/10 (100%)	8/10 (80%)	10/10 (100%)	10/10 (100%)
Family history	2/10 (20%)	7/10 (70%)	10/10 (100%)	10/10 (100%)
Vitals	10/10 (100%)	10/10 (100%)	10/10 (100%)	8/10 (80%)
Cognitive test	2/10 (20%)	3/10 (30%)	3/10 (30%)	3/10 (30%)
Physical Measure Score	10/10 (100%)	6/10 (60%)	10/10 (100%)	10/10 (100%)
Medication	1/10 (10%)	2/10 (20%)	0/10 (0%)	0/10 (0%)
Brain Volume in MRI	2/10 (20%)	7/10 (70%)	6/10 (60%)	8/10 (80%)
Brain Thickness in MRI	5/10 (50%)	5/10 (50%)	1/10 (10%)	7/10 (70%)
Other health condition	1/10 (10%)	0/10 (0%)	0/10 (0%)	2/10 (20%)

Overall, the repeated-split analysis complements the primary full-cohort evaluation by explicitly separating feature-selection optimization from predictive assessment and by quantifying the stability of the selected subsets across independent data splits. It therefore provides an empirical assessment of both predictive robustness and feature-selection reproducibility for the proposed ReinMax procedure and the alternative gating mechanisms.

9. Simulation study results

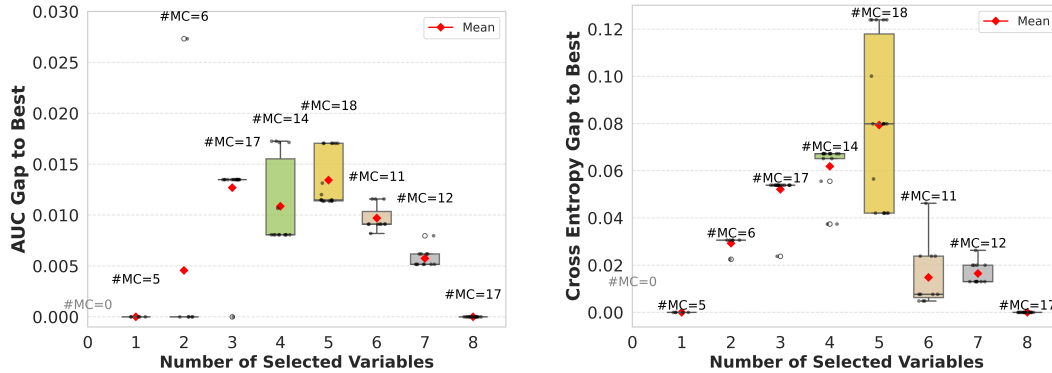
9.1. Fine-tuning dataset format

Instruction: You are an expert in answering whether this patient has diabetes. Please answer in Yes or No.

Input: The Age is 37. The Number of times pregnant is 6. The Diastolic blood pressure is 52. The Triceps skin fold thickness is 23. The Plasma glucose concentration at 2 hours in an oral glucose tolerance test (GTT) is 161. The 2-hour serum insulin is 94. The Body mass index is 23.2. The Diabetes pedigree function is 0.912.

Output: Yes

9.2. Oracle benchmarking: Support vector machine To further show the difference between LLM optimal variable groups and that from the ML method, we present the Support Vector Machine (SVM) results here. For each Monte Carlo sample, the gap is measured between the LLM performance on variables selected via p-maps and the best-performing subset of the same size, based on exhaustive search. Each red diamond indicates the mean gap within that group size. We can see SVM performance has the similar trend as the logistic regression results, indicating that the optimal subset of LLM is different from that identified by the original tabular data by ML method.



(a) AUC gap between the subset selected by SVM and the optimal subset with the same variable size through LLM combinatorial search.

(b) Cross-entropy gap between the subset selected by SVM and the optimal subset with the same variable size through LLM combinatorial search.

Fig 2: Evaluation of SVM feature selection. For each Monte Carlo sample, the gap is measured between the LLM performance on variables selected via p-maps and the best-performing subset of the same size, based on exhaustive search. Each red diamond indicates the mean gap within that group size. Subset sizes with no Monte Carlo samples are marked with #MC = 0.

10. Proof In this section, we analyze the theoretical properties of the relaxed variable-selection formulation introduced in the main paper. Recall that each variable group $k \in \{1, \dots, K\}$ is associated with a binary selection indicator $s_k \in \{0, 1\}$. These indicators form the selection vector $\mathbf{s} = (s_1, \dots, s_K)$, which yields the masked embedding matrix

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

where $\mathbf{H}$ denotes the padding token embedding matrix. The LLM prediction loss under mask $\mathbf{s}$ is defined as

$$f(\mathbf{y}, \mathbf{X}, \mathbf{s}) = \mathcal{L}_{\text{LLM}}(\mathbf{y}, p_{\theta}(\cdot | \mathbf{X}(\mathbf{s}))).$$

The discrete ℓ_0 formulation seeks to minimize:

$$\min_{\mathbf{s} \in \{0, 1\}^K} \left\{ f(\mathbf{y}, \mathbf{X}, \mathbf{s}) + \lambda \|\mathbf{s}\|_0 \right\}.$$

To enable gradient-based optimization, this formulation is relaxed by replacing each deterministic s_k with a Bernoulli random variable: $s_k \sim \text{Bernoulli}(\text{Logit}(\beta_k))$. Consequently, the relaxed objective takes the form:

$$\mathcal{L}(\mathbf{y}, \mathbf{X}, \beta) = \mathbb{E}_{\mathbf{s}}[f(\mathbf{y}, \mathbf{X}, \mathbf{s})] + \lambda \sum_{k=1}^K \text{Logit}(\beta_k).$$

In what follows, we establish that: (i) the relaxation is exact at the global optimum; (ii) this equivalence persists at the population level; and (iii) a PAC-Bayesian generalization bound holds for the relaxed estimator.

10.1. Proof of Lemma 1 To prove this result, we note that the relaxed objective $\mathcal{L}(\mathbf{y}, \mathbf{X}, \beta)$ is simply a convex combination of the discrete objective $f(\mathbf{y}, \mathbf{X}, \mathbf{s}) + \lambda \|\mathbf{s}\|_0$ across all possible configurations of $\mathbf{s} \in \{0, 1\}^K$. As a result, the global minimum of the continuous relaxation exactly matches that of the discrete formulation. To be more specific, define the discrete objective for a fixed selection vector $\mathbf{s} \in \{0, 1\}^K$ as

$$F(\mathbf{s}) = \mathcal{L}_{\text{LLM}}(\mathbf{y}, p_\theta(\cdot | \mathbf{X}(\mathbf{s}))) + \lambda \sum_{k=1}^K s_k.$$

Under our Bernoulli parameterization, $s_k \sim \text{Bernoulli}(p_k)$, with $p_k = \sigma(\beta_k)$. The relaxed objective can then be written as

$$\mathcal{L}(\mathbf{y}, \mathbf{X}, \beta) = \mathbb{E}_{\mathbf{s} \sim \mathbb{P}_\beta} [F(\mathbf{s})] = \sum_{\mathbf{s} \in \{0, 1\}^K} \mathbb{P}_\beta(\mathbf{s}) F(\mathbf{s}),$$

because

$$\lambda \sum_{k=1}^K p_k = \mathbb{E}_{\mathbf{s}} \left[\lambda \sum_{k=1}^K s_k \right].$$

Thus, for every β , the relaxed objective is simply a convex combination of the 2^K deterministic objective values $\{F(\mathbf{s})\}$. Therefore,

$$\mathcal{L}(\mathbf{y}, \mathbf{X}, \beta) \geq \min_{\mathbf{s} \in \{0, 1\}^K} F(\mathbf{s}),$$

since an average cannot be smaller than the smallest quantity being averaged.

Conversely, let $\mathbf{s}^* \in \arg \min_{\mathbf{s}} F(\mathbf{s})$. By letting each Bernoulli probability approach the corresponding deterministic value, i.e.,

$$p_k \rightarrow \begin{cases} 1, & s_k^* = 1, \\ 0, & s_k^* = 0, \end{cases}$$

the distribution $\mathbb{P}_\beta$ concentrates on $\mathbf{s}^*$, so that

$$\mathbb{E}_{\mathbf{s} \sim \mathbb{P}_\beta} [F(\mathbf{s})] \rightarrow F(\mathbf{s}^*).$$

Hence the relaxed problem has the same global optimum value as the original discrete problem. This argument does not rely on submodularity, convexity, monotonicity, or any other structural property of $F(\mathbf{s})$.

10.2. Proof of Corollary 1 Given the assumed integrability of the loss function, a simple application of Fubini's theorem yields the alternative expression:

$$\mathcal{L}(\beta) = \mathbb{E}_{\mathbf{s}} \mathbb{E}_{(\mathbf{y}, \mathbf{X})} \left[\mathcal{L}_{\text{LLM}}(\mathbf{y}, p_\theta(\cdot | \mathbf{X}(\mathbf{s}))) + \lambda \|\mathbf{s}\|_0 \right] = \mathbb{E}_{\mathbf{s}} \mathbb{E}_{(\mathbf{y}, \mathbf{X})} \left[f(\mathbf{y}, \mathbf{X}, \mathbf{s}) + \lambda \|\mathbf{s}\|_0 \right].$$

Consequently, $\mathcal{L}(\beta)$ is a convex combination of $\mathbb{E}_{(\mathbf{y}, \mathbf{X})} [f(\mathbf{y}, \mathbf{X}, \mathbf{s}) + \lambda \|\mathbf{s}\|_0]$ over all possible configurations of $\mathbf{s} \in \{0, 1\}^K$, and hence their global minimum values must coincide.

10.3. *Proof of Theorem 1* To establish this, we first note that in autoregressive LLMs with a finite vocabulary $\mathcal{V}$, the softmax function normalizes over all tokens in the vocabulary. This guarantees that the predicted probability of any token lies in $(0, 1)$, which in turn implies that the sequence-level loss function is inherently bounded:

$$0 \leq f(\mathbf{y}, \mathbf{X}, \mathbf{s}) \leq \mathfrak{c} := -T \cdot \log \varepsilon.$$

Here, ε denotes the minimum representable nonzero probability assigned by the softmax (for instance, $\varepsilon \approx 10^{-8}$ in FP32 arithmetic, yielding a per-token maximum of $-\log \varepsilon \approx 18.42$), and T is the number of tokens in the output sequence. Because T is typically a small, finite integer, $\mathfrak{c}$ serves as a finite bound independent of the sample size n .

Applying Hoeffding's inequality yields:

$$\mathbb{P}_{(\mathbf{y}, \mathbf{X})} [|\Delta_n(\mathbf{s})| > t] \leq 2 \exp \left(-\frac{2nt^2}{\mathfrak{c}^2} \right),$$

where we define

$$\Delta_n(\mathbf{s}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{s}) - \mathbb{E}_{(\mathbf{y}, \mathbf{X})} [f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{s})]$$

to streamline notation. The sub-Gaussian tail established above implies a bound on the Orlicz norm:

$$\mathbb{E}_{(\mathbf{y}, \mathbf{X})} \left[\exp \left(\frac{n}{\mathfrak{c}^2} \Delta_n(\mathbf{s})^2 \right) \right] \leq \mathfrak{c}',$$

where $\mathfrak{c}'$ is a universal constant (specifically, $\mathfrak{c}' \leq 3$). Because the right-hand side is independent of $\mathbf{s}$, taking the expectation over any data-independent prior distribution $\mathbb{P}_0$ yields:

$$\mathbb{E}_{\mathbf{s} \sim \mathbb{P}_0} \mathbb{E}_{(\mathbf{y}, \mathbf{X})} \left[\exp \left(\frac{n}{\mathfrak{c}^2} \Delta_n(\mathbf{s})^2 \right) \right] \leq \mathfrak{c}'.$$

By exchanging the order of expectations via Fubini's theorem and applying Markov's inequality, we obtain that with probability at least $1 - \delta$ over the data distribution:

$$\mathbb{E}_{\mathbf{s} \sim \mathbb{P}_0} \left[\exp \left(\frac{n}{\mathfrak{c}^2} \Delta_n(\mathbf{s})^2 \right) \right] \leq \frac{\mathfrak{c}'}{\delta}.$$

Next, we apply the Donsker-Varadhan inequality to the posterior distribution $\mathbb{P}$, which gives:

$$\mathbb{E}_{\mathbf{s} \sim \mathbb{P}} \left[\frac{n}{\mathfrak{c}^2} \Delta_n(\mathbf{s})^2 \right] \leq \text{KL}(\mathbb{P} \parallel \mathbb{P}_0) + \log \mathbb{E}_{\mathbf{s} \sim \mathbb{P}_0} \left[\exp \left(\frac{n}{\mathfrak{c}^2} \Delta_n(\mathbf{s})^2 \right) \right].$$

Combining this with our Markov bound, and applying Jensen's inequality (which guarantees $|\mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\Delta_n(\mathbf{s})]|^2 \leq \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\Delta_n(\mathbf{s})^2]$), implies that with probability at least $1 - \delta$:

$$|\mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\Delta_n(\mathbf{s})]| \leq \mathfrak{c} \sqrt{\frac{\text{KL}(\mathbb{P} \parallel \mathbb{P}_0) + \log(\mathfrak{c}'/\delta)}{n}}.$$

10.4. *Proof of Corollary 2* By the definition of the KL divergence,

$$\text{KL}(\mathbb{P} \parallel \mathbb{P}_0) = \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\log \mathbb{P}(\mathbf{s}) - \log \mathbb{P}_0(\mathbf{s})] \leq \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [-\log \mathbb{P}_0(\mathbf{s})],$$

where the inequality follows from the fact that $\log \mathbb{P}(\mathbf{s}) \leq 0$ for all $\mathbf{s} \in \{0, 1\}^K$. Since the prior $\mathbb{P}_0$ factorizes as i.i.d. $\text{Bernoulli}(p_0)$ with $p_0 = s_0/K$, its probability mass function is given by

$$\mathbb{P}_0(\mathbf{s}) = \prod_{k=1}^K p_0^{s_k} (1 - p_0)^{1-s_k} = p_0^{\|\mathbf{s}\|_0} (1 - p_0)^{K - \|\mathbf{s}\|_0}.$$

Taking the negative logarithm yields

$$-\log \mathbb{P}_0(\mathbf{s}) = -\|\mathbf{s}\|_0 \log p_0 - (K - \|\mathbf{s}\|_0) \log(1 - p_0).$$

Substituting this back into our initial bound gives

$$\begin{aligned} \text{KL}(\mathbb{P} \parallel \mathbb{P}_0) &\leq \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} \left[-\|\mathbf{s}\|_0 \log p_0 - (K - \|\mathbf{s}\|_0) \log(1 - p_0) \right] \\ &= -\mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\|\mathbf{s}\|_0] \log p_0 - (K - \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\|\mathbf{s}\|_0]) \log(1 - p_0). \end{aligned}$$

Let $m = \mathbb{E}_{\mathbf{s} \sim \mathbb{P}} [\|\mathbf{s}\|_0]$, and recall that the premise of the corollary guarantees $m \leq s_0$. Set $q = p_0 = s_0/K$ and define $\alpha = m/K$. Note that $\alpha \leq s_0/K = q$. The previous inequality can now be rewritten as

$$\text{KL}(\mathbb{P} \parallel \mathbb{P}_0) \leq K \left(-\alpha \log q - (1 - \alpha) \log(1 - q) \right).$$

To maximize this upper bound with respect to α , we define the function

$$g(\alpha) := -\alpha \log q - (1 - \alpha) \log(1 - q), \quad \alpha \in [0, 1].$$

Its derivative with respect to α is

$$g'(\alpha) = -\log q + \log(1 - q) = \log \left(\frac{1 - q}{q} \right).$$

In the sparse regime of interest where $q \leq 1/2$ (equivalently, $s_0 \leq K/2$), we have $(1 - q)/q \geq 1$, which implies $g'(\alpha) \geq 0$. Thus, $g(\alpha)$ is monotonically non-decreasing on $[0, q]$. Because $\alpha \leq q$, we obtain

$$g(\alpha) \leq g(q) = -q \log q - (1 - q) \log(1 - q) = H(q),$$

where $H(q)$ is the binary entropy function. Consequently, the KL divergence is bounded by

$$\text{KL}(\mathbb{P} \parallel \mathbb{P}_0) \leq K H(q).$$

It is a well-established information-theoretic bound that for any $q \in (0, 1/2]$,

$$H(q) \leq q \log \frac{e}{q}.$$

Applying this bound with $q = s_0/K$ yields

$$\text{KL}(\mathbb{P} \parallel \mathbb{P}_0) \leq K q \log \frac{e}{q} = s_0 \log \frac{eK}{s_0},$$

which closes the proof.

REFERENCES

- BENGIO, Y., LÉONARD, N. and COURVILLE, A. (2013). Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. arXiv:1308.3432.
- CUI, H., SHEN, Z., ZHANG, J., SHAO, H., QIN, L., HO, J. C. et al. (2025). LLMs-Based Few-Shot Disease Predictions Using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction. In *AMIA Annual Symposium Proceedings* 319–328.
- GAO, S., LIN, C.-H., HUA, T., TANG, Z., SHEN, Y., JIN, H. et al. (2024). DISP-LLM: Dimension-Independent Structural Pruning for Large Language Models. In *Proceedings of the 38th International Conference on Neural Information Processing System* 72219–72244.
- 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.
- 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.