

---

# LLM-Lasso: A Robust Framework for Domain-Informed Feature Selection and Regularization

---

Erica Zhang <sup>\*1</sup> Ryunosuke Goto <sup>\*2</sup> Naomi Sagan <sup>3</sup> Jurik Mutter <sup>4</sup> Nick Phillips <sup>4</sup> Ash Alizadeh <sup>4</sup>  
Kangwook Lee <sup>5</sup> Jose Blanchet <sup>1</sup> Mert Pilanci <sup>†3</sup> Robert Tibshirani <sup>†26</sup>

## Abstract

We introduce LLM-Lasso, a novel framework that leverages large language models (LLMs) to guide feature selection in Lasso  $\ell_1$  regression. Unlike traditional methods that rely solely on numerical data, LLM-Lasso incorporates domain-specific knowledge extracted from natural language, enhanced through a retrieval-augmented generation (RAG) pipeline, to seamlessly integrate data-driven modeling with contextual insights. Specifically, the LLM generates penalty factors for each feature, which are converted into weights for the Lasso penalty using a simple, tunable model. Features identified as more relevant by the LLM receive lower penalties, increasing their likelihood of being retained in the final model, while less relevant features are assigned higher penalties, reducing their influence. Importantly, LLM-Lasso has an internal validation step that determines how much to trust the contextual knowledge in our prediction pipeline. Hence it addresses key challenges in robustness, making it suitable for mitigating potential inaccuracies or hallucinations from the LLM. In various biomedical case studies, LLM-Lasso outperforms standard Lasso and existing feature selection baselines, all while ensuring the LLM operates without prior access to the datasets. To our knowledge, this is the first approach to effectively integrate conventional feature selection techniques directly with LLM-based domain-specific reasoning.

## 1. Introduction

Feature selection remains a cornerstone of statistical learning, enabling models to focus on the most relevant predictors while reducing complexity and improving interpretability (Guyon et al., 2007; Chandrashekhar & Sahin, 2014; Li & Liu, 2015). Among the various methods for feature selection, Lasso regression has gained widespread adoption for various reasons. It delivers a feature selection approach while simultaneously building a predictive model. The Lasso approach is interpretable and computationally efficient because it automatically selects a suitable linear model with a sparse set of coefficients. Selection is performed by solving a straightforward convex optimization problem that promotes sparsity by penalizing the size of the regression coefficients (Tibshirani, 1996; Bühlmann & Van De Geer, 2011; Hastie et al., 2015). As with any supervised learning model, the traditional Lasso approach is based only on the training data. It is natural to consider expert knowledge to inform the feature selection task. However, this is challenging to do in a systematic and scalable way that safeguards against introducing potential biases. *The goal of this paper is to meet this challenge. We do this by taking advantage of current advances in generative AI.*

The development of large language models (LLMs) trained on a large scale of unstructured text offers a transformative opportunity to augment traditional feature selection techniques in systematic and scalable way. Transformer-based pre-trained LLMs, such as GPT-4 (OpenAI, 2023b) and LLaMA-2 (Touvron et al., 2023) have demonstrated impressive abilities in encoding domain knowledge and contextual relationships and generalizing to a wide range of unseen tasks in a variety of domains (Vaswani et al., 2017; Brown et al., 2020; Radford et al., 2019; Manikandan et al., 2023), including various challenging reasoning tasks (Wei et al., 2022; Lewkowycz et al., 2022; Suzgun et al., 2023), prediction tasks that require domain-specific knowledge (Petroni et al., 2019; Dinh et al., 2022; Chen & Zou, 2024; Theodoris et al., 2023; Cui et al., 2024), and, more recently, feature selection (Choi et al., 2022; Jeong et al., 2024; Li et al., 2024; Liu et al., 2024; Han et al., 2024). Based on the idea that LLMs might possess relevant knowledge for the

---

<sup>\*</sup> Equal Contribution <sup>†</sup> Equal Advising <sup>1</sup>Department of Management Science and Engineering, Stanford University, Stanford, USA <sup>2</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, USA <sup>3</sup>Department of Electrical Engineering, Stanford University, Stanford, USA <sup>4</sup>Divisions of Oncology and Hematology, Stanford University School of Medicine, Stanford, USA <sup>5</sup>Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, USA  
<sup>6</sup>Department of Statistics, Stanford University, Stanford, USA.

task at hand, (Dinh et al., 2022) proposed fine-tuning an LLM with training data, feature names, and task descriptions and showed improvements in prediction performance. Specific to feature selection, (Choi et al., 2022) introduces the LMPriors framework, which selects features by analyzing log-probability differences when generating “Y” (Yes) or “N” tokens, effectively admitting or rejecting certain features based solely on task descriptions, feature names, and a few-shot examples. Based on this, (Jeong et al., 2024) and (Liu et al., 2024) enable feature selection with proprietary LLMs where internal token probabilities are inaccessible. Specifically, (Jeong et al., 2024) proposes three prompting strategies that rely only on textual information, bypassing the need for data access, to directly utilize the output of the generated text without further processing. Meanwhile, (Liu et al., 2024) introduces a framework that leverages LLMs for direct feature filtering based on test scores. These methods for incorporating LLMs into feature selection have demonstrated promising results, showing that LLMs can rival leading statistical feature selection techniques, even in zero-shot settings where they lack direct access to the data (Choi et al., 2022; Jeong et al., 2024). Collectively, these studies underscore the potential of LLMs to encode a rich set of relevant, task-specific information, augmenting traditional supervised learning approaches and enabling their application across a wide range of downstream tasks.

In this work, we build on these insights by introducing LLM-Lasso, a novel framework for LLM-powered feature selection that integrates LLM-derived penalty factors into Lasso penalty terms, allowing the seamless fusion of knowledge-based insights with traditional data-driven supervised learning methodologies. LLM-Lasso focuses on feature selection in the context of a specific family of models (the ones in the Lasso framework). This strategy allows us to ensure that the feature selection approach is effective for the downstream task of selecting a model in the family. Using other feature selection approaches that are not aligned with the downstream task of interest could result in features that are not relevant to the task. For example, the use of random forests could result in the choice of a feature that is perceived relevant because of its non-linear interactions with other features. But if the task at hand is to select a linear model, such a feature may not be useful.

LLM-Lasso assumes black-box access to the LLMs and utilizes an optional retrieval-augmented generation (RAG) pipeline (Lewis et al., 2020; Shuster et al., 2022; Wu et al., 2024; Siriwardhana et al., 2023) to extract domain-specific knowledge via LLMs, which is then used to inform Lasso regularization. This approach enables fine-tuning through cross-validation, ensuring adaptability and robustness. Although previous research on the adoption of LLMs for feature selection has shown promising results, these methods face a major bottleneck: they rely solely on context de-

scriptions of the task and features to make standalone feature selection decisions, without incorporating data-driven safeguards to ensure robustness against inaccuracies in the generated responses. This limitation makes these methods vulnerable to LLM hallucinations, that is, fabricating nonexistent facts, a common weakness even in the most advanced LLMs (Huang et al., 2024; Yao et al., 2024). This can also raise concerns about their reliability, particularly in scenarios where the data is unexpected or errors have occurred during data collection. All of the aforementioned inaccuracies can be especially detrimental in fields such as biomedicine, where precision and reliability are critical.

**Main Contributions.** In this paper, we address the key bottleneck of robustness and demonstrate the effectiveness of LLM-Lasso through experiments focused on various tasks, including an unpublished biomedical dataset with feature dimensions that are at least an order of magnitude larger than those in previous studies. Focused on logistic regression classification tasks, our results show that LLM-Lasso outperforms standard Lasso in feature selection accuracy and predictive performance. While biomedicine is a key use case, our framework is broadly applicable to other domains where external knowledge aids feature selection and extends to general supervised learning methods. Our main contributions are as follows.

1. We introduce LLM-Lasso, a scalable and robust framework for LLM-powered feature selection that effectively combines contextual knowledge with data-driven insights, enabling the direct integration of LLMs into traditional supervised learning methods.
2. We build into LLM-Lasso an internal validation step that determines how much to trust the contextual knowledge in our prediction pipeline. This validation step is tested using adversarial examples.
3. We show that LLM-Lasso consistently outperforms standard Lasso and other popular feature selection methods across various datasets.
4. We demonstrate LLM-Lasso’s clinical applicability by identifying key diagnostic factors distinct from those selected by traditional Lasso or standard LLM-based methods, enhancing precision and facilitating scientific discovery.

The paper is structured as follows: Section 2 reviews Lasso, RAG, and presents a schematic of our procedure. Section 3 details our methodology. Section 4 determines penalty factor selection via simulations and exemplify robustness of LLM-Lasso through adversarial experiments. Section 5 evaluates LLM-Lasso across diverse datasets and LLMs. Finally, Section 6 summarizes our findings

## 2. Preliminaries

### 2.1. Supervised Data-Driven Learning

We consider a generic data-driven supervised learning procedure. Given a dataset  $\mathcal{D}$  consisting of  $n$  data points  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$  drawn from an underlying distribution  $p(\cdot|\theta)$ , our goal is to estimate parameters  $\theta \in \Theta$  through a learning procedure, defined as  $f : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \Theta$  that minimizes the predictive error on observed data. Specifically, the learning objective is defined as follows:

$$\hat{\theta}_f := f(\mathcal{D}) = \arg \min_{\theta} \mathcal{L}(\theta, \mathcal{D}), \quad (1)$$

where  $\mathcal{L}(\cdot, \mathcal{D}) := \sum_{i=1}^n \mathcal{L}(\cdot, (x_i, y_i))$ , and  $\mathcal{L}$  is a loss function quantifying the error between predictions and true outcomes. Here,  $\hat{\theta}_f$  is the parameter that best explains the observed data pairs  $(x_i, y_i)$  according to the chosen loss function  $\mathcal{L}(\cdot)$ .

**Feature Selection.** Feature selection aims to improve model  $f$ 's predictive performance while minimizing redundancy. State-of-the-art techniques fall into four categories: (i) filter methods, which rank features based on statistical properties like Fisher score (Duda et al., 2001; Song et al., 2012); (ii) wrapper methods, which evaluate model performance on different feature subsets (Kohavi & John, 1997); (iii) embedded methods, which integrate feature selection into the learning process using techniques like regularization (Tibshirani, 1996; Lemhadri et al., 2021); and (iv) hybrid methods, which combine elements of (i)-(iii) (Singh & Singh, 2021; Li & Ren, 2022). This paper focuses on embedded methods via Lasso, benchmarking against approaches from (i)-(iii).

### 2.2. Language Modeling

Language modeling aims to approximate the true distribution of natural language  $p_{\text{text}}(x)$  by learning  $p_{\text{LM}}(x)$ , a probability distribution over text sequences  $x = (X_1, \dots, X_{|x|})$ . Modern large language models, trained on diverse datasets (Gao et al., 2020), exhibit strong generalization across domains, acquire contextual knowledge, and perform zero-shot learning—solving new tasks using only task descriptions—or few-shot learning by leveraging a small number of demonstrations (Brown et al., 2020).

**Retrieval-Augmented Generation (RAG).** Retrieval-Augmented Generation (RAG) enhances the performance of generative language models by integrating a domain-specific information retrieval process (Lewis et al., 2020). The RAG framework comprises two main components: *retrieval*, which extracts relevant information from external knowledge sources, and *generation*, where an LLM generates context-aware responses using the prompt combined with

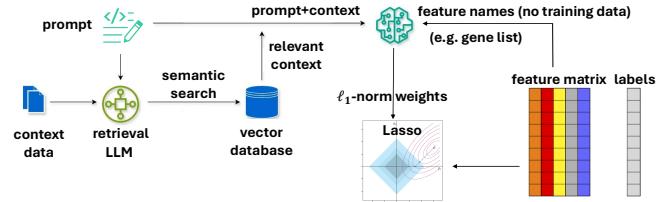


Figure 1. Retrieval Augmented Generation (RAG) based  $\ell_1$ -norm weights (penalty factors) for Lasso. Only feature names—no training data—are included in LLM prompt.

the retrieved context. Documents are indexed through various databases, such as relational, graph, or vector databases (Khattab & Zaharia, 2020; Douze et al., 2024; Peng et al., 2024), enabling efficient organization and retrieval via algorithms like semantic similarity search to match the prompt with relevant documents in the knowledge base. RAG has gained much traction recently due to its demonstrated ability to reduce incidence of hallucinations and boost LLMs' reliability as well as performance (Huang et al., 2023; Zhang et al., 2023).

### 2.3. Task-Specific Data-Driven Learning

LLM-Lasso aims to bridge the gap between data-driven supervised learning and the predictive capabilities of LLMs trained on rich metadata. This fusion not only enhances traditional data-driven methods by incorporating key task-relevant contextual information often overlooked by such models, but can also be especially valuable in low-data regimes, where the learning algorithm  $f : \mathcal{D} \rightarrow \Theta$  (seen as a map from datasets  $\mathcal{D}$  to the space of decisions  $\Theta$ ) is susceptible to overfitting.

The task-specific data-driven learning model  $\tilde{f} : \mathcal{D} \times \mathcal{D}_{\text{meta}} \rightarrow \Theta$  can be described as a metadata-augmented version of  $f$ , where a link function  $h(\cdot)$  integrates metadata (i.e.  $\mathcal{D}_{\text{meta}}$ ) to refine the original learning process. This can be expressed as:

$$\tilde{f}(\mathcal{D}, \mathcal{D}_{\text{meta}}) := \mathcal{T}(f(\mathcal{D}), h(\mathcal{D}_{\text{meta}})),$$

where the functional  $\mathcal{T}$  takes the original learning algorithm  $f(\mathcal{D})$  and transforms it into a task-specific learning algorithm  $\tilde{f}(\mathcal{D}, \mathcal{D}_{\text{meta}})$  by incorporating the metadata  $\mathcal{D}_{\text{meta}}$ .

There are multiple approaches to formulate  $\mathcal{T}$  and  $h$ . For instance, LMPriors (Choi et al., 2022) designed  $h$  and  $\mathcal{T}$  such that  $h(\mathcal{D}_{\text{meta}})$  first specifies which features to retain (based on a probabilistic prior framework), and then  $\mathcal{T}$  keeps the selected features and removes all the others from the original learning objective of  $f$ . Note that this approach inherently is restricted as it selects important features solely based on  $\mathcal{D}_{\text{meta}}$  without seeing  $\mathcal{D}$ .

In contrast, we directly embed task-specific knowledge into

the optimization landscape through regularization by introducing a structured inductive bias. This bias guides the learning process toward solutions that are consistent with metadata-informed insights, without relying on explicit probabilistic modeling. Abstractly, this can be expressed as:

$$\hat{\theta}_{\tilde{f}} := \tilde{f}(\mathcal{D}, \mathcal{D}_{\text{meta}}) = \arg \min_{\theta} \mathcal{L}(\theta, \mathcal{D}) + \lambda R(\theta, \mathcal{D}_{\text{meta}}), \quad (2)$$

where  $\lambda$  is a regularization parameter,  $R(\cdot)$  is a regularizer, and  $\theta$  is the prediction parameter.

### 3. Methodology

In this section, we introduce the LLM-Lasso framework by outlining its two key components: (i) the core statistical model that integrates an LLM-informed penalty into Lasso; and (ii) the general pipeline for training a task-specific LLM on expert knowledge base. An overview schematic is shown in Figure 1.

#### 3.1. The LLM-Lasso

We focus on the supervised learning framework introduced earlier in Section 2.1 with input feature  $X \in \mathbb{R}^{n \times p}$  and response  $Y \in \mathbb{R}^n$ . The Lasso is a shrinkage method that places an  $\ell_1$  penalty on the coefficient, which causes some of the coefficient to be exactly zero. The objective function of the Lasso is given by:

$$\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}. \quad (3)$$

To incorporate prior knowledge of the relationship between  $X$  and  $Y$  into the learning of a prediction model  $f : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \Theta$ , one can enhance the Lasso by assigning penalty factors to each coefficient in the  $\ell_1$  penalty (Zou, 2006). The objective function of the Lasso with penalty factors is:

$$\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \right\}. \quad (4)$$

While penalty factors can be manually assigned based on prior knowledge, this approach becomes impractical when  $p$  is large. To address this, we leverage LLMs to streamline the integration of task-specific knowledge by generating LLM-informed penalty factors or importance scores for all predictors using domain-specific insights. The key modeling challenge is determining how best to effectively inform the underlying data-driven shrinkage method. In the following, we introduce two approaches for modeling LLM-informed penalty factors, whose performance is evaluated through simulation studies in Section 4.

**Inverse importance penalty factors.** First, simply taking the inverse of the importance scores and their powers is useful. In this case, the penalty factors would take the form  $(\mathcal{I}_j)^{-\eta}$ , where  $\mathcal{I}_j$  is the importance score for feature  $j$  and the power  $\eta \geq 0$  would be determined via cross-validation. A large  $\eta$  would indicate heavy reliance of the LLM-Lasso on the importance scores, whereas a value of  $\eta$  close to 0 would indicate minimal reliance on the importance scores.

**ReLU-form penalty factors.** Another approach to defining the penalty factors involves interpolating between penalty factors derived from the LLM and those used in Lasso regression with equal  $\ell_1$ -norm weights by applying a rectification operation. Specifically, we use a rectified linear unit (ReLU) to achieve this. Suppose  $\tilde{w}_{(j)}$  is the penalty factor after ReLU-form processing of the  $j^{\text{th}}$  most important feature, as determined from the scores obtained from the LLM. Here, the largest penalty factor (and therefore the penalty factor of the least important feature) would be  $\tilde{w}_{(p)}$ , which is greater than 1. We can define  $\tilde{w}_{(j)}$  as  $\tilde{w}_{(j)} = 1 + \frac{(j-(1-\gamma)p)_+}{\gamma p} \cdot (\tilde{w}_{(p)} - 1)$ , where  $\gamma \in (0, 1)$  is the ReLU threshold. The lower the importance of a feature, the more likely its coefficient will be 0. The threshold  $\gamma$  can be chosen via cross-validation.

#### 3.2. Task-Specific LLM

To develop a task-specific LLM that provides accurate answers grounded in rigorous and extensive domain knowledge, we focus on two key aspects: prompt engineering and knowledge-base embedding via RAG.

##### 3.2.1. PROMPT ENGINEERING.

Prompting is an efficient and effective approach for adapting pretrained LLMs to tackle new tasks not encountered during training (Radford et al., 2019; Liu et al., 2023). In our experiment, we employ a zero-shot approach for large-scale experiment on biomedical dataset, where the acquisition of ground truth is often infeasible, and a few-shot approach for small-scale experiments. By default, we use greedy decoding—i.e., sampling with temperature  $T = 0$ —due to its simplicity and deterministic behavior, making it well-suited for replication and ablation studies. In addition, we incorporate chain-of-thought (CoT) prompting (Wei et al., 2022), a technique shown to significantly enhance performance on complex reasoning tasks.

For all classification tasks, our full prompt template consists of three components—user, retriever (if RAG is used), and system—and is defined as follows:

$$\mathcal{P}^{\text{full}} = \text{prompt}(\mathcal{Q}^{\text{user}}(\mathcal{A}(\phi, c)), \mathcal{C}^{\text{retriever}}(k, \mathcal{R}(\phi, c)), \mathcal{H}^{\text{system}}),$$

where (i).  $\mathcal{Q}^{\text{user}}$  stands for user query, which is comprised of

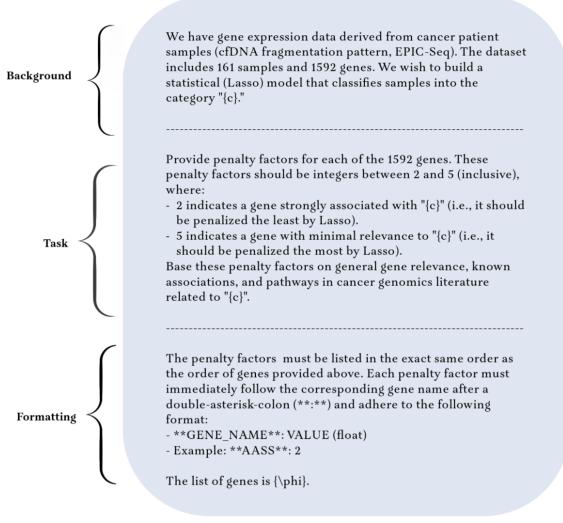


Figure 2. An example task description ( $\mathcal{A}$ ) prompt.

$\mathcal{A}$ , a task description prompt that takes features  $\phi$  and categories  $c$  as inputs; (ii).  $\mathcal{C}^{\text{retriever}}$  represents the top  $k$  retrieved contexts via a semantic similarity search of retrieval prompt  $\mathcal{R}(\phi, c)$  with the retrieval knowledge base; and (iii).  $\mathcal{H}^{\text{system}}$  summarizes past queries and responses, enacted through a conversational buffer. Under this framework, prompt engineering consists of three components:  $\mathcal{A}$  (task description), and  $\mathcal{R}$  (retrieval prompt). Component  $\mathcal{A}$  follows the general structure in Figure 2, where it is composed of a background description of the dataset, the assigned the task, and formatting instructions. We refer the readers to Appendix B for a more detailed description.

### 3.2.2. KNOWLEDGE-BASE EMBEDDING VIA RAG

We use the standard RAG pipeline to create an optional task-specific knowledge base embedding for our prediction task. We provide a brief overview of what RAG is and our specific pipeline. RAG provides LLMs with informative contextual information by selectively choosing the relevant documents from a database. This is critical, as today’s LLMs still struggle to handle very long contexts and generally cannot take an entire database as part of the input prompt. We now describe the specific RAG pipeline we use.

**Preprocessing (Embedding and Indexing)** Given a knowledge base consist of  $N$  text documents,  $\{D_i\}_{i=1}^N$ , we obtain their  $d$ -dimensional semantic embeddings  $\{d_i\}_{i=1}^N = \{E(D_i)\}_{i=1}^N$  via an embedding function  $E : \text{Text} \rightarrow \mathbb{R}^d$ . Here, we use the OpenAI embeddings off-the-shelf (OpenAI, 2024). Once the semantic embedding vectors are obtained, we apply the the Hierarchical Navigable Small World (HNSW) algorithm (Malkov & Yashunin, 2018), imple-

mented in chromadb, to enable sublinear complexity for semantic similarity search.

**Retrieval** At retrieval time, given a query vector  $q \in \mathbb{R}^d$ , the semantic similarity between  $q$  and the stored embeddings  $\{d_i\}_{i=1}^N$  is computed as  $\text{Sim}(q, d_i) = \frac{q^T d_i}{\|q\|_2 \|d_i\|_2}$ . The top  $k$  documents with the highest similarity scores are retrieved and supplied as context  $\mathcal{C}$ .

Throughout the paper, we adhere to the following naming convention: LLM-Lasso (Plain) refers to a pipeline without RAG, while LLM-Lasso (RAG) denotes a pipeline incorporating RAG. The performance of RAG in our framework highly depends on the retrieval prompt and the relevance of the retrieved documents. Figure 1 illustrates LLM-Lasso (RAG). Due to space constraints, a detailed discussion is provided in Appendix D.

## 4. Simulations

In this section, we aim to answer the following questions through simulation experiments: (i). What is the optimal penalty factor formulation for incorporating domain knowledge from LLMs into Lasso? (ii). Is LLM-Lasso robust to adversarial datasets where the features provided to the LLM are uninformative with respect to the data, or where the data itself is misaligned with the features presented?

### 4.1. Penalty Factor Form Simulations

We run simulations to find the adequate form of penalty factors. Based on the simulations, we use the inverse importance penalty factors to compare the LLM-Lasso to the baseline models. We defer the details to Appendix A.

### 4.2. Adversarial Simulations

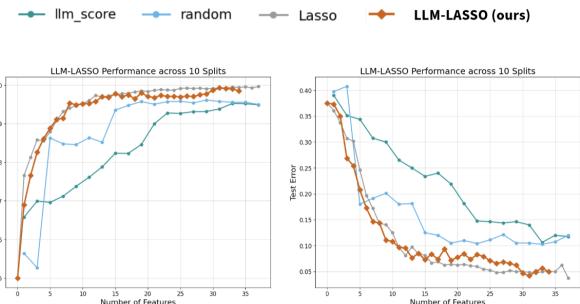


Figure 3. Adversarial simulation experiment using the DLBCL vs. MCL dataset

To showcase the robustness of our method in scenarios in which the LLM fails to produce meaningful results, we perform adversarial data corruption simulations. As a base

dataset, we use the myeloid cell leukemia (MCL) vs. diffuse large B-cell lymphoma (DLBCL) task from the Lymphoma dataset (Table 3). Of the 1592 gene features, we select the 800 most relevant based on presence in documents retrieved from the OMIM (Online Mendelian Inheritance in Man) knowledge base (see Section 5.4.1). We replace those genes with random base64 strings, ensuring via OMIM that the strings are not real gene names (see Figure 4).

We perform classification via LLM-Lasso and LLM-Score, as described in Section 5.4, using the GPT-4o model. Both methods are given the corrupted gene name list. For illustrative purposes, we also include a random feature selection baseline. The resulting misclassification error and AUROC plots can be found in Figure 9. Though half the genes names given to the LLM are corrupted, the accuracy of LLM-Lasso remains comparable to Lasso, whereas LLM-Score performs noticeably worse than random feature selection. We observe that for both LLM-Lasso and LLM-Score, the LLM analysis of the corrupted genes is heavily based on hallucinations, examples of which are in Figure 5. LLM-Lasso, however, remains robust to the corruptions, while the accuracy of LLM-Score degrades substantially.

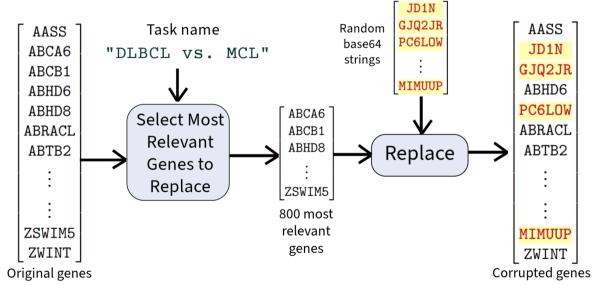


Figure 4. Gene name corruption for adversarial simulations.

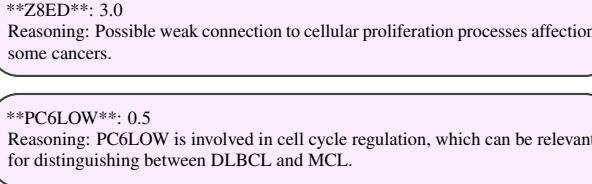


Figure 5. GPT-4o hallucination for corrupted gene names: an LLM-Lasso penalty factor, followed by an LLM-Score importance score. Even though both genes, Z8ED and PC6LOW are fake, the LLM hallucinates justification for their relevance to the task.

## 5. Experiments

In this section, we demonstrate the effectiveness of our proposed framework, LLM-Lasso, through a series of experiments. These include small-scale tests ( $\sim 20$  features) and large-scale experiments ( $> 1000$  features) which leverage

an unpublished biomedical dataset.

### 5.1. Model Details

For the experiment, we sample a combination of closed-source and open-source LLMs:

1. o1 (OpenAI, 2024) –<sup>1</sup>,
2. GPT-4o (OpenAI, 2023b): –,
3. GPT-3.5 (OpenAI, 2023a): –,
4. DeepSeek-R1 (DeepSeek AI, 2025): 671B parameters,
5. LLaMA-3.1 (Meta AI, 2025b): 405B parameters,
6. LLaMA-3 (Meta AI, 2025a): 8B parameters,
7. Qwen Models (Alibaba DAMO Academy, 2025): 72B parameters.

We use all GPT models via OpenAI API calling and all open-source models via OpenRouter API calling via cloud-based inference. We implement RAG using the langchain-community (LangChain Community, 2024) code-base and a self-query retriever as our base method for query construction via Chroma vectorstore.

### 5.2. Baselines

To robustly evaluate our model’s performance, we compare it against baselines from both LLM-based feature selectors and traditional data-driven feature selection methods, with representatives chosen from each of the three main categories, that is, filter, wrapper, and embedded:

1. LLM-Score (Jeong et al., 2024).
2. Filtering by Mutual Information (MI) (Lewis, 1992).
3. Recursive Feature Elimination (RFE) (Guyon et al., 2002).
4. Minimum Redundancy Maximum Relevance selection (MRMR) (Ding & Peng, 2005).
5. Lasso (Tibshirani, 1996).
6. Random feature selection.

For standalone feature selectors such as LLM-Select, MI, RFE, MRMR, and Random, we follow the evaluation procedures outlined in (Jeong et al., 2024) to ensure a fair comparison: approximately 10% of the total features are selected using each method, and their performance is evaluated by measuring the test performance of a downstream  $\ell_2$ -penalized logistic regression model, with hyperparameters chosen via grid search and cross-validation.

### 5.3. Small-Scale Experiments

We begin with a preliminary evaluation of our plain LLM-based method against baselines using small-scale, low-dimensional public datasets across various domains.

<sup>1</sup>We note that the official parameter counts for closed-source OpenAI models have not been disclosed. Therefore, we omit the model parameter counts.

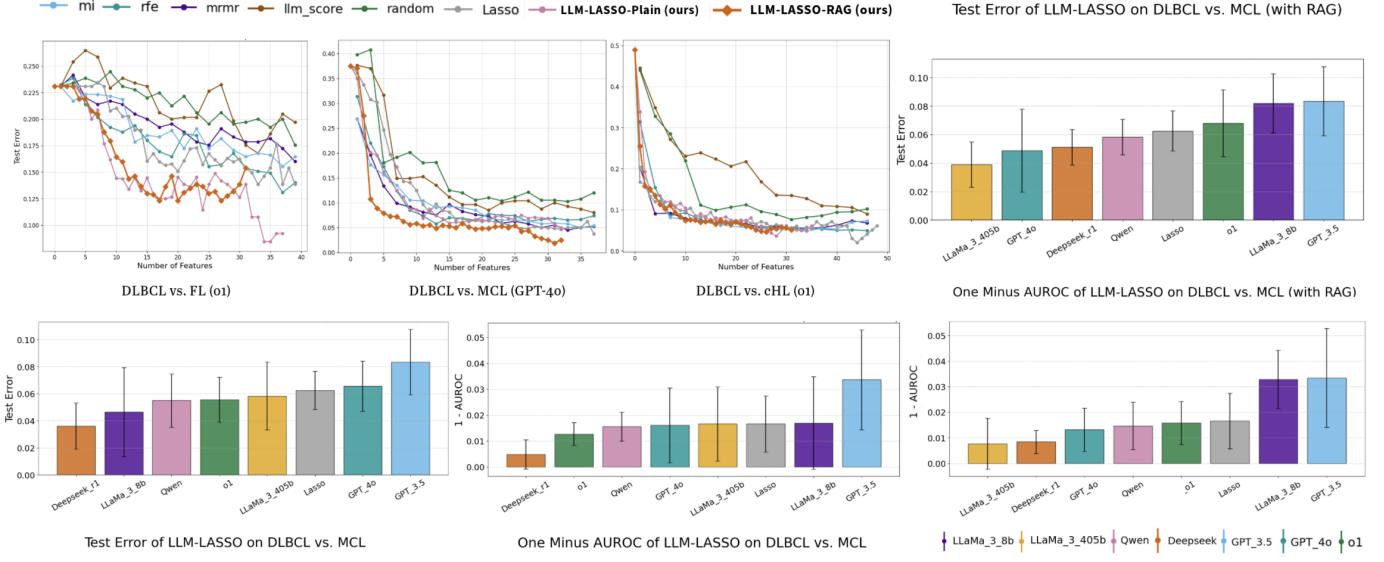


Figure 6. Large-Scale Experiments on Lymphoma Datasets: LLM-Lasso vs. Baselines and Model Ablation Across LLMs. The model ablations display mean misclassification and (one minus) AUROC at 20 features, with error bars for the standard deviation. Lasso is plotted for reference.

This includes three binary classification datasets (Bank, Diabetes, Glioma) and two regression datasets (Wine Quality, Spotify 2024\*). An asterisk (\*) denotes a dataset published after the pretraining-data cutoff for all sampled LLMs (see Table 2), included to mitigate concerns about pretraining-data memorization. A summary of the datasets used can be found in Table 4. We follow the evaluation procedures outlined in Section 5.2 for standalone feature selectors and in Section 5.4 for Lasso-based models. To ensure a fair assessment in the presence of class imbalance, we report the error rate across ten splits along with the AUROC. As shown in Figure 7, GPT-4o-based LLM-Lasso consistently outperforms all sampled datasets and baselines, even when not using the best-performing LLM (see Appendix E for a model ablation study on Spotify).

#### 5.4. Large-Scale Experiments

Gene expression levels can aid in cancer diagnosis and prediction. Moreover, identifying genes predictive of specific cancers or subtypes enhances our understanding of cancer pathophysiology and may facilitate drug discovery by prioritizing key predictive genes. To demonstrate the application of our proposed framework and to show that the strong performance of LLM-Lasso carries over to high-dimensional, complex datasets, we consider the task of cancer diagnosis and classification with gene expression data as features across a range of biomedical tasks using subsets of the following unpublished lymphoma dataset.

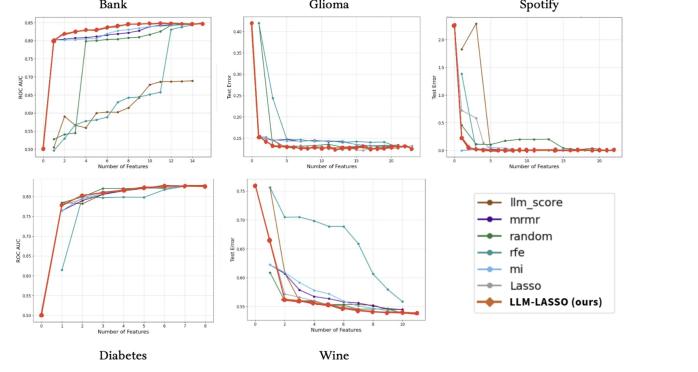


Figure 7. Small-scale experiments on public datasets using GPT-4o for LLM-Lasso (Plain) and LLM-Score.

**Lymphoma (Unpublished)** Follicular lymphoma (FL) is a relatively indolent form of lymphoma that usually does not require intervention, but it could occasionally transform into the more aggressive diffuse large B-cell lymphoma (DLBCL). Using an unpublished dataset of 1592 gene expression levels from 130 lymphoma samples, we use LLM-Lasso to classify tumor samples into DLBCL and FL. Though less clinically significant, we also perform the task of classifying 161 lymphoma samples into DLBCL and mantle cell lymphoma (MCL) and classifying 196 samples into DLBCL and classical Hodgkin lymphoma (cHL) using 1592 gene expression levels. The datasets used are summarized in Table 3.

#### 5.4.1. BUILDING A KNOWLEDGE-BASE FOR RAG

We utilize **OMIM** (Online Mendelian Inheritance in Man), an open-source database of human genes and their disease associations, to build our RAG knowledge base. Using the OMIM API, we extract gene symbols, preferred titles, clinical synopses, and detailed genetic and phenotypic relationships, storing the data in structured JSON format for efficient retrieval. The JSON data is indexed, chunked using a recursive text splitter, and ingested into Chroma database to populate the vector store. See Appendix D for more implementation details.

#### 5.4.2. EVALUATION

**Evaluation of LLM performance** We evaluate the performance of LLMs in two parts: (i). the performance comparison of the models surveyed in Section 5.1, and (ii). the performance of RAG in of retrieval quality and relevance.

The histograms of Figure 6 display the average misclassification error and AUROC of LLM-Lasso at 20 features, for the models listed in Section 5.1. Lasso is plotted as a baseline. Larger and more powerful models generally perform better, especially with RAG. Some key exceptions are LLaMa-3-8b, which achieves the second-lowest test error without RAG, and DeepSeek-R1, for which RAG degrades performance. We hypothesize that some models have more nuanced abilities to parse the medical documents provided by RAG, whereas others are harmed by the increased context from the retrieved documents.

As shown in the misclassification error plots of Figure 6, LLM-Lasso RAG outperforms the plain LLM-Lasso in all tasks except DLBCL vs. FL. Appendix E.1 provides an illustrative example justifying this, in which a RAG-enhanced GPT-4o model cites concrete gene interactions. RAG, however, does not unilaterally improve performance. This can result from the knowledge base being ill-suited to the task, issues with long contexts, and retrieval of irrelevant information, as is further discussed in Appendix E.1.

Scaling LLM generation of penalty factors to over a thousand genes also presents unique challenges, as we discuss in Appendix E.2.

**Evaluation of Prediction Performance** To test the prediction performance of the LLM-Lasso, the data is centered and split into the training set and the test set. On the training set, we perform 10-fold cross-validation across the hyperparameter  $\eta \in (0, 1, 2, \dots, 10)$  for penalty factors of the form  $\mathcal{T}^{-\eta}$ . Using the hyperparameter with the lowest cross-validation misclassification rate, we evaluate model performance (RAG LLM-Lasso, plain LLM-Lasso, and baselines) on the test set. For each number of selected features, we record the best misclassification rate, repeating the process

across 10 random splits and plotting the mean. The same procedure is applied using test metric AUROC. Results from large-scale experiments are shown in Figure 6, with a close-up comparison of RAG-enhanced LLM-Lasso vs. Lasso in Figure 8. RAG LLM-Lasso outperforms both the baselines and plain LLM-Lasso, achieving lower misclassification rates and higher AUROC with fewer selected genes.

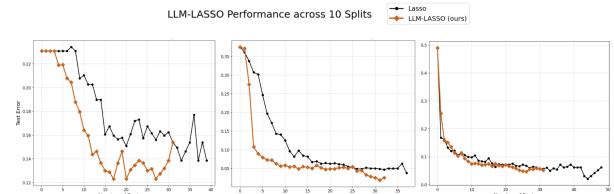


Figure 8. Close-up comparison of Lasso vs. RAG-enhanced LLM-Lasso on various Lymphoma datasets.

**Feature Contribution** In the experiments we run multiple LLM-Lasso regressions, and thus we are unable to extract a single list of selected features and their coefficients. For better interpretability, we introduce a feature contribution metric that takes the proportion that each feature appears across the full path of the number of features. A feature contribution of 1 means the feature appeared in all the models, while that of 0 means the feature appeared in none of the models. We create heatmaps of the union of genes with top 10 feature contributions for the Lasso, Plain LLM-Lasso, and RAG LLM-Lasso, as well as the polarity of the coefficients, represented as letters in the heatmaps (“F” coefficients in the direction of FL and “D” for DLBCL) (Figure 14). In the clinically relevant problem of classifying FL and DLBCL, there are several genes with high feature contributions that have relevance in cancer genomics and hematology/oncology, especially in the GPT-4o LLM-Lasso heatmap. For example, *AICDA*, *BCL2*, and *BCL6*, all of which have high feature contributions in the RAG LLM-Lasso, have been implicated in the transformation of FL to DLBCL (Lossos et al., 2004; Green et al., 2013). In the o1 LLM-Lasso heatmap, although *AICDA* is included as the top gene in the RAG LLM-Lasso, many of the other genes are less relevant to the DLBCL literature (Pasqualucci & Dalla-Favera, 2018).

## 6. Discussion and Conclusion

The LLM-Lasso is a simple, tunable model that incorporates domain-specific knowledge from LLMs and outperforms state-of-the-art feature selection models. The LLM-Lasso achieves superior performance with a small number of features, not only improving predictive performance compared to baseline models but also providing clues to identify important features. The LLM-Lasso protects against potential

inaccuracies or hallucinations from the LLM by tuning hyperparameters, as with the case with the FL experiment. Furthermore, the cross-validated power of the hyperparameter  $\mathcal{I}^{-\eta}$  allows us to tune the extent to which the LLM-Lasso relies on the penalty factors provided by the LLM.

In the experiments conducted in this particular paper, the identified genes had biomedical significance, suggesting that the LLM-Lasso could provide important clues that can lead to novel discoveries or confirm known feature-target relationships in biomedicine or in any field. More empirical investigations are needed to examine how well the LLM-Lasso prioritizes features.

The implementation of the full LLM-Lasso pipeline is made available at <https://github.com/pilancilab/LLM-Lasso>.

## Author Contributions

**Erica Zhang** proposed and implemented the complete RAG pipeline for LLM-Lasso, including scraping and curating the OMIM-based knowledge base for retrieval (Sections 3.2, 5.4.1). She developed the entire Python code base spanning all surveyed LLMs and baseline methods (Sections 5.1-5.3). She conducted the small-scale experiments and performed various model ablation studies with Naomi (Figures 6-8). Additionally, she led the Python-based large-scale experiments and was largely responsible for LLM or task-specific data-driven learning related sections in the manuscript (Sections 1-2, Appendices B, C.1-C.2, D.1-D.2).

**Ryunosuke Goto** led efforts to develop the statistical model of the pipeline, including the evaluation of various penalty factor forms through simulations and development of the cross-validation pipeline for determining the hyperparameter for each penalty factor form (Sections 3.1, 4.1, 5.4, Appendix A). He proposed the use of OMIM and contributed to prompt engineering. He led large-scale experiment efforts, including preparing the dataset, applying the LLM-Lasso in R, and visualizing and interpreting the results with Jurik, Nick, and Ash (Section 6, Appendix C.2.2, C.3.1, D.3).

**Naomi Sagan** implemented the regularized Lasso and the cross-validation pipeline in Python (Appendix D.4), performed the adversarial data corruption simulations (Section 4.2), and contributed to the LLM-Lasso evaluation process (Section 5.4.2). She and Erica were responsible for boosting LLM performance, including prompt engineering and improvements/customizations to the RAG pipeline (Appendix E), and they finalized Python-based experiments in Section 5.4 (Figures 6-8). She helped contribute to the experimental portion of the manuscript.

**Jurik Mutter** provided the dataset for the large-scale experiments in Section 5.4 and assisted in interpreting the biological implications of the data derived from the LLM-Lasso

analysis. Participated in occasional meetings and offered constructive feedback throughout the research process.

**Nick Phillips** assisted in providing validation dataset for the large-scale experiments in Section 5.4 and assisted in interpreting the biological implications of the data derived from the LLM-lasso analysis. He participated in meetings and offered feedback throughout the research process.

**Ash Alizadeh** helped provide the original motivation for the noninvasive cancer classification problem from blood samples. He conceived and proposed the use of and calculation of inferred expression of selected genes of interest that he curated toward this goal using EPIC-Seq using plasma cell-free DNA. His lab provided the corresponding data from multi-cancer classification cohort. He participated in regular meetings, he guided fellows and students working on the project, and he provided comments on the form and details of the manuscript, including biological interpretations.

**Kangwook Lee** participated in the initial idea conception, assisted the students in designing controlled experiments, and provided comments on the manuscript.

**Jose Blanchet** participated in weekly meetings, and helped in guiding students working on the project. In collaboration with others authors helped write the Introduction and assisted in writing and editorial work of the manuscript. Proposed adversarial prompt.

**Mert Pilanci** participated in the conception of the original idea, and helped develop the regularization and prompting strategy. He attended weekly meetings, provided guidance to the students on the project and offered feedback on the manuscript.

**Robert Tibshirani** helped to conceive the original idea and helped develop the statistical model for the pipeline. He participated in weekly meetings, and guided students working on the project. He provided comments on the form and details of the manuscript.

**Acknowledgement.** R.T. was supported by the National Institutes of Health (NIH) grant 5R01EB001988-16 and the National Science Foundation (NSF) grant 19DMS1208164. M.P. was supported in part by the NSF CAREER Award under Grant CCF-2236829; in part by the U.S. Army Research Office Early Career Award under Grant W911NF-21-1-0242; in part by the Office of Naval Research under Grant N00014-24-1-2164

## Impact Statement

This paper contributes to advancing machine learning and statistics by improving the robustness of LLM-based feature selection, particularly in the biomedical domain. By reducing susceptibility to overfitting, our approach enhances the

reliability and generalizability of feature selection methods, improving the trustworthiness and interpretability of AI. These improvements have positive implications for prioritizing features in biomedicine, potentially leading to novel biomedical discoveries.

## References

Alibaba DAMO Academy. Qwen models, 2025. URL <https://damo.alibaba.com/qwen>. Foundation models developed by Alibaba DAMO Academy.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020. URL <https://arxiv.org/abs/2005.14165>.

Bühlmann, P. and Van De Geer, S. Statistics for high-dimensional data: Methods, theory and applications. *Springer*, 2011.

Chandrashekhar, G. and Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Chen, J. and Zou, X. Prediction tasks using embeddings derived from domain-specific literature with large language models. *Advances in Data Science and Analytics*, 8(2):200–215, 2024.

Choi, K., Cundy, C., Srivastava, S., and Ermon, S. Lmpriors: Pre-trained language models as task-specific priors, 2022. URL <https://arxiv.org/abs/2210.12530>.

Cui, H., Wang, C., Maan, H., and Wang, B. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 2024.

DeepSeek AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

Ding, C. and Peng, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(02):185–205, 2005.

Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., yong Sohn, J., Papailiopoulos, D., and Lee, K. LIFT: Language-interfaced fine-tuning for non-language machine learning tasks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=s\\_PJMEGIUfa](https://openreview.net/forum?id=s_PJMEGIUfa).

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvassy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library, 2024. Available at <https://faiss.ai>.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley & Sons, New York, 2 edition, 2001.

Friedman, J., Tibshirani, R., and Hastie, T. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, December 2020. URL <https://arxiv.org/abs/2101.00027>.

Green, M. R., Gentles, A. J., Nair, R. V., Irish, J. M., Kihira, S., Liu, C. L., Kela, I., Hopmans, E. S., Myklebust, J. H., Ji, H., et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood, The Journal of the American Society of Hematology*, 121(9):1604–1611, 2013.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. *Feature Extraction: Foundations and Applications*. Springer, 2007.

Han, S., Yoon, J., Arik, S. O., and Pfister, T. Large language models can automatically engineer features for few-shot tabular learning, 2024. URL <https://arxiv.org/abs/2404.09491>.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, November 2023. doi: 10.48550/arXiv.2311.05232. URL <https://doi.org/10.48550/arXiv.2311.05232>.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T.

T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2024. ISSN 1558-2868. doi: 10.1145/3703155. URL <http://dx.doi.org/10.1145/3703155>.

Jeong, D. P., Lipton, Z. C., and Ravikumar, P. Llm-select: Feature selection with large language models, 2024. URL <https://arxiv.org/abs/2407.02694>.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

Khattab, O. and Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *arXiv preprint*, arXiv:2004.12832, April 2020. doi: 10.48550/arXiv.2004.12832. URL <https://doi.org/10.48550/arXiv.2004.12832>.

Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997.

LangChain Community. Langchain community resources, 2024. URL <https://docs.langchain.com>.

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Lewis, D. D. Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttrler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

Lewkowycz, A., Andreassen, A. J., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V. V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022.

Li, D., Tan, Z., and Liu, H. Exploring large language models for feature selection: A data-centric perspective, 2024. URL <https://arxiv.org/abs/2408.12025>.

Li, J. and Liu, H. Feature selection: An ever-evolving frontier in statistical learning. *IEEE Transactions on Neural Networks and Learning Systems*, 26(1):1–14, 2015.

Li, X. and Ren, J. Micq-ipso: An effective two-stage hybrid feature selection algorithm for high-dimensional data. *Neurocomputing*, 501:328–342, Aug 2022. doi: 10.1016/j.neucom.2022.05.048. URL <https://doi.org/10.1016/j.neucom.2022.05.048>.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 2023.

Liu, S., Lv, F., Liu, X., et al. Ice-search: A language model-driven feature selection approach. *arXiv preprint*, arXiv:2402.18609, 2024. URL <https://arxiv.org/abs/2402.18609>.

Lossos, I., Levy, R., and Alizadeh, A. Aid is expressed in germinal center b-cell-like and activated b-cell-like diffuse large-cell lymphomas and is not correlated with intraclonal heterogeneity. *Leukemia*, 18(11):1775–1779, 2004.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2018. doi: 10.1109/TPAMI.2018.2889473.

Manikandan, H., Jiang, Y., and Kolter, J. Z. Language models are weak learners. *arXiv preprint arXiv:2306.14101*, 2023.

Meta AI. Llama 3 - 8b instruct model. <https://huggingface.co/meta-llama/llama-3-8b-instruct>, 2025a. Accessed: 2025-01-16.

Meta AI. Llama 405b, 2025b. URL <https://ai.meta.com/llama>. Large-scale language model with 405 billion parameters.

OpenAI. Gpt-3.5 technical report, 2023a. URL <https://openai.com/research/gpt-3-5>. Accessed: YYYY-MM-DD.

OpenAI. Gpt-4 technical report, 2023b. URL <https://openai.com/research/gpt-4>.

OpenAI. Openai embeddings, 2024. URL <https://platform.openai.com/docs/guides/embeddings>.

OpenAI. Openai o1 system card, 2024. URL <https://cdn.openai.com/o1-system-card.pdf>. Large language model.

Pasqualucci, L. and Dalla-Favera, R. Genetics of diffuse large b-cell lymphoma. *Blood, The Journal of the American Society of Hematology*, 131(21):2307–2319, 2018.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., and Tang, S. Graph retrieval-augmented generation: A survey, 2024. URL <https://arxiv.org/abs/2408.08921>.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473. Association for Computational Linguistics, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL <https://www.R-project.org/>.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. OpenAI Blog, 2019.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

Shuster, K., Humeau, S., Komeili, M., and Weston, J. Improving fact-checking with retrieval-augmented generation. In *Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Singh, N. and Singh, P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemometrics and Intelligent Laboratory Systems*, 217:104396, 2021. ISSN 0169-7439. doi: <https://doi.org/10.1016/j.chemolab.2021.104396>. URL <https://www.sciencedirect.com/science/article/pii/S0169743921001647>.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., and Nanayakkara, S. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023. doi: 10.1162/tacl\_a\_00530. URL <https://aclanthology.org/2023.tacl-1.1/>.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012.

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature medicine*, 13(3):361–366, 2007.

Suzgun, M., Scales, N., Schärli, N., Gehrman, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics*, 2023.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618 (7965):616–624, 2023.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Ranzato, M., Roux, A. S. A., Koura, P. S., Gong, K., Rozière, B., Belgrave, D., Hoseiny, M. E., Sakhaei, P., Babaeizadeh, M., Bakas, S., de Las Casas, D., Xu, T., Larcher, R., Lacroix, T., Lample, G., and Conneau, A. LLaMA 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL <https://arxiv.org/abs/2307.09288>. Accessed: YYYY-MM-DD.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N., and Xue, C. J. Retrieval-augmented generation for natural language processing: A survey, 2024. URL <https://arxiv.org/abs/2407.13193>.

Yang, J. and Hastie, T. A fast and scalable pathwise-solver for group lasso and elastic net penalized regression via block-coordinate descent, 2024. URL <https://arxiv.org/abs/2405.08631>.

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., Liu, Y.-Y., and Yuan, L. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2024. URL <https://arxiv.org/abs/2310.01469>.

Zhang, Y., Khalifa, M., Logeswaran, L., Lee, M., Lee, H., and Wang, L. Merging generated and retrieved knowledge for open-domain qa. In Bouamor, H., Pino, J.,

and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4710–4728, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.286. URL <https://aclanthology.org/2023.emnlp-main.286>.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

## A. Penalty Factor Simulations Details

We run simulations to find the adequate form of penalty factors. Based on the simulations, we use the inverse importance penalty factors to compare the LLM-Lasso to the baseline models. The details can be found in A. using datasets outlined in Table 1. The data are split into the importance score-generating set and the cross-validation set. The hypothetical importance scores are generated by running a Lasso regression on the score-generating set and assigning the absolute values of the coefficient of each feature is assigned to be the score of that feature. Then, the hypothetical importance scores are scaled so that the maximum score was 1 and the minimum score is 0.1. On the cross-validation set, we run the hypothetical LLM-Lasso using different forms of penalty factors: (i). the inverse of the importance scores and their powers and (ii). ReLU penalty factors with different thresholds. For ReLU penalty factors, we set the maximum penalty factor such that the least important feature received a coefficient of 0 for all values of  $\lambda$ . We perform 5-fold cross-validation across the hyperparameter  $\gamma \in (0.1, 0.2, \dots, 0.9)$  for the ReLU penalty factors and  $\eta \in (0, 1, \dots, 10)$  for penalty factors of the form  $\mathcal{I}^{-\eta}$ . We obtain cross-validation misclassification rates across the spectrum of regularization parameters  $\lambda$  with a  $\lambda_{\min}$  to  $\lambda_{\max}$  ratio of 0.01. Cross-validation is performed such that the difference in the area of the plot of the misclassification rate across numbers of features with respect to the Lasso is maximized. The best misclassification rate is obtained for each number of features selected. We perform the above for 10 data splits and plot the mean of the best misclassification rate for each number of features. The above procedure is repeated for areas under the receiver-operating characteristic curve (AUROCs) as the cross-validation metric.

Dataset	n	p
Cancer microarray (Ramaswamy et al., 2001)	52	1000
Small-round-blue-cell tumor (Khan et al., 2001)	83	1000
Lung cancer (Spira et al., 2007)	187	1000

Table 1. Summary of simulation datasets for penalty factor form.

The datasets we use in the simulations are summarized in Table 1. In the simulations, we perform the task of classifying samples into tumor tissue or healthy tissue (lung cancer dataset) or cancer subtypes (cancer microarray dataset, rhabdomyosarcoma vs others; small-round-blue-cell tumor (SR-BCT) dataset, lymphoma vs leukemia) using gene expression levels. We select features with the top 1000 variances as predictors. The simulation show an advantage of the inverse importance penalty factors over the ReLU penalty factors, as well as compared to the Lasso (Figures 9 and 10). Thus, in the experiments, we use the inverse importance penalty factors to compare the LLM-Lasso to baseline models.

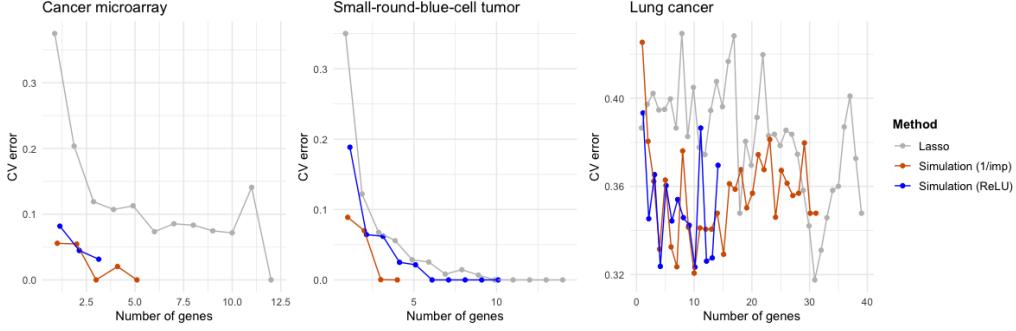


Figure 9. Test error in simulations

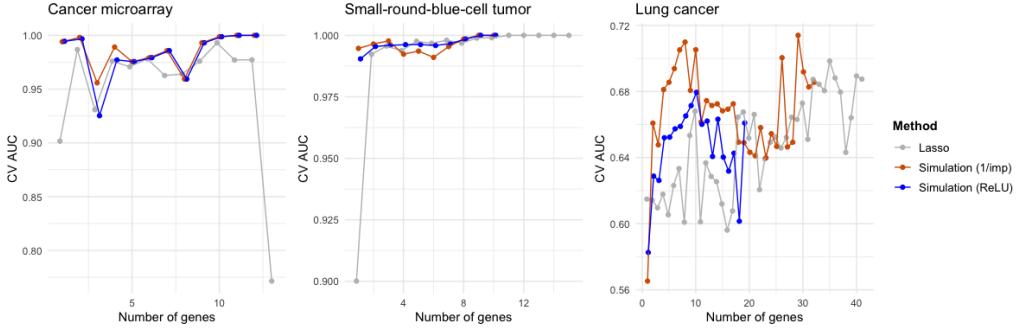


Figure 10. AUROC in simulations

## B. Prompt Construction

Not only is prompting shown to be significant to the performance of LLMs Throughout our experiment with the biomedical dataset, we set the system message to the generation LLM as “assistant,” with instruction: “you are an expert assistant with access to gene and cancer knowledge.”

As we recall in Section 3.2.1 that our full prompt follows the following structure:

$$\mathcal{P}^{\text{full}} = \text{prompt}(\mathcal{Q}^{\text{user}}(\mathcal{A}(\phi, c)), \mathcal{C}^{\text{retriever}}(k, \mathcal{R}(\phi, c)), \mathcal{H}^{\text{system}}).$$

The design choice for the user therefore primarily resides in (i). the construction of the task description prompt  $\mathcal{A}(\phi, c)$  and (ii). the construction of a customized retrieval prompt  $\mathcal{R}(\phi, c)$  in the case when RAG is used.

### B.0.1. TASK DESCRIPTION

The general format of text description follows Figure 2. However, there are many ways one can format each of the three sections, that is, background description, a task description, and formatting rules. In the following, we go through each component in depth.

**Background Description.** We include the following key elements in our background description prompt:

- Meta-data of the dataset. This includes details on how the data is collected, number of samples, and number of features.

- User Intention. This includes a description of our goal for data analysis. For conducting classification experiments using LLM-Lasso, for instance, we remark: “*We wish to build a statistical (Lasso) model that classifies samples into category diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL).*”

**Task Description.** The task section specifies the exact request made to the generation LLM. For LLM-Lasso, this involves a description of the penalty factors. As penalty factors can be less intuitive to understand than the straightforward importance scores, through our experiments, we experimented with a number of prompts to describe to the LLM the meaning of “penalty factors” in an effort to boost prediction performance by facilitating understanding. To this end, we employ four prompting strategies—bayesian, ReLU, adversarial, and  $\ell_1$ -generated—to guide the interpretation of the penalty factors. We found that this part of the prompt has a direct and considerable impact on the predictive ability of the LLM. We discuss each of the three component in detail in the following.

### I. The Bayesian Approach for Prompt Construction.

The Lasso with penalty factors can be interpreted from a Bayesian perspective, where the penalty factor serves as the scaling parameter of a Laplace prior. A larger penalty factor results in a tighter distribution around zero, encouraging sparsity. Under this framework, the corresponding prompt for the oncology prediction task is:

I would like you to provide penalty factors greater than or equal to 0 to use on each coefficient of a Lasso estimator based on domain knowledge for a regression or classification task. Suppose  $\beta_k$  is the regression coefficient for feature  $k$ . We interpret Lasso with penalty factors  $\lambda_k$  as yielding a maximum a posteriori estimate under Laplace priors with parameters  $\lambda_k$ . This means that, before observing the data, the ratio of log-tail probabilities  $\log P(\|\beta_i\| > t) / \log P(\|\beta_j\| > t)$  is equal to  $\lambda_i / \lambda_j$  for each  $i, j$  and for all  $t$ . Therefore, the penalty factors represent relative log-tail probabilities of coefficients. For example, if feature  $A$  has a penalty factor of  $\lambda$  and feature  $B$  has a penalty factor of  $2\lambda$ , this implies that the log-likelihood of the absolute value of the regression coefficient for  $A$  exceeding any threshold is twice that of  $B$ . Thus, the larger the penalty factor for a coefficient, the less “important” the coefficient is.

### II. The ReLU-form Approach for Prompt Construction.

Another prompting framework for interpreting the penalty factor is to directly describe the process which we code our underlying Lasso model with penalty using the ReLU-form penalty.

We plan to use your scores with a Lasso-regularized multinomial classifier, implemented via the R package `glmnet`. The scores will generate penalty factors (weights on the  $\ell_1$  norm), which will be used in `glmnet`. Higher importance genes will be assigned smaller penalty factors, while lower importance genes will receive larger penalty factors.

Let  $\mathbf{x}_{\text{all}}$  denote the feature matrix (number of observations by number of genes) and  $\mathbf{y}_{\text{all}}$  the multinomial class outcome. Similarly, let  $\mathbf{x}_{\text{test}}$  and  $\mathbf{y}_{\text{test}}$  be the test set feature matrix and class outcome, respectively. Let  $\mathbf{scores}$  be the  $p$ -vector of gene importance scores provided by ChatGPT.

The details of our plan are implemented in the following R code: [omitted]

### III. The Adversarial Approach for Prompt Construction.

The penalty factor can also be interpreted as part of an adversarial game to enhance out-of-sample prediction robustness. Here, the penalty factor scales the cost of perturbing covariates under a weighted  $\ell_\infty$  norm. Larger penalty factors make changes to a covariate more “expensive,” limiting adversar-

ial alterations, while smaller factors make them cheaper, reflecting lower importance. The adversary operates within a fixed budget, distributing total weights across covariates to balance importance and vulnerability.

You are tasked with helping perform a what-if (adversarial) analysis to improve out-of-sample prediction on a logistic regression model for classification. Here is how this analysis works: (1). For each sample, every covariate (gene expression level) can be modified (increased or decreased), but the cost of changing each covariate is scaled by a weight that we assign now. (2). The “size” of a change to a single sample is measured by the weighted  $L_\infty$  norm: if  $\delta_i$  is the change to covariate  $i$ , and  $w_i$  is the weight for covariate  $i$ , then the size of the change is:  $\max_i(|\delta_i| \times w_i)$ . Across the dataset, the average of these sizes is constrained by a fixed budget. (3). A larger weight on covariate  $i$  makes changes to that covariate more “expensive” to the adversary, limiting how drastically it can be altered under the same overall budget. A smaller weight makes it cheaper to perturb that feature, which might be acceptable if the gene is less important. (4). You must distribute a total of 100 weight units among all covariates:  $\sum_i w_i = 100$ . Given this setup, your job is to choose weights for each predictor. Your goal is to provide a plausible weighting scheme that balances the importance of each predictor against potential adversarial changes.

### IV. o1-Generated Approach for Prompt Construction.

In addition to using different theoretical angles to explain to the generation LLM the notion of penalty factors, we consult advanced LLMs, such as o1 from OpenAI for their advice on constructing a prompt to explain the penalty factors in a way that would be most conducive for an LLM to perform:

Provide penalty factors for each of the genes. These penalty factors should be integers between 2 and 5 (inclusive), where: 2 indicates a gene strongly associated with “category” (i.e., it should be penalized the least by Lasso). 5 indicates a gene with minimal relevance to “category” (i.e., it should be penalized the most by Lasso).

We note that the range (2, 5) is arbitrary and can be chosen by the user. It is noticeably that the o1-generated prompt is significantly shorter and simpler than the other three approaches and focuses on direct instructions (i.e. more important features should be penalized more) using use-case examples rather than attempting to explain the intuition behind the penalty factors. While the o1-generated prompt seems to introduce no in-depth understanding of penalty factors at the shallow level, empirically, we find that this prompt consistently encourages better prediction perfor-

mance across a range of LLMs from simple to advanced on penalty factor production for feature selection. Overall, our empirical findings suggest that: o1-generated prompt > Bayesian prompt > ReLU prompt > Adversarial prompt, where we use a descending order of performance.

**Output Format Instructions.** Our experiments revealed that selecting appropriate output format instructions is crucial not only for the accuracy of the score collection process but also for maintaining the quality of the scores produced. This is especially important for smaller models with fewer parameters (e.g., `llama-3-8b-instruct`), which often struggle to follow prompt instructions and understand the concepts and guidance provided.

In practice, we found that directly using text responses and providing LLMs with clear text formatting rules is more effective in regulating their behavior and ensuring a smooth score collection process compared to requesting responses in raw JSON format, as commonly used in LangChain’s pipeline. For all LLMs, we attach a format instruction to the end of every prompt, with slight modifications tailored to the specific task. Below is an example of a format instruction used for the task of outputting penalty factors for gene selection in cancer or lymphoma prediction.

#### Formatting Rules:

1. **Score Representation:** Use a direct floating-point number (e.g., 0.5). Avoid scientific notation (e.g.,  $10^{**}(-2)$  or  $1e-2$ ) and additional formatting.
2. **Include All Genes:** Assign a penalty factor for every gene in the input list, preserving the order of input.
3. **Reasoning:** After each penalty factor, add a concise reasoning about the gene’s role in predicting {category}.
4. **Consistency:** Ensure uniform formatting. Example:

AASS: 1

Reasoning: This gene is highly expressed in cancer pathways and has been associated with {category}. Assigned a low penalty factor.

BRCA1: 5

Reasoning: BRCA1 is not significantly relevant for {category}. Assigned a high penalty factor.

Do not include disclaimers about lacking full data; rely on general cancer genomics and pathway relevance.

As outlined in the formatting prompt, three strategies were found to be particularly effective:

1. **Highlighting common errors:** We include a list of frequent formatting mistakes made by LLMs, identified through trial and error. These include, for example, using scientific notation instead of floating-point numbers, which complicates the score collection algorithm, and applying inconsistent additional formatting to the scores.
2. **Providing examples:** Examples demonstrating the desired score and explanation format significantly improve the LLMs’ understanding of the task. This is particularly important when querying for penalty factors instead of importance scores. While a dedicated prompt explains the concept of penalty factors, smaller models like `llama-3-8b-instruct` often struggle with the counterintuitive nature of penalty scores—where lower values indicate higher significance and vice versa. Including examples of both low and high penalty scores helps address this challenge and ensures better compliance.
3. **Using a firm tone:** We employ strict language to enforce adherence to the rules. Commands such as “Do not say that it’s not possible to compute precise penalty factors without access to the actual gene expression values” and “Responses not following these guidelines will be considered invalid” have proven effective in ensuring LLMs behave consistently and follow the guidance provided.

Figure 11 is an example of the full user prompt used in the study of classifying patients into DLBCL and FL, which employs o1-generated explanation of penalty factors.

#### B.0.2. RETRIEVAL PROMPT

The default pipeline in Langchain for retrieval query is to directly perform semantic similarity search on the user’s original prompt to the generation LLM. This becomes problematic, however, when the main user prompt is large and overshadows the important information that sheds light on what documents should be retrieved. As an example, when passing in directly the full user prompt for retrieval in the high-dimensional oncology classification tasks, semantic similarity search sometimes retrieves information on the description of the dataset, for example, contexts regarding cfDNA fragmentation pattern and EPIC-Seq, instead of what we are actually curious about, that is, the relevance of certain gene, say AASS, with classifying lymphoma subtypes, say, diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL).

In order to pinpoint the retriever to the specific retrieval

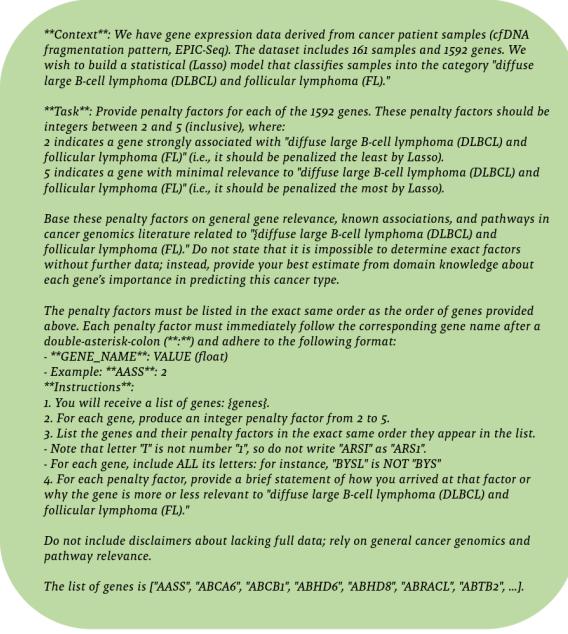


Figure 11. Example of a full user prompt for experiment study DLBCL vs FL.

documents that are actually relevant to the task at hand, we use a customized retrieval prompt. For the lymphoma classification tasks using micro-array gene data, due to the high-dimensional nature of the dataset, we batch process the genes (see Appendix E for discussion) and use the following prompt that takes each gene  $g_i \in \{g_1, \dots, g_B\}$  in each batch of size  $B$  and the target classification category  $c$ :

Retrieve information about gene  $\{g\}$ , category  $\{c\}$ , especially in the context of  $\{g\}$ 's relevance to  $\{c\}$ .

An example prompt using this format is show in Figure 12.

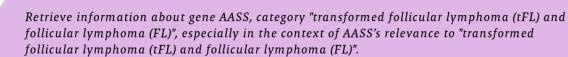


Figure 12. Example of a completed retrieval prompt.

For each pass of retrieval search with gene and lymphoma pair, we retrieve top  $k$  relevant documents. After collecting the contexts for all the genes in the batch, we then filter for the unique documents and then append them to the full prompt in prompt component  $C$ . It turns out that the specific

implementation of this procedure is an art: we want to strike a balance between overwhelming the generation LLM with long-context and potentially minimally informative documents and excessive cautious retrieval that does not inform the LLM by much.

## C. Experiment Details

### C.1. Model Details

We provide more details of the LLMs sampled. Table 2 summarizes the cut-off dates in each LLMs in used.

### C.2. Dataset Details

In this section, we give more details on the datasets used in the Experiment Section.

#### C.2.1. SMALL-SCALE EXPERIMENT DATASETS

We source a wide range of small-scale datasets for feature selection in classification and regression. We use \* to indicate the datasets that are released after the cutoff dates for all models sampled (see Table 2 for an overview of the model cutoff dates). For all small-scale datasets, we remove features whose values are not numerical and not categorical and remove rows and columns with missing values. We remark that the purpose of the small-scale experiment is not meant to demonstrate performance on the specific task but rather to show case the ability of the feature selector candidate, even in the absence of some potentially informative features and data.

Figure 13 describes the model ablation study on the spotify regression dataset for feature selection. As we can see, the GPT-4o model we used in Figure 7 is not the top performing model, yet we still outperformed the sampled feature selection baselines. This demonstrates once again the strength of our method. In addition, as we have seen in Figure 6, Deepseek r1 continues to dominate in performance against other models and smaller models such as LLaMa-3-8b-instruct and GPT-3.5 has worse performance comparing to their more advanced model counterparts.

#### C.2.2. LARGE-SCALE EXPERIMENT DATASETS

The datasets used in the large-scale experiments are outlined in Table 3, where  $n, p$  denotes resp. sample size and number of features. We note that the dataset has not been published and is currently confidential. Therefore, we do not disclose further details.

Model Name	Company	Cut-off Date	Source
GPT-3.5 (Turbo)	OpenAI	2021.09	Source
GPT-4o (2024-08-06)	OpenAI	2023.10	Source
o1	OpenAI	2023.10	Source
Llama-3-8B	Meta	2023.12	Source
Llama-3.1-405B	Meta	2023.12	Source
DeepSeek-R1	DeepSeek	2024.07	Source
Qwen Models-72B	Alibaba	2023.09	Source

Table 2. Surveyed LLMs Cutoff Dates Overview

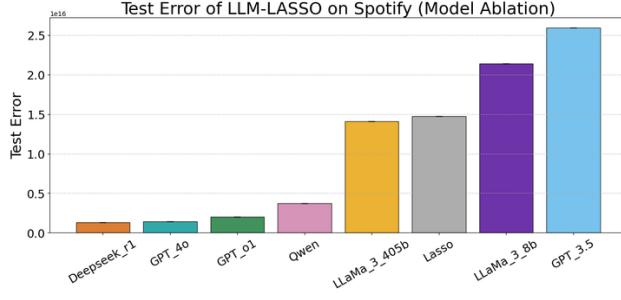


Figure 13. Model ablation on the Spotify dataset with test error rate computed at 25% of total features.

Dataset	n	p
Lymphoma (FL vs DLBCL)	130	1592
Lymphoma (MCL vs DLBCL)	161	1592
Lymphoma (cHL vs DLBCL)	196	1592

Table 3. Summary of large-scale experiment datasets.

### C.3. Supplemental Experiment Results

#### C.3.1. FEATURE CONTRIBUTIONS

Figure 14 shows heatmaps of feature contributions in the FL vs DLBCL experiment for GPT-4o and o1 RAG LLM-Lasso. Although the o1 model performs very well in terms of misclassification error and AUROC (Figure 6), the genes selected by GPT-4o may align more with the oncology literature, as per the discussion in the Feature Contribution paragraph of Section 5.4.2.

Dataset	Year	n	p	Source
Spotify*	2024	4600	29	Source
Wine	2009	6497	11	Source
Diabetes	1998	768	8	Source
Bank	2012	45211	51	Source
Glioma	2022	839	23	Source

Table 4. Summary of small-scale experiment datasets.

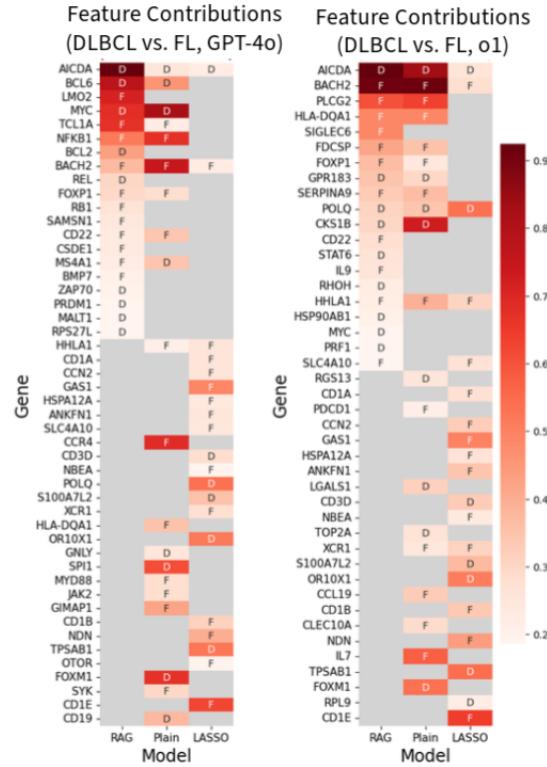


Figure 14. Heatmaps of feature contributions in the FL vs DLBCL experiment for GPT-4o and o1 LLM-Lasso, with and without RAG.

#### C.3.2. DEFERRED PLOTS

In this subsection, we present the deferred plots in the large-scale experiment section. Figure 15 illustrates the AUROC performance of our model against various baseline across the three high-dimensional lymphomal datasets. It is evident that the strong performance demonstrated by Figure 6 carries over to the AUROC metric.

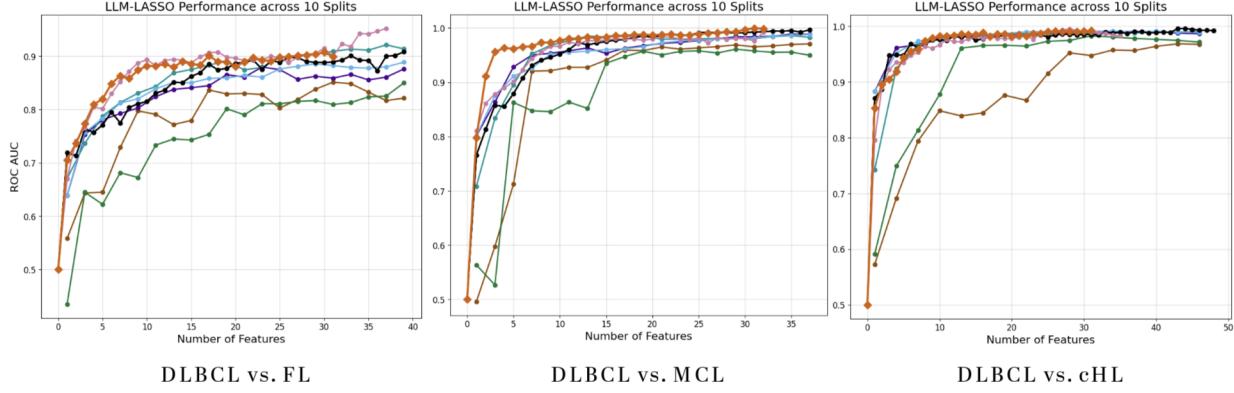


Figure 15. AUROC performance across 10 splits for the lymphoma datasets.

## D. Implementation Details

### D.1. Handling Token Limits in LLMs

Input and output token limits in closed-source pretrained LLMs pose significant challenges, especially when querying large sets of predictors, as they restrict the information processed or returned in a single interaction. The absence of memory retention further complicates output aggregation. This limitation affects both closed-source GPT models via the OpenAI API and cloud-hosted open-source models, which also lack persistent memory.

However, ensuring that an LLM has sufficient output tokens is critical for its performance. For instance, (Wei et al., 2022) showed that step-by-step reasoning improves effectiveness, and we observe that LLMs struggle under tight token limits or when limits are exceeded (see Appendix D). To handle large feature sizes, batch-querying with an appropriate batch size is necessary to stay within token limits. However, this approach introduces challenges: without memory retention, the LLM cannot access previously processed features or their scores, leading to inconsistencies when aggregating batch results and potential scale mismatches. To address these issues and ensure accurate feature diagnoses while preserving essential output tokens, we propose two strategies that require no fine-tuning or parameter modifications.

**Text-based Summary.** A straightforward approach to address this challenge is to batch-query the features while enabling memory retention in the LLM by augmenting the user query  $Q_{\text{user}}$  with a summarization of chat history  $\mathcal{H}^{\text{system}}$ , stored in a conversation buffer constrained by the max token limit. Several open-source Python packages support this functionality. In our implementation, we use LangChain’s `ConversationBufferMemory`. While not ideal for score-collection scenarios—since summarization often

omits full scores and context due to token constraints—we find that including memory increases the likelihood of the LLM assigning scores on a consistent scale and provides marginal improvements in prediction performance (See Appendix D for more details).

**Statistical Estimation.** Another approach to address this problem is to use statistical techniques to infer the true score from batch scores without injecting memory into each batch. We introduce the following method. We note that to balance batch size and the number of queries, we heuristically select a batch size of  $\lceil \sqrt{p} \rceil$ , where  $p$  is the total number of features and  $\lceil \cdot \rceil$  is the ceiling notation.

**Scaling.** Given batch scores  $B_1(s), \dots, B_{\lceil \sqrt{p} \rceil}(s)$ , the scaling method involves selecting the maximum score from each batch,  $s_{\max,1}, \dots, s_{\max,\lceil \sqrt{p} \rceil}$ , and passing these maximum scores as a new batch to the LLM for rescore, yielding  $\tilde{s}_{\max,1}, \dots, \tilde{s}_{\max,\lceil \sqrt{p} \rceil}$ . The final score is then computed by weighting and concatenating the batch scores. Specifically, each batch  $B_i$  is weighted by  $\frac{\tilde{s}_{\max,i}}{\sum_{j=1}^{\lceil \sqrt{p} \rceil} \tilde{s}_{\max,j}}$ , which is the normalized rescored maximum candidate from that batch relative to the rescored maximum candidates across all batches.

### D.2. Score Collection

LLMs, especially smaller models can make formatting mistakes. For instance, they may not include all necessary genes, or may include extra genes (e.g., ones mentioned in retrieval context) as part of the genes to score. For OpenAI models, we use structured outputs to directly receive the scores as a Python object. In models where this streamlined score collection feature is not available, we rely on the output formatting from the prompt (see Appendix B) and search for floating point scores that immediately follow the double-asterisk-colon sign (with or without space). If we

fail to collect scores, we retry until the correct scores for the batch are collected.

### D.3. R implementation

Once the importance scores are obtained from the LLM, the LLM-Lasso can be implemented in R (R Core Team, 2024). One can pass the penalty factors, transformed into the form of choice, such as the ReLU-form or inverse importance and their powers ( $\mathcal{I}^{-n}$ ), into the `cv.glmnet` function in the package `glmnet` (Friedman et al., 2010). The penalty factors can be passed into an argument called `penalty.factor`, which specifies the penalty factors to be assigned to each feature.

### D.4. Python Implementation

The full end-to-end pipeline of LLM-Lasso is implemented in Python. The score collection is done via OpenAI APIs for GPT models and o1, and via OpenRouter otherwise. Langchain is used for the retrieval component of RAG.

For computing the data-driven baseline metrics (such as mutual information and MRRM), we first produce a set of randomly generated 50/50 train and test splits and save them to CSV files. These splits are used for both LLM-Lasso and each data-driven baseline. Then, baseline scores can be computed via our Python implementations, relying on `scikit-learn`.

The implementation of the LLM-Lasso model, given the importance scores, is based on the package `adelie` (Yang & Hastie, 2024). We have a custom fork of `adelie` that adds AUROC and misclassification error metrics to the output of `adelie.cv.cv_grnnet`. For penalty factors, we consider powers of the inverse importance ( $\mathcal{I}^{-n}$ , where  $n$  is some small integer). Cross-validation, with folds determined internally by `adelie`, determines which power of the inverse importance to use. Results are averaged across the same folds as used to compute the baselines.

For more details, refer to our open-sourced Git Repository.

## E. Experiment Supplements

### E.1. Evaluation of RAG

To illustrate the benefits of RAG, we query GPT-4o with the question “Can you concisely tell me about the gene BCL6?” twice: once without RAG, and once with the OMIM RAG pipeline described in Section 5.4.1. The corresponding outputs are displayed in Figure 16. Without RAG, the response is more generic, whereas the RAG-enhanced response is more detailed and scientific, specifically describing how BCL6 is related to lymphoma and citing interactions with proteins and other genes.

Prompt

Can you concisely tell me about the gene BCL6?

Without RAG

BCL6 (B-cell lymphoma 6) is a gene that encodes a transcriptional repressor protein involved in the regulation of immune responses. It plays a critical role in the development and function of B cells, particularly in the formation of germinal centers where B cells mature and undergo antibody diversification. BCL6 is crucial for processes like somatic hypermutation and class switch recombination. Abnormal expression or translocations involving BCL6 are associated with certain types of lymphomas, especially diffuse large B-cell lymphoma (DLBCL). As a transcriptional repressor, BCL6 regulates the expression of genes involved in cell cycle control, apoptosis, and differentiation

With RAG

The BCL6 gene encodes a 79-kD protein that functions as a zinc finger-transcription factor, primarily involved in B-cell differentiation. It is predominantly expressed in the B-cell lineage, specifically in mature B cells and germinal centers. BCL6 is implicated in the pathogenesis of certain B-cell lymphomas, particularly diffuse large B-cell lymphoma (DLBL), where its expression may be deregulated due to chromosomal translocations affecting the 3q27 region. These translocations can lead to fusion with other genes, such as the Ikaros gene, potentially contributing to lymphoma development by preventing normal B-cell differentiation.

Figure 16. Output of GPT-4o responses, with and without RAG.

We, however, find that there are cases where RAG fails to improve performance. There are several explanations for this. In some cases, the RAG knowledge base is not compatible with the downstream classification or regression task. This can cause few documents to be retrieved, in which case RAG can only marginally improve performance. For some tasks, e.g., for DLBCL vs. FL, irrelevant documents are retrieved. This harms performance by increasing the context that the LLM has to parse, while requiring it to sift through the context for relevant details. We also notice that some models can be overly reliant on the documents retrieved, assigning high penalty factors to relevant genes and citing that the genes did not appear in the provided context. This can be problematic in cases where the retriever fails to return all relevant documents, or the knowledge base lacks information about many genes.

### E.2. Scalability of Our Method

In this section, we discuss challenges and insights related to generating penalty factors for all 1592 genes in the Lymphoma dataset (Table 3). First, difficulties with long contexts (and context limits for some models) prevent us from generating all scores with a single query (see Section D.1 for discussion). Instead, we generate scores in batches of 40 genes, where  $40 \approx \sqrt{1592}$  balances batch size with number of queries. As a result, it is essential to ensure consistency in the penalty factors produced across batches (i.e., that a penalty factor of 5 corresponds to the same degree of relevance across batches). Without explicit handling of consistency, the LLM-produced penalty factors exhibit mean shifts between batches, as well as differences in orders of

magnitude.

To this extent, we constrain the penalty factors to be in a pre-determined range, which we encode in our prompt (see Figure 2). Empirically, the range 2 – 5 (inclusive) produces good results for the tasks in Table 3. If the range is too small, the penalty factors produced are close to those used in plain Lasso, preventing a large improvement of LLM-Lasso over Lasso. If the range is too large, some models (e.g., GPT-4o) provide penalty factors too close to the extremes. This can lead to convergence issues in the downstream algorithm, and increase the impact of spuriously low or high scores.