
Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning

Jaehyun Nam^{*1}, Kyuyoung Kim^{*1}, Seunghyuk Oh¹
Jihoon Tack¹, Jaehyung Kim², Jinwoo Shin¹

¹Korea Advanced Institute of Science and Technology, ²Carnegie Mellon University
{jaehyun.nam, jinwoos}@kaist.ac.kr

Abstract

Learning effective representations from raw data is crucial for the success of deep learning methods. However, in the tabular domain, practitioners often prefer augmenting raw column features over using learned representations, as conventional tree-based algorithms frequently outperform competing approaches. As a result, feature engineering methods that automatically generate candidate features have been widely used. While these approaches are often effective, there remains ambiguity in defining the space over which to search for candidate features. Moreover, they often rely solely on validation scores to select good features, neglecting valuable feedback from past experiments that could inform the planning of future experiments. To address the shortcomings, we propose a new tabular learning framework based on large language models (LLMs), coined *Optimizing Column feature generator with decision Tree reasoning (OCTree)*. Our key idea is to leverage LLMs' reasoning capabilities to find good feature generation rules without manually specifying the search space and provide language-based reasoning information highlighting past experiments as feedback for iterative rule improvements. Here, we choose a decision tree as reasoning as it can be interpreted in natural language, effectively conveying knowledge of past experiments (*i.e.*, the prediction models trained with the generated features) to the LLM. Our empirical results demonstrate that this simple framework consistently enhances the performance of various prediction models across diverse tabular benchmarks, outperforming competing automatic feature engineering methods.

1 Introduction

Learning useful representations from raw data is key to the success of deep learning algorithms, and their effectiveness has been demonstrated across multiple domains, *e.g.*, vision [1, 2, 3, 4] and language [5, 6]. However, in the tabular domain, deep learning approaches are often perceived as less effective [7, 8, 9, 10]. For instance, tree-based approaches utilizing raw column features of tabular data [11, 12] often outperform deep learning models in tabular prediction tasks such as classification and regression [13, 14, 15]. As a result, practitioners commonly resort to using tree-based methods coupled with manual feature engineering, such as computing the product of two column features [16].

The generation of column features, even with domain knowledge, can be challenging and costly. For instance, manual validation to identify useful features is infeasible due to the exponentially many possible combinations to explore [17]. To address this issue, existing feature engineering methods [17, 18, 19] use additional filtering schemes [20, 21, 22] to evaluate and select the useful features automatically. While these approaches reduce manual effort and improve feature quality, they still present several challenges. First, practitioners often resort to using a manually defined search space over which to generate candidate features due to the inherent ambiguity in what constitutes

^{*}Equal contribution

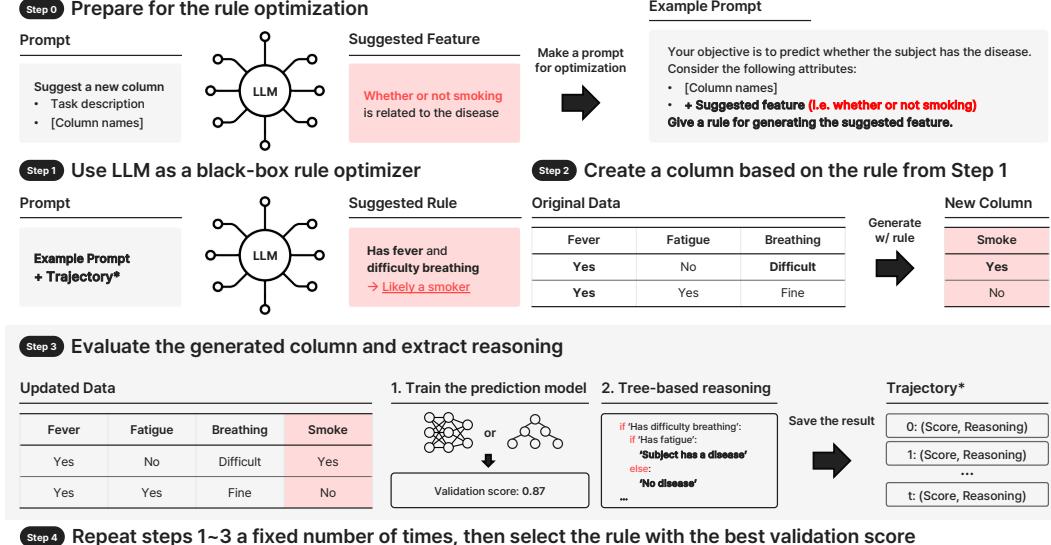


Figure 1: **An overview of OCTree.** (Step 0) Prompt the LLM to propose a name for the new column. (Step 1) Generate a rule by prompting the LLM with feedback on previously generated rules and relevant information for reasoning about the data. (Step 2) Generate a new column feature based on the proposed rule. (Step 3) Train a prediction model on the new data and compute the validation score and tree-based reasoning, provided as feedback for iterative improvements. (Step 4) Repeat steps 1-3 a fixed number of times and select the rule with the best validation score.

informative features [17, 23]. However, this still requires substantial computation for validating candidate features, particularly as the number of features and the complexity of the search space grow. Furthermore, they overlook more effective experimental designs, solely relying on validation scores to select good features, even though information from past experiments is useful for better selection.

Motivated by this, we propose to approach this problem from a novel perspective: optimization to discover effective generation rules, leveraging the language understanding and reasoning capabilities of large language models (LLMs). Recent research has demonstrated that LLMs can optimize various non-differentiable problems using prompts that describe the optimization task in natural language [24, 25, 26]. This suggests the potential for LLMs to automatically generate and iteratively refine feature generators without the need for manually specifying the rule space. For example, the reasoning capabilities of LLMs allow incorporating feedback on their previous outputs into the process for iterative refinement. Moreover, linguistic contexts, such as column names (*e.g.*, ‘Gender’ and ‘Age’) and categorical values (*e.g.*, ‘Female’ and ‘Male’), could be naturally integrated into the optimization [27, 28, 29, 30], which is difficult, if not impossible, with conventional methods.

Contributions. In this work, we leverage LLMs to generate novel column features for tabular prediction tasks, proposing *Optimizing Column feature generator with decision Tree reasoning (OCTree)*, a generic framework for automatic feature generation using LLMs. Figure 1 illustrates an overview of our framework. Our approach begins by prompting an LLM to propose a name for a novel column feature based on the task description, such as ‘Trading Volume’ for stock price prediction. This initial suggestion guides the LLM in exploring and refining the corresponding feature values. Then, we further leverage the reasoning capability of LLMs to produce a *good* rule that generates values for the newly introduced column feature based on the existing ones. Specifically, starting from an initial rule r_0 , we let the LLM iteratively improve the current rule r_t by using extracted reasonings d_0, d_1, \dots, d_t and validation scores s_0, s_1, \dots, s_t attained by the prediction model as feedback. Here, we let d_t be the language description of a decision tree under the entire dataset and the new feature generated by r_t . Namely, we propose using such *decision tree reasoning* to provide the LLM with effective knowledge of the past experiments, *i.e.*, the prediction model trained with the generated features, providing learned knowledge of the entire dataset. This procedure is iterated for a fixed number of times, after which we select the rule with the highest validation score.

We assess the effectiveness of OCTree through extensive evaluations on real-world datasets (*e.g.*, stock price and patient mortality prediction) from various sources, including recent Kaggle competitions.

Our experimental results demonstrate that OCTree consistently enhances the performance of various prediction models, including gradient-boosted decision trees [11] and deep neural networks [31, 32], for both classification and regression tasks. We also assess OCTree on datasets where language descriptions are unavailable, *i.e.*, all feature values and column names are anonymized during preprocessing. Even on these datasets, OCTree reduces relative prediction errors by an average of 5.0% compared to the best baseline model, *i.e.*, XGBoost on the 19 classification tasks benchmarked by Grinsztajn et al. [13]. Here, we use the Llama 2 7B model fine-tuned on high-quality dialogue data, for making the model understand and generate contextually relevant and coherent rules. We also show that OCTree outperforms recent automatic feature engineering methods [17, 18]. Lastly, we verify that the features generated for one type of model (*e.g.*, simpler) can often be used to enhance the performance of other types of models (*e.g.*, more complex).

2 Related work

Tabular learning with LLMs. Recent developments in LLMs have encouraged investigation into their potential applications to tabular prediction tasks. Dinh et al. [27] and Hegselmann et al. [28] fine-tuned the GPT-3 [33] and T0 [34], respectively, by serializing tabular data into natural language. Nam et al. [30] utilizes unlabeled data expressed in natural language with LLMs via prompting for few-shot semi-supervised tabular classification tasks. More recently, Yan et al. [35] introduced tabular-specific tokenization to pre-train a single language model with multiple tabular datasets. Instead of using LLMs as prediction models, we explore whether they can effectively generate informative column features useful for tabular prediction tasks. Specifically, we propose to improve various prediction models by generating novel column features using LLM as an optimizer [25].

LLMs as optimizers. The use of prompting techniques has enabled LLM to demonstrate its potential for optimization. This is achieved by describing the optimization problem in natural language and instructing LLM to generate new solutions iteratively based on the previously found solutions and their evaluated scores. In particular, Yang et al. [25] uses LLM to optimize linear regression, the traveling salesman problem, and prompt optimization (*i.e.*, finding instructions to guide LLM to generate the best results). Motivated by these observations, we leverage an LLM to optimize an effective column feature generator. However, unlike previous work, we additionally use decision tree reasonings as feedback, which provides learned knowledge of the dataset in a natural language.

Automatic feature engineering. Automatic feature engineering refers to generating features from raw tabular data without human labor, which can improve the performance of prediction tasks [19]. A variety of automatic feature engineering methods have been proposed [36, 37, 38], and recently, Horn et al. [23] have used iterative feature subsampling using beam search to select informative features. Zhang et al. [17] have proposed feature boosting and pruning algorithms to efficiently and accurately filter the generated features. After the emergence of LLM, Hollmann et al. [19] proposed a context-aware automatic feature engineering method that leverages LLM to generate semantically meaningful features based on the description of a given task. Unlike previous works, we use LLM’s optimization and reasoning capabilities to find good feature generation rules without having to manually define the search space. Furthermore, it is worth noting that our method works in both context-aware and context-agnostic ways (Hollmann et al. [19] requires language-based context information), and can also be used to suggest meaningful features to human annotators.

3 Optimizing column feature generator with decision tree reasoning

In this section, we introduce a framework for automatically generating column features by leveraging the language understanding and reasoning capabilities of LLMs. In a nutshell, our framework utilizes LLMs as optimizers to propose and refine the rules for generating column features. Specifically, we iteratively refine the rules through LLMs to achieve better validation scores, using (1) the validation scores of previously proposed rules and (2) decision tree-based reasoning (extracted from the training data) as input of LLMs to facilitate the optimization task. We first frame the tabular feature generation problem as an optimization of the rule generator (Section 3.1), and then present the core component, coined *Optimizing Column feature generator with decision Tree reasoning (OCTree)* (Section 3.2).

Problem setup. Formally, the goal of tabular prediction tasks is to train a prediction model $f : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{X} is the input space and $\mathbf{x} \in \mathcal{X}$ is an M -dimensional column feature with corresponding

column names $\mathcal{C} = \{c_1, \dots, c_M\}$. For example, ‘ x_1 : Female’ and ‘ x_2 : 36’ are features for the columns ‘ c_1 : Gender’ and ‘ c_2 : Age’, respectively. In classification tasks, $\mathbf{y} \in \mathcal{Y} = \{0, 1\}^K$ represents the label space with K classes, while in regression tasks, $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}$. We use c_{target} to denote the name of the corresponding column, containing the values of the label \mathbf{y} .

3.1 Tabular feature generation as rule generator optimization

We start by framing column feature generation as an optimization problem for the rule generator, where the rule defines a mapping from the original set of features to a new feature. In this regard, our objective is to generate an informative one-dimensional column feature, \mathcal{X}' , by optimizing the rule $r : \mathcal{X} \rightarrow \mathcal{X}'$, *i.e.*, a novel column feature that enhances the performance of the prediction model when trained with the new feature, $f : \mathcal{X} \oplus \mathcal{X}' \rightarrow \mathcal{Y}$. Formally, based on the generated rule w.r.t to the training dataset $r := g(\mathcal{D}_{\text{train}})$, our optimization problem is as follows:

$$\min_r \mathcal{L}_{f^*}(\mathcal{D}_{\text{val}} \oplus r) \text{ subject to } f^* = \arg \min_f \mathcal{L}_f(\mathcal{D}_{\text{train}} \oplus r), \quad (1)$$

where $\mathcal{D} \oplus r := \{\mathbf{x}_i \oplus r(\mathbf{x}_i), \mathbf{y}_i\}_{i=1}^N$, \oplus denotes concatenation of the column features, N is the number of samples in the dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, and g is the rule generator (*i.e.*, LLM \mathcal{M} in our case) applied to the training dataset $\mathcal{D}_{\text{train}}$. \mathcal{L}_f is an objective function evaluated using the prediction model f , such as the mean absolute error measured with a decision tree regressor for a regression task. In short, we optimize a rule to achieve the best validation score measured with $\mathcal{D}_{\text{val}} \oplus r$ on f^* .

However, such a bi-level optimization is often non-differentiable (or extremely compute-heavy) as it requires computing gradients through the optimization process of f^* , making it challenging to learn the rule. Moreover, the generated rule may also require non-differentiable operations such as logical conjunction between categorical features (*e.g.*, ‘Is a Smoker = Has Fever \wedge Has Difficulty Breathing’ in Figure 1). One way to tackle this issue is to use black-box optimizations, such as evolutionary strategies [39] or reinforcement learning [40]. However, these approaches still face several limitations. For instance, they require a manually pre-defined search space for the operations to be used (and it becomes computationally heavy as the search space grows) and may overlook the effective experimental design of feedback for improvement.

To tackle this issue, we propose using LLMs as optimizers to iteratively improve the rule by prompting them with a trajectory of feedback, *i.e.*, the history of previously generated rules along with their corresponding validation scores and reasoning information. This optimization via prompting method [25] has emerged as an efficient and effective tool for black-box optimization. Furthermore, LLMs offer several advantages, *e.g.*, leveraging the semantics of column and feature values for better optimization and the flexibility to operate without a restricted pre-defined search space for rules.

3.2 Generating column features with OCTree

We now describe the core algorithm. First, we prompt the LLM to propose the name of a new column to generate based on the task description, *e.g.*, ‘Smoking Status’ in Figure 1. We then compute the validation score of the prediction model and extract reasoning information from the decision tree fitted with the training dataset to initialize the optimization trajectory, which is used as feedback for the LLM. Here, this novel *decision tree reasoning* effectively conveys knowledge of past experiments, which further facilitates the optimization process. Finally, we continue updating the trajectory as the optimization progresses, allowing the LLM to iteratively improve the rule.

Column name generation. OCTree first generates the *column name* c_{new} of a novel feature. OCTree achieves this by prompting the LLM \mathcal{M} with a prompt p_{col} to recommend a new column name that would be useful for predicting the target (see Appendix A.1): $c_{\text{new}} = \mathcal{M}(p_{\text{col}}(\mathcal{C}, c_{\text{target}}))$. With their language understanding capabilities, LLMs generate reasonable column names. For example, the LLM might propose using trading volume as a new feature for predicting stock prices.

Initialize optimization and extract reasoning. OCTree then generates the initial rule r_0 for deriving a novel column feature from the original column names \mathcal{C} . OCTree achieves this by prompting the LLM \mathcal{M} with p_{init} to propose a rule to predict c_{new} with c_1, \dots, c_M (see Appendix A.2): $r_0 = \mathcal{M}(p_{\text{init}}(\mathcal{C}, c_{\text{new}}))$. Then, we obtain the initial score s_0 evaluated with f^* : $s_0 = \mathcal{L}_{f^*}(\mathcal{D}_{\text{val}} \oplus r_0)$. Also, we extract the decision tree reasoning d_0 with CART [41] fitted on training dataset:

$$d_0 = \text{CART}(\mathcal{D}_{\text{train}} \oplus r_0).$$

CART is a binary tree that recursively splits data based on criteria, such as Gini impurity, to predict outcomes. We use CART because (i) tree-based algorithms, which typically use ensembles of simple decision trees like CART, still outperform deep learning on tabular prediction tasks, and (ii) it is more interpretable than tree ensembles and can be easily expressed in natural language. For example, as illustrated in Figure 1, CART can be expressed in natural language using the if-else syntax. Intuitively, the decision tree reasoning extracted by CART provides valuable insights learned from the entire training dataset. The reasoning explicitly shows the columns that are considered more significant (as nodes in the tree) and the corresponding values (as thresholds of the nodes) used for prediction.

Optimization with decision tree reasoning. To optimize the rule, we describe the task in natural language and provide the trajectory $\mathcal{T}_t = \{(s_i, d_i, r_i)\}_{i=0}^t$, which includes a history of previously proposed rules with the corresponding scores and reasonings. Specifically, we generate a new rule r_{t+1} with the LLM \mathcal{M} using the prompt p_{gen} (see Appendix A.3). The prompt p_{gen} asks \mathcal{M} to generate a new rule that is not in \mathcal{T}_t that would give a better score than the scores in \mathcal{T}_t : $r_{t+1} = \mathcal{M}(p_{\text{gen}}(\mathcal{T}_t, \mathcal{C}, c_{\text{target}}))$. Optimization trajectories are presented in ascending order of score since LLM is more likely to generate entries similar to those that appear later in the list [25, 42]. Finally, we append the score $s_{t+1} = \mathcal{L}_{f^*}(\mathcal{D}_{\text{val}} \oplus r_{t+1})$, decision tree reasoning $d_{t+1} = \text{CART}(\mathcal{D}_{\text{train}} \oplus r_{t+1})$, and rule r_{t+1} to the optimization trajectory $\mathcal{T}_t : \mathcal{T}_{t+1} = \mathcal{T}_t \cup \{(s_{t+1}, d_{t+1}, r_{t+1})\}$. We then optimize rule r with a fixed number of iterations and select the rule with the highest validation score.

Generating multiple features. The optimization steps can be repeated in order to generate multiple useful features. For example, after generating additional column ‘Smoking Status,’ ‘Physical Activity Level’ can be generated based on the original features and ‘Smoking Status.’ Formally, we first generate a new column $\mathcal{X}' = r_{\text{opt}}(\mathcal{X})$, where r_{opt} is the optimized rule for generating a column feature \mathcal{X}' . Then, one can generate a new input space $\mathcal{X}^{\text{new}} = \mathcal{X} \oplus \mathcal{X}'$. Using $\mathcal{D}^{\text{new}} \subseteq \mathcal{X}^{\text{new}} \times \mathcal{Y}$, OCTree iteratively performs the optimization process to generate new column features again. We sequentially introduce multiple new column features until the validation score stops improving.

4 Experiments

In this section, we evaluate the effectiveness of OCTree across a range of tabular classification and regression tasks using diverse datasets. Our findings demonstrate that OCTree consistently improves the performance of various prediction models (Section 4.1). Furthermore, ablation studies validate the effectiveness of the proposed decision tree reasoning and demonstrate the utility of the generated features across different types of prediction models (Section 4.2).

Datasets. First, we select real-world datasets with language descriptions from diverse sources: the Disease, Academic, Enefit, and Tesla Stock datasets were recently released on Kaggle, and the Clinical Trial dataset is from the US National Library of Medicine. These prediction tasks are very practical and compelling in domains such as healthcare (*e.g.*, diagnostics), academia (*e.g.*, student dropout), and finance (*e.g.*, stock price prediction). In addition, to verify that OCTree is also applicable without language descriptions, we select 19 classification datasets benchmarked by Grinsztajn et al. [13]. Overall, when selecting a dataset, it is important to reflect the heterogeneous nature of tabular data [43]: whether it contains both categorical and numerical features, or is limited to only one type. Additionally, the task type, *i.e.*, classification or regression, is also considered. We conduct experiments on datasets that exhibit both characteristics to ensure OCTree’s general applicability to various types of tabular data. Details are provided in Appendix B.

Baselines. To validate our method, we use three baselines. We first consider XGBoost [11], one of the most competitive tree-based methods, as they are still known to be effective in tabular domains. Secondly, we enhance multilayer perceptron (MLP) [31], the base architecture of deep learning models. Finally, we show that OCTree improves on a recently proposed tabular model, HyperFast [32], which is practical for rapid model deployment because classification can be performed instantly. Implementation details, including hyperparameter search space, are provided in Appendix C.

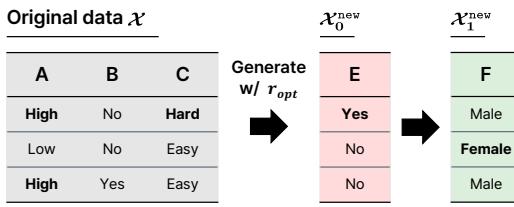


Figure 2: **Generation of multiple features.** The optimization process is repeated to generate multiple column features in sequence.

Table 1: Performance improvement by OCTree on datasets with language descriptions. We report test error rates (%) on three classification tasks (*) and mean absolute error ($\times 10^{-3}$) on two regression tasks (^). The lowest error is denoted in **bold**. Values in parentheses indicate relative error rate reduction from the baseline. We report the mean error and standard deviation across the three random splits, except for two regression tasks (time series tabular data), which are split by time index. N/A indicates not applicable since HyperFast [32] is a classification model.

Method	LLM	Tesla [†]	Enefit [†]	Disease*	Clinical*	Academic*
<i>XGBoost [11]</i>						
Baseline	-	6.61	8.00	28.09 ± 7.9	46.27 ± 5.0	14.15 ± 0.6
OCTree	Llama 2	5.56 (15.9%)	8.00 (0.0%)	26.19 ± 7.2 (6.8%)	45.07 ± 4.1 (2.6%)	14.11 ± 0.5 (0.3%)
OCTree	GPT-4o	5.48 (17.1%)	7.82 (2.3%)	25.72 ± 6.6 (8.4%)	43.75 ± 4.4 (5.4%)	13.74 ± 0.1 (2.9%)
<i>MLP [31]</i>						
Baseline	-	7.41	33.53	38.10 ± 3.6	41.77 ± 1.7	14.41 ± 0.8
OCTree	Llama 2	5.23 (29.4%)	29.99 (10.6%)	32.86 ± 5.7 (13.7%)	39.80 ± 2.3 (4.7%)	14.26 ± 0.7 (1.0%)
OCTree	GPT-4o	5.01 (32.4%)	21.68 (35.3%)	30.95 ± 5.8 (18.8%)	39.25 ± 0.5 (6.0%)	14.22 ± 0.5 (1.3%)
<i>HyperFast [32]</i>						
Baseline	-	N/A	N/A	28.57 ± 10.0	43.64 ± 1.1	14.67 ± 0.7
OCTree	Llama 2	N/A	N/A	28.10 ± 9.2 (1.6%)	41.45 ± 1.7 (5.0%)	14.49 ± 0.5 (1.2%)
OCTree	GPT-4o	N/A	N/A	27.14 ± 3.8 (5.0%)	42.00 ± 1.5 (3.8%)	14.49 ± 0.5 (1.2%)

Common setup. For all datasets, 60% of the data is used for training, 20% for validation, and 20% for testing. Following Gorishniy et al. [31], we use learned embeddings for categorical features when training MLPs (note that HyperFast handles categorical features automatically). For all experiments, we use CART [41] with a maximum depth of 4 to extract decision tree reasoning, provided to the rule generating LLM in the prompt. Unless noted otherwise, we use the Llama 2 model [44] at the 7B scale, fine-tuned on dialogue data to enhance its chat capability. Specifically, we fine-tune the pre-trained model on UltraChat [45], a dataset of dialogues that has been used to produce strong chat models such as UltraLM [45]. Our empirical findings suggest that open models even at moderate scales can be effective, particularly when equipped with strong chat capabilities. Further analysis on performance comparison depending on the types of LLMs is presented in Section 4.1.

4.1 Main results

Results on datasets with language descriptions. We first describe the details of the experiments when clear language descriptions are provided (*e.g.*, column names and categorical features). In these cases, the LLM generates a logical rule in *natural language*. Since the logical rule is easily converted to Python code, we prompt the LLM to convert it. In this case, we use gpt-4o with a temperature of 0.0 (since the Llama 2 model often fails to do this; see Appendix A.4 for used prompt).

As shown in Table 1, our framework consistently improves various types of baseline models. For instance, when we generate column ‘Trading Volume’ for the Tesla Stock dataset using Llama 2 during training of the XGBoost, the relative error decreases by 15.9%. Additionally, our framework is compatible with any kind of LLM. In particular, while Llama 2, the open-source model, improves the baseline models with a high margin, one can even improve with the advanced model. For example, using GPT-4o, the most recent LLM by OpenAI, our method reduces the relative error by 17.1% on the Tesla Stock dataset for the XGBoost. Remark that we select recently released datasets (*e.g.*, Enefit is a recent Kaggle competition dataset) to curate datasets that LLMs would not have seen.

Results on datasets without language descriptions. In practice, language descriptions that clearly describe the given prediction task are not always available for all datasets. For example, feature names and values are often changed to arbitrary meaningless symbols in many financial and medical datasets to protect confidentiality [46]. Even though the language description does not exist, our framework can be easily extended to this kind of case by using arithmetic rules as feature generators. This is because the superiority of our framework comes from the optimization capability of LLMs [24, 25], using decision tree reasoning as explicit feedback, rather than utilizing language description.

Specifically, to ensure that the datasets do not contain language descriptions, we use ordinal encoding for categorical features and then normalized all features with a min-max scaler to convert the original numeric values to different values. In addition, we use generic indicators for the column names. For

Table 2: **Performance improvement by OCTree on datasets without language descriptions.** We report test error rates (%) on 19 classification tasks from Grinsztajn et al. [13]. The lowest error is denoted in **bold**. Values in parentheses indicate relative error rate reduction from the baseline and N/I indicates no gain. We report the mean error and standard deviation across the three random splits.

Dataset	XGBoost [11]		MLP [31]		HyperFast [32]	
	Baseline	OCTree (Ours)	Baseline	OCTree (Ours)	Baseline	OCTree (Ours)
electricity	8.32 \pm 0.0	6.65 \pm 0.1 (20.1%)	15.64 \pm 0.3	14.82 \pm 0.4 (5.2%)	15.25 \pm 0.5	14.70 \pm 0.5 (3.6%)
rl	23.61 \pm 0.8	19.32 \pm 0.4 (18.2%)	32.03 \pm 4.2	28.30 \pm 1.7 (11.6%)	33.77 \pm 1.3	33.50 \pm 1.2 (0.8%)
compass	22.91 \pm 0.5	18.89 \pm 0.4 (17.6%)	27.41 \pm 1.0	26.78 \pm 0.1 (2.3%)	25.74 \pm 0.6	24.91 \pm 1.1 (3.2%)
covertype	9.10 \pm 0.2	7.96 \pm 0.0 (12.5%)	8.73 \pm 0.4	8.25 \pm 0.3 (5.5%)	9.86 \pm 1.6	9.21 \pm 1.3 (6.6%)
phoneme	10.89 \pm 0.5	10.15 \pm 0.7 (6.8%)	12.06 \pm 0.8	10.98 \pm 0.6 (9.8%)	10.55 \pm 0.7	10.57 \pm 0.9 (N/I)
kddCup09	19.86 \pm 1.1	19.07 \pm 1.4 (4.0%)	24.30 \pm 0.3	24.30 \pm 1.6 (0.0%)	25.75 \pm 0.7	24.46 \pm 1.1 (5.0%)
pol	1.69 \pm 0.2	1.62 \pm 0.2 (4.0%)	1.37 \pm 0.3	1.27 \pm 0.3 (7.3%)	1.70 \pm 0.4	1.55 \pm 0.2 (8.8%)
Magic	14.25 \pm 0.3	13.75 \pm 0.4 (3.5%)	14.60 \pm 0.2	14.50 \pm 0.0 (0.7%)	14.95 \pm 0.2	14.34 \pm 0.5 (4.1%)
california	9.45 \pm 0.6	9.13 \pm 1.0 (3.4%)	11.91 \pm 0.3	11.37 \pm 0.1 (4.5%)	11.75 \pm 0.7	11.02 \pm 0.6 (6.2%)
house_16H	11.66 \pm 0.5	11.32 \pm 0.2 (3.0%)	13.07 \pm 0.2	12.54 \pm 0.6 (4.1%)	12.77 \pm 0.3	12.29 \pm 0.4 (3.8%)
eye_movements	35.06 \pm 0.7	34.17 \pm 2.0 (2.6%)	40.03 \pm 1.2	39.86 \pm 1.9 (0.4%)	41.33 \pm 1.5	40.29 \pm 1.7 (2.5%)
road-safety	21.14 \pm 0.0	20.65 \pm 0.1 (2.3%)	22.17 \pm 0.4	21.87 \pm 0.1 (1.4%)	24.54 \pm 0.3	24.07 \pm 0.4 (1.9%)
kdd_ipums_la	10.89 \pm 1.0	10.69 \pm 1.0 (1.8%)	13.13 \pm 1.3	11.72 \pm 1.5 (10.7%)	16.15 \pm 0.3	13.55 \pm 1.4 (16.1%)
MiniBooNE	5.48 \pm 0.2	5.42 \pm 0.1 (1.2%)	9.69 \pm 0.3	7.35 \pm 0.2 (24.1%)	6.61 \pm 0.4	6.54 \pm 0.2 (1.1%)
credit	22.02 \pm 0.3	21.78 \pm 0.3 (1.1%)	24.43 \pm 0.6	23.23 \pm 0.7 (4.9%)	25.06 \pm 1.1	24.30 \pm 1.8 (3.0%)
Higgs	27.95 \pm 0.7	27.91 \pm 0.2 (0.1%)	29.43 \pm 0.4	28.80 \pm 0.2 (2.1%)	30.04 \pm 0.2	29.73 \pm 0.5 (1.0%)
jannis	20.61 \pm 0.1	20.64 \pm 0.1 (N/I)	22.28 \pm 0.1	22.51 \pm 0.1 (N/I)	24.29 \pm 0.4	23.65 \pm 0.3 (2.6%)
wine	19.11 \pm 3.3	19.18 \pm 3.9 (N/I)	21.53 \pm 3.1	21.59 \pm 1.4 (N/I)	19.18 \pm 2.7	19.31 \pm 2.2 (N/I)
bank-marketing	20.09 \pm 0.3	20.31 \pm 0.6 (N/I)	21.11 \pm 0.4	21.09 \pm 0.4 (0.1%)	21.25 \pm 1.0	21.66 \pm 0.8 (N/I)

example, if the number of columns is $M = 5$, we use $\mathcal{C} = \{\text{'x1'}, \text{'x2'}, \dots, \text{'x5'}\}$ for the column names. For the initial rule, we simply use the product of the two most important features measured by XGBoost, *e.g.*, $x_6 = x_1 \times x_5$. The generated rule is, therefore, in the form of an arithmetic formula.

As shown in Table 2, our framework improves the baseline models even without language descriptions. Specifically, our framework reduces errors by an average of 5.0% over the XGBoost classifier. We hypothesize that our method successfully finds good feature generation rules since LLM automatically generates informative new column features in a black-box manner from the large search space of arithmetic rules since our method does not need to manually specify the search space of the rules.

We experiment with several open LLMs to evaluate their performance when employed as the rule generator in our framework. As shown in Table 3, while all of these models result in improvements over the baseline on the datasets without language descriptions, we find our own model (*i.e.*, Llama 2 7B fine-tuned on UltraChat) to be particularly effective. This demonstrates that our framework can be implemented effectively even with open models at a moderate scale, provided that they have strong chat capabilities. We suspect that this is due to the enhanced ability of such models to understand and generate contextually relevant and coherent rules, which leads to improved optimization outcomes. Furthermore, Code Llama also demonstrates competitive performance; it is trained on code datasets, which include a variety of arithmetic rules, therefore beneficial for suggesting rules for datasets lacking language descriptions. This suggests that incorporating code datasets alongside dialogue datasets may further enhance the performance of the rule generator LLM.

Incorporating previous automatic feature engineering methods. As OCTree is orthogonal to existing automatic feature engineering methods, our method naturally integrates with them. The simplest approach is to first generate features with OCTree and subsequently employ these existing methods to further enhance the feature set. In Table 4, we first compare the performance of using OCTree alone. Here, we compare using XGBoost and MLP to show that our framework outperforms the others for both tree-based and deep learning models. As can be seen from the table, our method outperforms state-of-the-art automatic feature engineering methods (*i.e.*, AutoFeat [23] and OpenFE [17]). We exclude CAAFE [19] from our comparison since this method relies on a language-based context of the dataset, making it inapplicable to the datasets in Table 2. In contrast, other methods, including ours, are context-agnostic and can also be applied without any description (but our

Table 3: **OCTree with Llama 2 variants.** We report the average test error rates (%) and standard deviations across three random seeds on the 19 datasets without language descriptions.

Method	LLM	Avg. Err.
XGBoost	-	16.53 \pm 0.1
OCTree	Llama 2 Chat 7B	16.32 \pm 0.1
OCTree	Code Llama 7B	15.83 \pm 0.2
OCTree	Ours 7B	15.71 \pm 0.4

Table 4: **Comparison with automatic feature engineering methods.** We report a mean error (%) and standard deviation across 22 classification datasets used in Table 1 and 2. The lowest error is denoted in **bold** and the second lowest is underlined. Values in parentheses indicate relative error rate reduction from the baseline model. OCTree[†] indicates ours incorporated with classical methods.

Prediction model	Baseline	AutoFeat [23]	OpenFE [17]	OCTree (Ours)	OCTree [†] (Ours)
XGBoost [11]	18.30 ± 0.3	18.24 ± 0.3 (1.3%)	17.79 ± 0.2 (2.8%)	<u>17.45 ± 0.5 (4.6%)</u>	16.85 ± 0.3 (7.9%)
MLP [31]	20.88 ± 0.1	20.60 ± 0.5 (1.3%)	20.12 ± 0.5 (3.6%)	<u>19.91 ± 0.4 (4.6%)</u>	19.41 ± 0.5 (7.0%)

Table 5: **Ablation study of the proposed decision tree reasoning.** We report a mean error (%) and standard deviation across three random splits on two datasets with language description (*) and two datasets without language description (†). The lowest error is denoted in **bold**. Values in parentheses indicate relative error rate reduction from the baseline model.

Gen. Feat.	DT Reasoning	Disease*	Clinical*	electricity†	kddCup09†
-	-	28.09 ± 7.9	46.27 ± 5.0	8.32 ± 0.0	19.86 ± 1.1
✓	✗	27.62 ± 8.4 (1.7%)	45.61 ± 4.1 (1.4%)	6.89 ± 0.6 (17.2%)	19.47 ± 1.6 (2.0%)
✓	✓	26.19 ± 7.2 (6.8%)	45.07 ± 4.1 (2.6%)	6.65 ± 0.1 (20.1%)	19.07 ± 1.4 (4.0%)

method benefits from clear language descriptions as shown in Table 1 if it is available). Furthermore, we highlight that using our method in combination with OpenFE (*i.e.*, OCTree[†] in Table 4) further improves the performance, resulting in a 7.9% reduction in relative error for XGBoost.

4.2 Ablations and analysis

Ablation study of the proposed components. Our framework consists of two main components: (i) generating new additional column features (Gen. Feat.), and (ii) providing explicit decision tree reasonings as feedback (DT reasoning) during the optimization process. As shown in Table 5, both components are crucial to our framework. First of all, the rules for introducing new column features are successfully optimized even without using explicit decision trees for feedback (*i.e.*, just providing a score as feedback to the LLM), therefore resulting in performance improvements. Secondly, one can get even better performance by providing the decision tree as feedback to the LLM. We hypothesize that providing a decision tree is actually the same as providing learned information from the entire training samples. Also, decision trees include information about important columns and their threshold values used for predictions in their nodes. In addition, deep learning models often lack interpretability (*i.e.*, non-trivial to represent them in prompts), whereas decision trees are used in prompts for LLMs because they are easily converted to natural language using if-else syntax.

Transferring generated rules to other prediction models. While we optimize feature generation rules to improve the performance of a certain single model, it is able to use these generated features in other models to improve their performance. For example, it is highly practical to generate features by evaluating with XGBoost, which has a relatively short evaluation time compared to large deep neural networks, and then use the generated features to improve large models. To verify such transferability of our framework, we first optimize the column generation rules by evaluating with XGBoost, and then train MLP and HyperFast along with the generated features. As shown in Table 6, the generated features are useful for improving the performance of MLP and HyperFast, and show high practicality when features need to be generated in a given amount of time.

Examination on LLM for column generation. During the initialization, LLM recommends a new column feature that is not considered in the original dataset. Here, we conduct an experiment to verify whether LLM actually introduces the most beneficial feature that can improve prediction performance. We perform an analysis from two perspectives: (i) whether LLM can distinguish column features that are more relevant to the target task when given candidates for new column names, and (ii) whether it is actually beneficial to use the real values of the column features if they are obtainable.

Our motivation is that LLM can understand the relationship between the target task and the column feature so that LLM can introduce the new columns needed for the task. To verify this, we first verify that LLM can distinguish column features that are more important for the task. For the experiment, we first corrupt the dataset by removing two existing features. We then prompt LLM to rank these two features (by providing as candidates) in order of most informative to the target task. Finally, we compare the performance of adding each feature to the corrupted dataset. As shown in Table 7, LLM

Table 6: **Performance improvement by transferring generated features.** We optimize the feature generation rule with XGBoost and transfer the generated features to improve MLP and HyperFast, *i.e.*, OCTree_{trans}. We report test error rates (%) and standard deviation across three random seeds on two datasets with language description (*) and two datasets without language description (†). The lower error is denoted in **bold**. Values in parentheses indicate relative error rate reduction from the baseline model and N/I indicates no improvement.

Dataset	MLP		HyperFast	
	Baseline	OCTree _{trans} (Ours)	Baseline	OCTree _{trans} (Ours)
Disease*	38.10 \pm 3.6	35.24\pm4.4 (7.5%)	28.57 \pm 10.0	27.62\pm5.8 (5.8%)
Clinical*	41.77\pm1.2	42.32 \pm 2.3 (N/I)	43.64 \pm 1.1	42.76\pm1.8 (2.0%)
electricity†	15.64 \pm 0.3	15.03\pm0.3 (3.9%)	15.37 \pm 0.4	14.88\pm0.2 (3.2%)
kddCup09†	24.30 \pm 0.3	23.47\pm0.5 (3.4%)	25.62 \pm 0.7	25.22\pm0.9 (1.6%)

Table 7: **LLM distinguishes important columns.** We report a mean error (%) and standard deviation across three random splits on the Disease dataset. Note that both GPT-4o and Llama 2 answer that the cough feature is more important than the cholesterol level feature for the target prediction task.

Column feature		Model
Cough	Cholesterol	XGBoost [11]
✗	✗	34.76 \pm 0.8
✗	✓	33.34 \pm 0.8
✓	✗	30.00 \pm 4.3
✓	✓	28.09 \pm 7.9

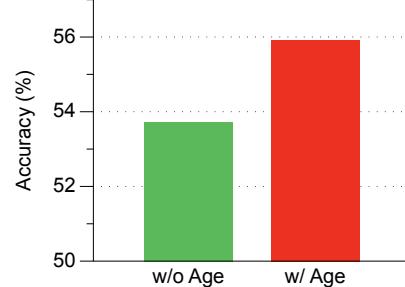


Figure 3: **Imputing with real values, *i.e.*, Age.** We report mean accuracy (%) across three random splits on the Clinical dataset, measured with XGBoost.

indeed distinguishes more informative column features. Here, in the Disease dataset, we remove ‘Cholesterol Level’ and ‘Cough’ from the original dataset. We then ask LLM what attribute is more important in predicting whether a patient has a disease. Of the two, LLM answers ‘Cough,’ which shows more performance improvement than using ‘Cholesterol Level’ as an additional feature. Therefore, generated column features from LLM are beneficial for the prediction.

Using this ability of LLM, we propose to generate a new column through LLM, which practitioners actually need for their target task, even without candidates. For example, on the Clinical Trial dataset, we verify that LLM indeed generates useful column features, which are beneficial for the target task (*i.e.*, predicting the patient’s mortality). During the column name generation step, LLM introduces ‘Age’ as an additional column name. First, we get the actual ages of patients for US National Library of Medicine. We then measure the performance gain by using the actual values obtained from the real world. As shown in Figure 3, imputing the generated column with real values shows an improvement, validating that OCTree actually recommends columns that are useful for the target task. Therefore, we suggest practitioners utilize OCTree (i) to identify additional column features that should be collected or, if this is not possible, (ii) optimize a column feature generation rule.

5 Conclusion

In this paper, we propose OCTree, a generic framework that leverages the power of LLMs (*e.g.*, reasoning capability) for automatically generating column features for tabular prediction tasks. We evaluate the effectiveness of OCTree across various prediction tasks and find that our method consistently improves the performance of diverse prediction models. As a future work, applying feedback-based alignment methods, such as reinforcement learning from human feedback, to further enhance LLMs as rule generators would be an exciting direction to explore.

Limitation. One potential limitation of our work is that evaluating the generated features involves computing the validation scores of the prediction model, which may be time-consuming, if the model requires extensive training. However, as demonstrated by the results in Table 6, this issue can be mitigated by initially generating features for a simpler model and then transferring them to the target model, thereby reducing runtime.

References

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [4] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [6] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.
- [7] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.
- [8] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.
- [9] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.
- [10] Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. *arXiv preprint arXiv:2305.06090*, 2023.
- [11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [12] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.
- [13] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520, 2022.
- [14] Kuan-Yu Chen, Ping-Han Chiang, Hsin-Rung Chou, Ting-Wei Chen, and Tien-Hao Chang. Trompt: Towards a better deep neural network for tabular data. *arXiv preprint arXiv:2305.18446*, 2023.
- [15] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Ganesh Ramakrishnan, Micah Goldblum, Colin White, et al. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 2023.
- [16] Valeria Cherepanova, Roman Levin, Gowthami Somepalli, Jonas Geiping, C Bayan Bruss, Andrew Gordon Wilson, Tom Goldstein, and Micah Goldblum. A performance-driven benchmark for feature selection in tabular deep learning. *Advances in Neural Information Processing Systems*, 2023.

- [17] Tianping Zhang, Zheyu Aqa Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu, Wei Cao, and Li Jian. Openfe: Automated feature generation with expert-level performance. In *International Conference on Machine Learning*, pages 41880–41901. PMLR, 2023.
- [18] Avinash Amballa, Anmol Mekala, Gayathri Akkinapalli, Manas Madine, Naga Pavana Priya Yarrabolu, and Przemyslaw A Grabowicz. Automated model selection for tabular data. *arXiv preprint arXiv:2401.00961*, 2024.
- [19] Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems*, 36, 2024.
- [20] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [21] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [22] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.
- [23] Franziska Horn, Robert Pack, and Michael Rieger. The autofeat python library for automated feature engineering and selection. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, pages 111–120. Springer, 2020.
- [24] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.
- [25] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.
- [26] Michael R Zhang, Nishkrit Desai, Juhan Bae, Jonathan Lorraine, and Jimmy Ba. Using large language models for hyperparameter optimization. *arXiv e-prints*, pages arXiv–2312, 2023.
- [27] Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35: 11763–11784, 2022.
- [28] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.
- [29] Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. Language models are weak learners. *Advances in Neural Information Processing Systems*, 2023.
- [30] Jaehyun Nam, Woomin Song, Seong Hyeon Park, Jihoon Tack, Sukmin Yun, Jaehyung Kim, and Jinwoo Shin. Semi-supervised tabular classification via in-context learning of large language models. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.
- [31] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34: 18932–18943, 2021.
- [32] David Bonet, Daniel Mas Montserrat, Xavier Giró-i Nieto, and Alexander G Ioannidis. Hyperfast: Instant classification for tabular data. *AAAI Conference on Artificial Intelligence*, 2024.
- [33] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

- [34] Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.
- [35] Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. Making pre-trained language models great on tabular prediction. In *International Conference on Learning Representations*, 2024.
- [36] James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE international conference on data science and advanced analytics (DSAA)*, pages 1–10. IEEE, 2015.
- [37] Wei Fan, Erheng Zhong, Jing Peng, Olivier Verscheure, Kun Zhang, Jiangtao Ren, Rong Yan, and Qiang Yang. Generalized and heuristic-free feature construction for improved accuracy. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 629–640. SIAM, 2010.
- [38] Liyao Li, Haobo Wang, Liangyu Zha, Qingyi Huang, Sai Wu, Gang Chen, and Junbo Zhao. Learning a data-driven policy network for pre-training automated feature engineering. In *The Eleventh International Conference on Learning Representations*, 2022.
- [39] Benjamin Doerr and Frank Neumann. A survey on recent progress in the theory of evolutionary algorithms for discrete optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 1(4):1–43, 2021.
- [40] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400, 2021.
- [41] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.
- [42] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.
- [43] Jaehyun Nam, Jihoon Tack, Kyungmin Lee, Hankook Lee, and Jinwoo Shin. Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. In *International Conference on Learning Representations*, 2023.
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [45] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.
- [46] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.
- [47] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Appendix: Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning

A Prompt examples

A.1 Generate a new column

In Listing 1, we present the prompt p_{col} which instructs an LLM to generate a new column name c_{new} . The prompt includes a detailed explanation of each column's feature, specifying its types and values. We restricted c_{new} to be either binary or categorical for convenience.

```
f'''### Your task ###
Your objective is to predict {Objective}. You have access to the following
↪ attributes:
- {Column #1 Name}: (Numerical value of {Column #1 Min} ~ {Column #1 Max})
- {Column #2 Name}: (Boolean)
- {Column #3 Name}: (Categorical value of {Value #1}, {Value #2})
...

To enhance prediction performance, what additional attributes should be
↪ considered? This attribute should be either binary (e.g. valued as 'Yes' or
↪ 'No') or categorical (e.g. valued as 'high', 'low', or 'moderate'). Please
↪ propose a new attribute that is not listed above.

### Answer ###
'''
```

Listing 1: Prompt for the generation of a new column c_{new} .

A.2 Initialize a rule

In Listing 2, we present the prompt p_{init} which instructs an LLM to create an initial rule for generating the new column c_{new} . An LLM generates the new column name c_{new} considering the column features in the dataset and the meaning of c_{new} .

```
f'''### Your task ###
You have access to following attributes:
- {Column #1 Name}: (Numerical value of {Column #1 Min} ~ {Column #1 Max})
- {Column #2 Name}: (Boolean)
- {Column #3 Name}: (Categorical value of {Value #1}, {Value #2}, {Value #3})
...

### Question ###
Give a good rule to predict the '{New Column Name}' ({Output Category #1} or
↪ {Output Category #2}) with the attributes listed above.

### Answer ###
'''
```

Listing 2: Prompt for the rule initialization.

A.3 Generate a rule

In Listing 3, we present the prompt p_{gen} which instructs an LLM to generate a better rule than rules generated by itself. The list of (rule, tree-based reasoning, score) tuples is provided in the prompt. Since the score in the following prompt is the accuracy of the XGBoost classifier, the trajectory is sorted in ascending score order. In the case of a regression task that uses mean absolute error, the trajectory is sorted in descending score order.

```
f'''I have some rules to predict {Objective} with attributes listed below.
- {Column #1 Name} (Numerical value of {Column #1 Min} ~ {Column #1 Max}):
    ↳ {Column #1 Feature Description}
- {Column #2 Name} (Boolean): {Column #2 Feature Description}
- {Column #3 Name} (Categorical value of {Value #1}, {Value #2}, {Value #3}):
    ↳ {Column #3 Feature Description}
...
We also have corresponding decision trees (CART) to predict {Objective} from the
attributes listed above along with predicted {New Column Name}.
The rules are arranged in ascending order based on their scores evaluated with
XGBoost classifier, where higher scores indicate better quality.

Rule to predict {Objective}:
{Rule #1}
Decision tree (CART):
{Decision Tree #1}
Score evaluated with XGBoost classifier:
{Score #1}

Rule to predict {Objective}:
{Rule #2}
Decision tree (CART):
{Decision Tree #2}
Score evaluated with XGBoost classifier:
{Score #2}

...
Give me a new rule to predict {Objective} that is different from the old ones (but
should use the listed attributes above) and has a score as high as possible.

Improved rule:'''
```

Listing 3: Prompt for the rule generation.

A.4 Translate a rule into Python code

In Listing 4, we present the prompt p_{code} which instructs an LLM to translate the rule into Python code. To eliminate the potential for errors, we restrict it in multiple ways. First, we specify the types of variables as far as we can. Next, we instruct LLM to consider the types of feature values before calculating them (e.g. avoid adding categorical value and numerical value).

```
f'''## Rule ##
{Rule to translate}

### Your Task ###
Change the rule into executable Python code. Consider the type of each feature.
Input data (Numpy array (not a Dict)): [{all_column_names}]
- {Column #1 Name}: (Numerical value of {Column #1 Min} ~ {Column #1 Max})
- {Column #2 Name}: (Boolean)
- {Column #3 Name}: (Categorical value of {Value #1}, {Value#2})
...
Output: {Output Category #1} or {Output Category #2}
Give instantly executable code without example usage. Consider the feature value
→ types (avoid calculating categorical value and numerical value). Start with
→ 'def {Rule Function Name}(data):'

### Only python code ###
'''
```

Listing 4: Prompt for translating rule in natural language into Python code.

B Dataset details

In this section, we provide further details on the datasets.

B.1 Using language descriptions.

1. Disease¹

A classification task predicting the patient's diagnosis of the disease given the following attributes:

- **Binary (Yes or No):** Fever, Fatigue
- **Numerical (Range):** Age (25 ~ 90)
- **Categorical:** Gender (Female or Male), Blood Pressure (High, Normal, or Low), Cholesterol Level (High, Normal, or Low)

2. Clinical Trial²

A classification task predicting patient mortality in clinical trials given the following attributes:

- **Binary (Yes or No) - Historical Disease:** Deep Vein Thrombosis, Pulmonary Embolism, Antiandrogen Therapy, Cardiac Failure, Respiratory Failure, Venous Insufficiency, Coronary Artery Disease, Myocardial Infarction, Hypertension, Peripheral Arterial Occlusive Disease
- **Binary (Yes or No) - Medication:** Dexamethasone, Ondansetron, Heparin, Fluorouracil, Ranitidine, Cisplatin, Metoclopramide, Carboplatin, Furosemide

3. Academic³

A classification task predicting whether the student dropout or not given the following attributes:

- **Numerical:** Marital Status, Daytime/Evening Attendance, Previous Qualification, Nationality, Father's Qualification, Father's Occupation, Displaced, Debtor, Tuition Fees up to Date, Gender, Scholarship Holder, Age at Enrollment, International, Curricular Units 1st Sem (Approved), Curricular Units 1st Sem (Grade), Curricular Units 2nd Sem (Approved), Curricular Units 2nd Sem (Grade).

4. Enefit⁴

A regression task predicting the energy consumption of the day given the following attributes:

- **Numerical:** Prediction Unit Id, Day, Hour, Lowest Price Per MWh, Highest Price Per MWh, Installed Capacity, Euros Per MWh, Local Forecast Temperature, Local Forecast Dewpoint, Local Forecast Cloudcover Total, Local Forecast 10 Metre U Wind Component, Local Forecast 10 Metre V Wind Component, Local Forecast Direct Solar Radiation, Local Forecast Surface Solar Radiation Downwards, Local Forecast Total Precipitation.

5. Tesla Stock⁵

A regression task predicting the highest stock price of the target day given the following attributes:

- **Numerical:** Open Price of 2 Days Before, Highest Price of 2 Days Before, Lowest Price of 2 Days Before, Close Price of 2 Days Before, Open Price of 1 Day Before, Highest Price of 1 Day Before, Lowest Price of 1 Day Before, Close Price of 1 Day Before, Open Price of the Target Day, Time Index.

¹<https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset>

²<https://data.projectdatasphere.org/projectdatasphere/html/content/119>

³<https://www.kaggle.com/datasets/missionjee/students-dropout-and-academic-success-dataset>

⁴<https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers>

⁵<https://www.kaggle.com/datasets/guillemservera/tsla-stock-data>

B.2 Without using language descriptions.

For our main result without language descriptions (see Table 2), we use the 19 classification datasets from the tabular benchmark proposed by Grinsztajn et al. [13]. We provide brief statistics of the dataset in this section. Due to limited compute resources, we uniformly subsample 50,000 samples when the original number of samples is larger than 50,000, following the dataset curation scheme from Grinsztajn et al. [13]).

Table 8: **Dataset statistics.** 19 classification datasets benchmarked by Grinsztajn et al. [13].

Dataset	OpenML ID	# Samples	# Features
rl	44160	4970	12
electricity	44156	38474	8
compass	44162	16644	17
wine	44091	2554	11
house_16H	44123	13488	16
MagicTelescope (Magic)	44125	13376	10
Higgs	44129	940160	24
jannis	44131	57580	54
credit	44089	16714	10
eye_movements	44157	7608	23
kddCup09_upselling (kddCup09)	44158	5032	45
road-safety	44161	111762	32
bank-marketing	44126	10578	7
phoneme	44127	3172	5
covertype	44159	423680	54
california	44090	20634	8
kdd_ipums_la_97-small (kdd_ipums_la)	44124	5188	20
MiniBooNE	44128	72998	50
pol	44122	10082	26

C Baseline details

In this section, we outline the hyperparameter search space for the baseline models. For each random split of every dataset, we find the optimal set of hyperparameters using a random sampler run for 400 trials. We utilize the Optuna library [47] for the hyperparameter tuning.

C.1 XGBoost

For XGBoost [11], we follow the hyperparameter search space used in Grinsztajn et al. [13].

Table 9: XGBoost [11] hyperparameters space.

Parameter	Distribution
Max depth	UniformInt [1, 11]
Num estimators	UniformInt [100, 6100, 200]
Min child weight	LogUniformInt [1, 1e2]
Subsample	Uniform [0.5, 1]
Learning rate	LogUniform [1e-5, 0.7]
Col sample by level	Uniform [0.5, 1]
Col sample by tree	Uniform [0.5, 1]
Gamma	LogUniform [1e-8, 7]
Lambda	LogUniform [1, 4]
Alpha	LogUniform [1e-8, 1e2]

C.2 Multilayer perception (MLP)

For MLP, we follow the hyperparameter search space used in Grinsztajn et al. [13] and the architecture used in Gorishniy et al. [31]. Specifically, the MLP architecture for tabular data additionally learns the feature embeddings of categorical features. MLPs are run up to 300 epochs, with early stopping, and the model that performs the best on the validation set is selected. If validation scores do not improve for 40 epochs, we early-stopped. For learning rate schedulers, we use the `ReduceOnPlateau` implementation in PyTorch.

Table 10: MLP [31] hyperparameters space.

Parameter	Distribution
Num layers	UniformInt [1, 8]
Layer size	UniformInt [16, 1024]
Dropout	Uniform [0, 0.5]
Learning rate	LogUniform [1e-5, 1e-2]
Category embedding size	UniformInt [64, 512]
Learning rate scheduler	[True, False]
Batch size	[256, 512, 1024]

C.3 HyperFast

For HyperFast [32], we consider the hyperparameter space used in the original paper.

Table 11: HyperFast [32] hyperparameters space.

Parameter	Distribution
N ensemble	[1, 4, 8, 16, 32]
Batch size	[1024, 2048]
NN bias	[True, False]
Stratify sampling	[True, False]
Optimization	[None, ‘optimize’, ‘ensemble_optimize’]
Optimize steps	[1, 4, 8, 16, 32, 64, 128]
Seed	UniformInt [0, 9]

D Compute resources

We ran our experiments on a variety of machines. We used:

- CPU: Intel(R) Xeon(R) Gold 6226R, GPU: RTX 3090
- CPU: Intel(R) Xeon(R) Gold 6426Y, GPU: RTX A4090
- CPU: Intel(R) Xeon(R) Gold 6426Y, GPU: RTX A6000

E Broader impacts

Since our method is particularly effective when collecting real data is expensive or restricted, practitioners in need might give big attention (*e.g.*, in the finance or medical domain, where real data is hard to obtain due to issues like privacy). However, as features introduced by OCTree are artificially generated, careful inspection of the generated features is needed.