FROM DECISION TREES TO STRONG REASONING MODELS FOR TABULAR DATA CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular data classification in high-stakes domains such as healthcare and finance requires both high accuracy and faithful, human-understandable reasoning. Symbolic decision trees provide explicit decision rules but offer limited semantic explanations, while adapting large language models (LLMs) to domain-specific tabular reasoning requires scalable and reliable reasoning supervision. We propose **ReSS**, a framework that combines symbolic decision structure with LLM semantic knowledge to curate reasoning data and train specialized tabular reasoning models. ReSS extracts decision paths from training instances whose tree predictions agree with their ground-truth labels. These paths, together with input features and labels, guide an LLM to construct natural-language rationales that follow the selected symbolic constraints and incorporate domain-specific interpretations. A scaffold-invariant augmentation strategy is proposed to expand the training dataset while preserving the symbolic constraints. The resulting data are used to fine-tune an open LLM that generates reasoning and predictions directly from input features, without requiring the decision tree at inference. We assess reasoning faithfulness through hallucination rate, explanation necessity, and explanation sufficiency. Experiments on medical and financial benchmarks demonstrate accuracy gains of up to 10% over decision trees and standard fine-tuning baselines, alongside grounded and consistent explanations. Sensitivity analyses show that predictive accuracy remains relatively stable across the evaluated tree configurations, demonstrating that useful reasoning supervision can be obtained from trees of varying predictive quality.

1 INTRODUCTION

In high-stakes domains such as healthcare (Clore et al., 2014), finance (Moro et al., 2014) where tabular data is common, predictive accuracy alone is often insufficient. Models are also expected to provide faithful and human-understandable reasoning that enables practitioners to inspect, trust, and act upon their predictions. Despite decades of progress in tabular learning (Breiman et al., 1984; Chen et al., 2016; Si et al., 2024), constructing models that jointly achieve strong performance and meaningful reasoning remains a fundamental challenge.

Traditional machine learning methods have long dominated tabular prediction, with tree-based approaches such as decision tree (Breiman et al., 1984), random forests (Breiman, 2001), XGBoost (Chen et al., 2016), and LightGBM (Ke et al., 2017) achieving strong empirical performance. From a reasoning perspective, however, these models remain limited. Tree ensemble methods do not expose an explicit, instance-level decision process, while single decision trees express reasoning solely through symbolic rules defined over low-level feature thresholds. Although such representations faithfully reflect the model's internal logic, they lack semantic grounding in domain concepts and therefore cannot be readily interpreted by non-expert users, limiting their practical utility in real-world, high-stakes decision-making settings. Recent advances in deep learning (DL) for tabular data have produced models that match or even surpass the performance of tree-based methods (Huang et al., 2020; Arik & Pfister, 2020; Abutbul et al., 2020; Si et al., 2024). However, these models usually provide limited or no interpretability regarding its decision process. Even when post-hoc explanations are available, they often fail to constitute truly human-understandable reasoning chains.

Recently, studies have leveraged LLMs for tabular data prediction (Dinh et al., 2022; Hegselmann et al., 2023). These methods transform tabular rows into natural language prompts to utilize the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LLM's prior knowledge for prediction. However, they essentially function as direct-prediction models rather than reasoning models, as they map inputs directly to final labels while bypassing the intermediate decision-making process. Consequently, they fail to provide a transparent, step-by-step logical chain to justify their predictions.

This raises an important question: *How to efficiently train a faithful reasoning model for tabular data prediction?* There are two key challenges to be addressed: (i) how to curate high-quality reasoning data for fine-tuning an LLM; and (ii) how to ensure that the reasoning traces generated by the learned model are both faithful and explainable. Curating high-quality reasoning data for high-stakes domains such as healthcare and finance typically requires domain experts and is therefore expensive, time-consuming, and difficult to scale. Moreover, while current LLMs are capable of producing chain-of-thought (CoT)-like outputs, these explanations are not guaranteed to be faithful and causal to the tabular data prediction (Kambhampati et al., 2025; Barez et al., 2025).

To address these challenges, we propose **ReSS**, a systematic framework that curates high-quality reasoning data via symbolic scaffolds and trains faithful reasoning models for tabular prediction. ReSS first trains a decision tree on tabular data and extracts instance-level decision paths from samples that are correctly predicted by the decision tree, which serve as explicit symbolic scaffolds capturing the model's underlying decision logic. These scaffolds, together with input features and labels, are used to guide an LLM to generate grounded, step-by-step natural-language reasoning that strictly adheres to the logical constraints encoded in the decision paths, yielding a reasoning dataset that is both faithful to the original decision process and interpretable by humans. To further enhance generalization and explainability, ReSS incorporates a scaffold-invariant data augmentation strategy that expands the reasoning dataset and introduces out-of-distribution examples by perturbing input features while preserving the symbolic decision paths and labels.

We summarize our contributions as follows:

- We propose ReSS, a systematic framework that leverages decision-tree paths as symbolic scaffolds to guide an LLM using its inherent domain knowledge to curate high-quality grounded reasoning data, and trains faithful reasoning models for tabular prediction.
- We introduce a scaffold-invariant data augmentation strategy that preserves symbolic decision logic while generating both in-distribution and out-of-distribution reasoning data, enabling improved generalization and more robust explainability.
- We propose a set of metrics to measure the faithfulness and explainability of generated CoT-like outputs by the trained reasoning model, including hallucination rate, explanation sufficiency, and explanation necessity.
- We conduct experiments on five real-world datasets, including two from medicine and three from finance. The results demonstrate the effectiveness of ReSS and the scaffold-invariant data augmentation, and validate the faithfulness and causal impact of the generated rationales.

2 Related work

While tabular data prediction has long been dominated by tree-based ensemble models, recent years have seen the emergence of new approaches based on DL and LLMs.

Deep Learning Methods for Tabular Prediction. Early DL approaches for tabular prediction typically rely on generic architectures such as MLPs, which lack inductive biases aligned with the structure of tabular decision manifolds and often struggle to match the performance of tree-based methods (Arik & Pfister, 2020). To overcome this limitation, subsequent work introduced tabular-specific inductive biases into neural architectures, including TabNet (Arik & Pfister, 2020), DNF-Net (Abutbul et al., 2020), TabTransformer (Huang et al., 2020), TabPFN (Hollmann et al., 2023). But these methods are not interpretable. InterpreTabNet (Si et al., 2024) attempts to provide intrinsic interpretability through instance-wise feature selection mechanisms, by employing an LLM to map from the learned feature mask onto natural language text describing the learned signal. Despite these advances, they do not aim to produce explicit, human-understandable reasoning processes.

LLMs for Tabular Prediction. Recent work has explored applying LLMs to tabular prediction by serializing tabular records into natural-language descriptions, allowing LLMs to operate in their native textual space. Depending on whether model parameters are updated, these approaches can be broadly categorized into inference-only and fine-tuning–based methods. TABLET (Slack & Singh, 2023)

studies inference-only prediction by prompting pretrained LLMs, but reports that simple tree-based baselines consistently outperform LLMs under comparable settings. Fine-tuning–based approaches, including LIFT (Dinh et al., 2022) and TabLLM (Hegselmann et al., 2023), demonstrate that language pretraining and meaningful feature semantics can improve tabular prediction, particularly in low-shot regimes. However, these gains are inconsistent across datasets, and overall performance still struggles to reliably surpass strong tree-based models. These existing methods focus solely on final predictions and do not generate explicit reasoning.

Large Reasoning Models. Recently, Large Reasoning Models (LRMs) trained by RL have demonstrated strong capabilities in solving complex tasks (Guo et al., 2025; Team et al., 2025), positioning them as a promising paradigm for tabular prediction. Beyond generating final predictions, LRMs can produce explicit reasoning chains. Many RL algorithms have been developed including, GRPO (Guo et al., 2025), REINFORCE++ (Hu et al., 2025), DisCO (Li et al., 2025), to name just a few. Recent works (Xu et al., 2025; Cai et al., 2025) have applied RL directly to a base model for tabular prediction. However, reasoning capabilities of existing LRMs do not readily transfer to tabular reasoning, which requires both semantic and numerical understanding (Cai et al., 2025). As a result, directly generating reasoning traces from existing LRMs without fine-tuning on the tabular reasoning data may lead to hallucinations, undermining the faithfulness and trustworthiness of the outputs. Werling et al. (2025) proposed an approach for curating reasoning traces from tabular data by constructing contrastive triplets of an example including the sample itself, the nearest samples with the same and different labels, and prompting an generalist LLM to transform the triplets into contrastive decision processes, which are then used to supervise fine-tuning of a target LLM. However, this reasoning data is specifically constructed for medical question-answering tasks.

Faithfulness of Reasoning Models. The reasoning traces generated from LRMs, although provide explicit explanation for the final prediction, could be unfaithful (Turpin et al., 2023; Arcuschin et al., 2025; Barez et al., 2025). These works show that the reasoning chains of a model can be misleading or suffer from subtle biases that are not reflected in the traces, hampering the trustworthiness of the prediction. Thus, to ensure that the answer is trustworthy, evaluating the faithfulness of model reasoning explanation is imperative for any task, including tabular prediction. An important quality of a faithful reasoning is consistency, where important sentences or features in the reasoning trace have strong causal impact (Paul et al., 2024; Atanasova et al., 2023) and replacing them with their counterfactual version should greatly affect the output of the model. However, existing works that use LRMs in tabular prediction (Xu et al., 2025; Cai et al., 2025; Werling et al., 2025) only focus on evaluating prediction accuracy, lacking a comprehensive test to quantify the faithfulness quality of their models' reasoning.

3 ReSS: Reasoning via Symbolic Scaffolds

Notations. We consider classification problems on tabular data. Each sample can be represented by a tuple $(\mathbf{x}, y)$, where $\mathbf{x}$ is a feature vector $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ and $y \in \mathcal{Y}$ is the class label. The features in $\mathbf{x}$ may be numerical, ordinal, or categorical. Given an example $\mathbf{x}$, our goal is to fine-tune an LLM that given input features $\mathbf{x}$ predicts the label y, along with a reasoning trace $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$ to support and explain the final prediction, where each $\mathbf{z}_k$ represents a reasoning step consisting of a sequence of reasoning tokens.

To input each tabular row into an LLM, we serialize the feature vector $\mathbf{x}$ into a sequence of natural-language tokens. Let us consider an example of diabetic prediction problem, where features may include age, BMI (Body Mass Index), DBP (Diastolic Blood Pressure), etc. Given a feature vector {Age: 31, BMI:10.5, DBP: 66, ... }, it is serialized to a text, e.g., *"'The age is 31, the body mass index is 10.5, the diastolic blood pressure is 66, ..."*. Without causing any confusion, let $\mathbf{x}, \mathbf{y}$ denote the serialized texts for the input features and the output label, respectively. Let $\pi_{\mathbf{w}}$ denote an LLM with parameters $\mathbf{w}$. We denote by $\pi_{\mathbf{w}}(\mathbf{y}|\mathbf{x})$ the probability of generating $\mathbf{y}$ given $\mathbf{x}$.

Before delving into the proposed approach, we first present two straightforward approaches for fine-tuning an LLM for tabular data prediction and discuss their limitations.

Baseline: Direct SFT. Given serialized data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, a straightforward approach to fine-tune an LLM is to perform the supervised fine-tuning (SFT) by minimizing the objective $-\frac{1}{n}\sum_{i=1}^n \log \pi_{\mathbf{w}}(\mathbf{y}_i|\mathbf{x}_i, \mathbf{t})$, where $\mathbf{t}$ denotes the task-related prompt. A limitation of this approach is that the model does not output any reasoning for the final prediction.

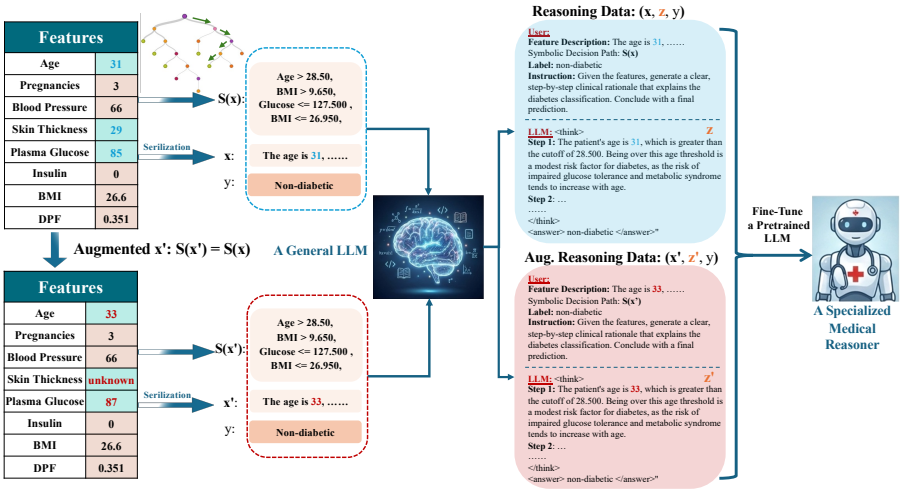

Figure 1: An illustration of the ReSS pipeline applied to the diabetes prediction problem.

Baseline: Direct Reasoning Curation and SFT. One way to address the above limitation is to first curate a reasoning data and then conduct SFT. However, curating high-quality reasoning data from domain experts in high-stakes domains is expensive, time-consuming, and hard to scale. A simple way is to leverage the knowledge of a general-purpose LLM. To this end, we can prompt an LLM with all features and the label $(\mathbf{x}, \mathbf{y})$ and a guidance prompt $\mathbf{t}_c$ to generate a rationale for how a label $\mathbf{y}$ is predicted given the features in $\mathbf{x}$. We let $\mathbf{R}(\mathbf{x}, \mathbf{y})$ denote the reasoning tokens generated by the LLM. Then, SFT is conducted by minimizing the objective $-\frac{1}{n}\sum_{i=1}^{n} \log \pi_{\mathbf{w}}(\{\mathbf{z}_i, \mathbf{y}_i\}|\mathbf{x}_i, \mathbf{t})$, where $\mathbf{z}_i = \mathbf{R}(\mathbf{x}_i, \mathbf{y}_i)$, $\{\mathbf{z}_i, \mathbf{y}_i\}$ denotes the concatenation of reasoning tokens and the final label tokens, and $\mathbf{t}$ denotes the task-related prompt.

3.1 Using Decision Tree Paths as Symbolic Scaffolds

Motivation of Using Symbolic Scaffolds. A consequence of the direct curation approach discussed above is that the generated rationale may contain many non-useful features. An example given in Figure 11 (Appendix C.2) shows that the rationale generated by ChatGPT could iterate through all eight features on the diabetes dataset. To explain its limitation and motivate the proposed approach, we consider a simple example below. Let us consider the input data $\mathbf{x} \in \{1, -1\}^{2d}$, which contains a sequence of $\{1, -1\}$ tokens. Suppose the y is given by $y = \prod_{i=1}^{d} x_{P(i)}$, where P is a fixed (but unknown) permutation of $\{1, \ldots, 2d\}$. Hence, y is just the parity of d fixed tokens in $\mathbf{x}$. The existing work (Wies et al., 2023) has shown that if we directly learn a model $\pi_{\mathbf{w}}(y|\mathbf{x})$ to predict y given $\mathbf{x}$, the learning requires an exponential number of samples in the order of $O(\exp(d))$. Hence, it is impossible to learn the correct model for predicting y given $\mathbf{x}$ with limited data. The challenge of the problems lies that finding the correct permutation P out of all permutations is exponentially hard. However, if we tell the learner that the final prediction y only relies on the features in $x_{P(1)}, \ldots, x_{P(d)}$, it could dramatically simplify the learning process.

This motivates the proposed approach of using **symbolic scaffolds** to guide the feature selection and decision process. A symbolic scaffold refers to an explicit external structure such as rules, diagrams, decision trees, or step-by-step procedures that humans use to organize and guide their reasoning processes. In developmental psychology (Vygotsky, 1978), symbolic scaffolds are known to support problem solving by breaking complex tasks into interpretable intermediate steps, thereby reducing cognitive load and enabling systematic inference. Rather than replacing reasoning, these symbolic aids shape how reasoning unfolds by constraining the space of possible actions and making intermediate states explicit. It is important to note that the concept of a **symbolic scaffold differs from that of chain-of-thought (CoT)** (Wei et al., 2022) in reasoning models. CoT concerns how reasoning is expressed, whereas a symbolic scaffold determines how the reasoning process is structured and constrained.

To construct a symbolic scaffold for a tabular data, it is natural to use a decision tree model. To this end, we first train a decision tree on the given tabular data. For each input instance $\mathbf{x}$, the tree induces a unique root-to-leaf symbolic decision path, which we formalize as an instance-specific symbolic

scaffold denoted by $S(\mathbf{x})$. In particular, the scaffold $S(\mathbf{x})$ is an ordered set of feature-level logical constraints that characterizes the sufficient conditions leading to the prediction for $\mathbf{x}$. Importantly, $S(\mathbf{x})$ is induced in a data-dependent manner: both the selected features and their associated thresholds arise from the empirical data distribution. As a result, the scaffold provides a compact, verifiable constraint space that can be directly used to guide rationale generation. An example of a scaffold $S(\mathbf{x})$ from the decision tree is given in Figure 1, "age >28.5, BMI >9.650, ... ". Similar as before, we serialize $S(\mathbf{x})$ into a sequence of tokens.

3.2 Symbolic Scaffold Informed Reasoning Dataset Curation

Given a symbolic scaffold that specifies the constraints of the decision process, we leverage the input features, the output label, and the corresponding symbolic scaffolds of correctly predicted instances to generate reasoning data. To this end, we prompt a generalist LLM (e.g., ChatGPT) by asking it to provide the rationale how the label $\mathbf{y}$ is predicted from the input features $\mathbf{x}$ following the symbolic decision path $S(\mathbf{x})$ that leads to the correct label. Generalist LLMs encode rich domain knowledge acquired from large-scale pretraining, which can be leveraged to translate symbolic rules into coherent natural language explanations. In our setting, we use a generalist LLM as a constrained generator that injects domain knowledge while strictly adhering to symbolic constraints as specified in the symbolic scaffold.

For each example $(\mathbf{x}, \mathbf{y})$ that is correctly predicted by the decision tree model, we let $\mathbf{R}(\mathbf{x}, \mathbf{y}, S(\mathbf{x}))$ denote the curated reasoning tokens. As a result, we collect a set of reasoning data $\{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i\in\mathcal{C}}$, where $\mathbf{z}_i = \mathbf{R}(\mathbf{x}_i, \mathbf{y}_i, S(\mathbf{x}_i))$, and $\mathcal{C} \in [1, \ldots, n]$ denotes the subset of data that is correctly predicted by the decision tree model.

3.3 Scaffold-Invariant Data Augmentation

One might note that we only use the data that is correctly predicted by the decision tree model. This is important for LLM to learn consistent decision rules. However, it will shrink the size of training data for SFT. Below, we present an effective data augmentation strategy.

A data augmentation is usually performed by perturbing the input features. However, a major challenge in data augmentation for tabular prediction is ensuring the consistency between perturbed features and the label, as unconstrained perturbations may cause an instance to no longer correspond to its original label. Our framework addresses this issue by enforcing **scaffold invariance** during augmentation. Specifically, for each correctly predicted training instance by the decision tree model, its symbolic scaffold $S(\mathbf{x})$ characterizes the decision logic leading to its label. This scaffold defines a conjunction of feature-level constraints and serves as an explicit invariance during data augmentation.

To generate an augmented sample, we randomly sample some features (e.g., by flipping a coin for each feature) and for each feature we randomly perturb its value in a valid set such that it does not violate the decision rule specified by the scaffold $S(\mathbf{x})$. Perturbed values are sampled within a feature-specific valid range. To improve the out-of-distribution generalizability, we also allow for a certain probability (e.g., 0.2) to perturb a sampled feature that do not appear in the decision path to the value `"unknown"`. This not only simulates out-of-distribution data but also facilitates the evaluation of explainability presented in next section. As a result, all perturbations satisfy the original decision-path constraints, ensuring that the symbolic scaffold remains unchanged. By preserving the decision scaffold while allowing controlled feature-level variation, our augmentation procedure generates diverse yet label-scaffold consistent samples.

For a perturbed sample $(\mathbf{x}', \mathbf{y})$ corresponding to an original data point $(\mathbf{x}, \mathbf{y})$, we follow the same procedure described in the previous subsection to generate the reasoning traces $\mathbf{z} = \mathbf{R}(\mathbf{x}', \mathbf{y}, S(\mathbf{x}'))$. Let $\mathcal{D} = \{(\mathbf{x}_j, \mathbf{z}_j, \mathbf{y}_j)\}_{j=1}^{m}$ denote the final set of reasoning data including the original data and the perturbed data, we conduct SFT on the resulting reasoning data $\mathcal{D}$.

4 Evaluation of Faithfulness and Explainability

Since the ReSS-trained reasoning model is derived from an explainable and faithful decision tree model, a natural question is whether the learned reasoning model preserves the faithfulness and explainability of the underlying decision tree. To answer this question, we need to evaluate whether the model-generated reasoning is faithful and explainable. Without systematic evaluation, reasoning

traces may appear plausible while relying on nonexistent features or implausible feature values, and including artificial rationales that are not causal for the final predictions. To this end, we consider the following three complementary criteria: hallucination, explanation sufficiency, and explanation necessity, which together characterize whether the reasoning is grounded in the input, complete, and causally relevant to the prediction.

Hallucination. We measure hallucination by identifying reasoning steps that violate the input evidence. Specifically, we consider three types: (i) *feature existence hallucination (FEH)*, where the reasoning refers to features absent from the input; (ii) *feature value hallucination (FVH)*, where incorrect or unsupported feature values are asserted; and (iii) *comparison hallucination (CH)*, where invalid numerical or categorical comparisons are made. The hallucination rate is computed at the step level: for each reasoning step, we mark it as hallucinated if it contains a violation, and report the fraction of hallucinated steps among all steps.

Explanation Sufficiency. To assess explanation sufficiency, we mask features that are not referenced in the generated reasoning traces by setting their values to `unknown` in the input, and prompting the model to re-generate reasoning and prediction, and evaluating the prediction accuracy. Specifically, we progressively mask the number of such features from $k = 1$ to 4. If the accuracy remains stable under these perturbations, the original reasoning traces are considered to be sufficient.

Explanation Necessity. To assess explanation necessity, we mask the features explicitly appearing in the generated reasoning traces by setting their values to `unknown` in the input, and measure the resulting performance degradation. Similarly, the number of masked features is varied from $k = 1$ to 4. A significant drop in accuracy suggests that the identified features are necessary for the prediction, while a minor change indicates redundant or spurious reasoning.

5 EXPERIMENTS

We conduct a comprehensive set of experiments to systematically evaluate the effectiveness of our proposed method.

Datasets. We conduct experiments on five high-stakes tabular datasets from the medical and financial domains. In medical domain, we use an Alzheimer's disease (AD) dataset with three classes: *normal cognition, mild cognitive impairment, and Alzheimer's disease* (Kwon et al., 2024), and a Diabetes dataset to predict *diabetes or non-diabetes* (Hegselmann et al., 2023). In financial domain, we use the Creditg dataset to classify credit risk as *good or bad* (Hegselmann et al., 2023), and HomeLoan dataset to predict *approved or not approved* decisions sourced from Kaggle[1]. Additionally, we incorporate a large-scale dataset *Give Me Some Credit* (Credit-D) (Grinsztajn et al., 2022) to predict *default or non-default* outcomes over a two-year period.

Detailed dataset statistics and data split information are summarized in Table 1(left). For the AD dataset, we use its official split. For Credit-D, considering its larger scale, we perform a single split using a fixed random seed (Seed 0). For other three small datasets, we use random 80%/10%/10% split for training, validation and testing, repeated three times.

To serialize tabular instances into natural-language text for LLMs, we adopt the Text Template proposed in (Hegselmann et al., 2023) for the Diabetes and Creditg datasets. For AD, HomeLoan, and Credit-D, we design dataset-specific templates, with details in Appendix A.

Baselines. We compare our method against baselines covering classical machine learning and LLM-based approaches that share the same prediction objective. Specifically, we include traditional machine learning methods (1) Decision Tree (Breiman et al., 1984) and (2) XGBoost (Chen et al., 2016). For baselines involving fine-tuning LLMs, we further include (3) Direct SFT and (4) Direct Reasoning Curation followed by SFT (DRC+SFT); (5) Direct RL approach that directly conducts RL on the base LLM to exploit the existing reasoning capabilities of the base LLM, corresponding to the approach in (Xu et al., 2025). Deep learning methods for tabular prediction, including TabNet (Arik & Pfister, 2020) and TabPFN (Hollmann et al., 2023), are compared separately in Appendix E.4, as they are pure prediction models that produce no interpretable reasoning and therefore target a different objective from ReSS. For all LLM fine-tuning methods, we use Qwen-2.5-3B-Instruct[2] as

[1] https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval

[2] Released under Apache License 2.0.

Table 1: **Left:** Dataset statistics and data split details. **Right:** Hallucination rates (%) across different datasets. We report three types: Feature Existence Hallucination (FEH), Feature Value Hallucination (FVH), and Comparison Hallucination (CH). Results are averaged over three random seeds.

Dataset	#Feature	Train	Val.	Test	$\text{Train}_{\text{aug}}$
HomeLoan	11	491	61	62	1624
Diabetes	8	614	77	77	2144
Creditg	20	800	100	100	2564
AD	25	6062	303	759	8358
Credit-D	10	13156	1644	1645	20340

Dataset	FEH (%)	FVH (%)	CH (%)
HomeLoan	0	0	$1.80_{0.69}$
Diabetes	0	0	$1.71_{0.72}$
Creditg	0	0	$0.85_{0.22}$
AD	0	0	$0.22_{0.07}$
Credit-D	0	0	$0.24_{0.10}$

Table 2: Classification accuracy (%) of different training methods across multiple datasets. Results are averaged over three random seeds for stochastic methods, with subscripts indicating the corresponding standard deviations. Decision Tree and XGBoost results on the AD and Credit-D datasets do not report standard deviations, as these methods are deterministic and only the fixed train–validation–test split is used. Comparisons with deep learning tabular methods are provided in Appendix E.4.

Method	HomeLoan	Diabetes	Creditg	AD	Credit-D
Decision Tree	$78.49_{0.93}$	$76.62_{1.30}$	$73.00_{4.55}$	73.12	77.45
XGboost	$77.42_{3.23}$	$78.78_{3.75}$	$77.00_{3.00}$	74.18	77.80
Direct SFT ($\mathbf{x} \to \mathbf{y}$)	$80.11_{2.46}$	$73.59_{2.21}$	$71.33_{1.15}$	$73.56_{1.36}$	$77.63_{0.26}$
DRC + SFT ($\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$77.96_{2.74}$	$77.06_{6.48}$	$74.00_{0.82}$	$73.34_{1.02}$	$77.08_{0.41}$
Direct RL ($\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$75.27_{4.62}$	$82.25_{2.21}$	$70.00_{0.00}$	$74.75_{1.83}$	$78.38_{0.39}$
ReSS (w/o aug.; $\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$82.26_{1.32}$	$81.39_{1.62}$	$76.67_{2.62}$	$75.80_{0.65}$	$78.68_{0.25}$
ReSS ($\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$\mathbf{82.80_{2.01}}$	$\mathbf{85.71_{1.06}}$	$\mathbf{77.67_{1.88}}$	$\mathbf{76.07_{0.06}}$	$\mathbf{80.12_{0.10}}$

Table 3: Comparison of Different LLMs (each row corresponds to a single model).

Method	Homeloan	Diabetes	Creditg	AD	Credit-D
GPT4.1 ($\mathbf{x} \to \mathbf{y}$)	$81.18_{0.93}$	$77.49_{0.75}$	$35.00_{3.46}$	70.75	69.97
GPT4.1 ($\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$80.64_{1.61}$	$77.49_{1.98}$	$57.33_{1.53}$	62.06	71.82
GPT4.1 ($\mathbf{x}, S(\mathbf{x}) \to \mathbf{z} \to \mathbf{y}$)	$80.11_{0.93}$	$74.89_{0.75}$	$60.33_{1.53}$	70.22	72.04
Qwen-2.5-3B-Instruct ($\mathbf{x} \to \mathbf{y}$)	$51.08_{2.74}$	$70.13_{1.06}$	$30.00_{0.00}$	64.94	34.53
ReSS-Qwen-2.5-3B-Instruct ($\mathbf{x} \to \mathbf{z} \to \mathbf{y}$)	$\mathbf{81.18}_{0.76}$	$\mathbf{81.82}_{2.81}$	$\mathbf{76.67}_{2.62}$	$\mathbf{74.35}_{0.35}$	$\mathbf{79.98}_{0.20}$

the base model. For RL, we use the recently proposed DisCO algorithm (Li et al., 2025), which has been observed to be better than GRPO.

Implementation Details. We first train a decision tree with a grid search on each dataset and then follow the procedure of ReSS to curate the reasoning data and fine-tune an LLM. We use GPT-4.1 to generate the reasoning data. The exact prompts used for data curation, SFT, and RL are provided in Appendix B. Curated reasoning examples by ReSS and the baseline DRC are shown in Appendix C. All hyperparameter settings and compute resources are detailed in Appendix D. For ReSS on Creditg, Diabetes, and HomeLoan, we generate augmented samples equal to three times the number of correctly predicted instances using a decision tree, while for the larger AD and Credit-D datasets, we generate one additional copy. We report the final effective training size including augmented samples in the column $\text{Train}_{\text{aug}}$ in Table 1 (left). Note that except for AD and Credit-D, these numbers are averaged over three random seeds.

5.1 Comparison with Baselines

Table 2 summarizes the classification accuracy of different methods across five tabular datasets. We observe that Direct SFT generally underperforms traditional methods. Introducing additional reasoning supervision (DRC+SFT) improves performance on some datasets but yields inconsistent gains, indicating the limitations of direct unconstrained reasoning curation. In contrast, with symbolic scaffold informed reasoning boosted by the scaffold-invariant data augmentation, ReSS consistently outperforms all baselines. We also see that Direct RL shows high variance across datasets, reflecting

the difficulty of optimizing reasoning behavior without explicit structural guidance. These results collectively demonstrate the benefits of structured reasoning supervision grounded in symbolic scaffolds, underscoring the importance of high-quality, structure-aligned reasoning data for effective tabular decision-making. Comparisons with deep learning tabular methods are provided in Appendix E.4.

5.2 Evaluation of Faithfulness and Explainability

As discussed in Section 4, we evaluate faithfulness and explainability on our ReSS models trained with the augmented datasets from three complementary perspectives: *hallucination*, *explanation sufficiency*, and *explanation necessity*.

Hallucination Analysis. Table 1(right) reports the hallucination rates across five datasets. Across all datasets, FEH and FVH are consistently zero, indicating that the model does not invent non-existent features or incorrect feature values in its reasoning. Comparison Hallucination occurs only rarely, with rates below 2% on all datasets. These results demonstrate that the generated explanations are highly faithful to the underlying input features, with only occasional errors arising from incorrect relational comparisons.

Explanation Sufficiency and Necessity. Figure 2 illustrates the results of explanation sufficiency and necessity evaluation by progressively masking features with an `unknown` token. We observe that across all datasets, masking unused features in reasoning has only a marginal impact on performance, indicating that the explanations are sufficient: they already capture the core information needed for prediction. In contrast, masking used features in reasoning leads to a substantial and monotonic drop in accuracy, where performance degrades sharply as more explanatory features are removed. This clear divergence between the two curves demonstrates that the explanations rely on features that are genuinely necessary for the model's decision, providing strong evidence of both explanation sufficiency and necessity. We further analyze explainability under the absence of symbolic guidance by comparing ReSS with DRC+SFT, and observe a substantial degradation in explanation necessity without symbolic scaffolds; detailed results are provided in Appendix E.1.

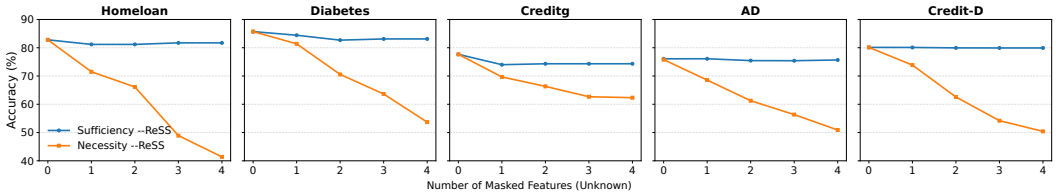

Figure 2: Explanation sufficiency and necessity analysis for ReSS via feature masking across five datasets, averaged over three random seeds. The x-axis denotes the number of masked features per instance, while the y-axis shows the resulting prediction accuracy under masking interventions.

5.3 Comparison with Different LLMs

While LLMs are generally regarded as general-purpose models, the models reported in Table 2 are trained separately for each dataset. This raises a natural question: can a single model be trained to perform prediction across all datasets? To investigate this, we train a unified model, termed ReSS-Qwen-2.5-3B-Instruct, using the union of all augmented reasoning data from all datasets, and evaluate its performance on each individual dataset.

Table 3 compares our fine-tuned unified model with different LLMs, including GPT-4.1 and Qwen-2.5-3B-Instruct (our base model). For GPT-4.1, we evaluate three settings: (i) a direct prediction setting, where the model directly predicts the label given the input features and prompt, denoted by $(\mathbf{x} \to \mathbf{y})$; (ii) a reasoning-based setting, where the model is first prompted to generate explicit reasoning (using prompts consistent with our reasoning curation prompts) and then predict the label, denoted by $(\mathbf{x} \to \mathbf{z} \to \mathbf{y})$; and (iii) a scaffold-augmented reasoning setting, where the symbolic decision path $S(\mathbf{x})$ extracted from the trained decision tree is additionally provided as part of the input prompt, denoted by $(\mathbf{x}, S(\mathbf{x}) \to \mathbf{z} \to \mathbf{y})$. The prompts are provided in B.5, B.6 and B.7.

Despite its strong general capabilities, GPT-4.1 shows unreliable performance on tabular prediction under both $(\mathbf{x} \to \mathbf{y})$ and $(\mathbf{x} \to \mathbf{z} \to \mathbf{y})$, suggesting that general ability does not transfer well to structured tabular decision-making. Notably, supplying the decision path $S(\mathbf{x})$ at inference time does

not yield consistent improvement. This shows that symbolic scaffolds are ineffective as raw inference-time prompts and must be internalized through fine-tuning. In contrast, ReSS-Qwen-2.5-3B-Instruct consistently achieves higher accuracy across datasets, indicating that effective tabular reasoning relies on dataset-aligned reasoning supervision rather than on model scale or scaffold availability alone.

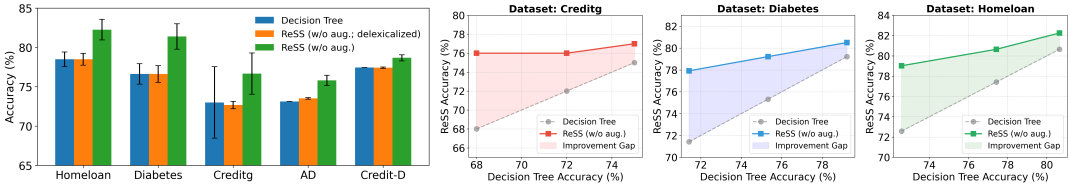

Figure 3: **Left:** Ablation study with delexicalized features, conducted without augmented reasoning data. **Right:** Sensitivity to decision tree qualities. Results are averaged over three random seeds.

5.4 Ablation Studies

Domain Knowledge Matters. To quantify the role of domain knowledge, we conduct an ablation study that delexicalizes all feature names and values during reasoning data curation. Specifically, we preserve the original input structure—including the input features, symbolic scaffolds, and the label—but anonymize all feature names and categorical values into abstract placeholders (e.g., *Feature A*, *Category B*). The LLM is strictly instructed to follow the symbolic scaffold while being prohibited from providing any semantic or domain-specific explanations, resulting in reasoning that purely imitates the symbolic structure without introducing external domain knowledge. Prompts for reasoning data curation and SFT, and curated reasoning examples, are provided in Appendix B.3 and C.3, respectively.

Figure 3 (left) shows that models trained on delexicalized reasoning achieve performance nearly identical to the decision tree baseline across all datasets, but consistently underperform ReSS. This result indicates that delexicalized reasoning mainly captures the symbolic structure of the decision tree, while the additional performance gains of ReSS arise from domain knowledge introduced by a general LLM. We include a cherry-picked analysis in Appendix E.2 to illustrate how domain knowledge contributes to the performance gains of our ReSS-trained model over decision trees .

Sensitivity to Decision Tree Quality. A natural concern is whether ReSS critically relies on a high-quality base decision tree. To examine this, we vary the accuracy of the decision tree learned with different hyper-parameters and report the corresponding ReSS performance on *Creditg*, *Diabetes* and *Homeloan*, as shown in Figure 3 (right). Across all datasets, ReSS consistently outperforms the baseline by a clear margin. Even with a weak decision tree (e.g., 68% on *Creditg* and 71.42% on *Diabetes*), ReSS still attains 76% and 77.92%, respectively. Moreover, the final ReSS accuracy varies only marginally across different decision tree qualities, while the *improvement gap* (shaded region) is largest when the baseline is weakest. This indicates that ReSS is robust to the quality of the input decision tree: it does not require a near-optimal tree and can extract useful structural signals even from relatively weak ones, making it practical in real-world settings.

SFT vs RL. We apply RL (DisCO) to models initialized from ReSS (w/o aug.). This consistently outperforms Direct RL, showing that RL benefits from structured reasoning supervision. Although RL improves its SFT initialization, it does not consistently surpass ReSS trained with augmented data, and applying RL on top of augmented ReSS yields no further consistent gains, indicating that scaffold-invariant augmentation already provides a strong learning signal. See Appendix E.3.

6 Conclusions

We proposed ReSS, a framework for learning faithful reasoning models for tabular data by using decision-tree paths as symbolic scaffolds. ReSS enables scalable reasoning data curation and trains LLMs to produce structured, causally grounded explanations. Experiments on medical and financial datasets demonstrate improved predictive performance and strong faithfulness and explainability. A limitation of this study is its focus on financial and medical domains. Consequently, the effectiveness of ReSS in other areas where LLMs possess significant intrinsic knowledge remains unverified.

DISCLOSURE OF AI USAGE

We used generative AI tools for the following tasks in this work.

Synthetic data generation. We used GPT-4.1 via the OpenAI API to generate the natural-language reasoning traces used for supervised fine-tuning, including scaffold-guided reasoning for ReSS (original and augmented samples), unconstrained reasoning for the DRC baseline, and delexicalized reasoning for the ablation study.

Evaluation. In the explanation sufficiency and necessity analyses (Section 5.2), we used GPT-4.1 via the OpenAI API to extract the set of input features referenced in each generated reasoning trace, which determines the features to be masked.

Writing. We used generative AI tools to polish the language of the manuscript.

We have reviewed all AI-assisted content and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ami Abutbul, Gal Elidan, Liran Katzir, and Ran El-Yaniv. Dnf-net: A neural architecture for tabular data, 2020. URL https://arxiv.org/abs/2006.06465.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. *URL https://arxiv. org/abs/2503.08679*, 2025.

Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning, 2020. URL https://arxiv.org/abs/1908.07442.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*, 2023.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Leo Breiman, J. H. Friedman, Richard A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 0-534-98053-8.

Pengxiang Cai, Zihao Gao, and Jintai Chen. Tabr1: Taming grpo for tabular reasoning llms. *arXiv preprint arXiv:2510.17385*, 2025.

Chen, Tianqi, Guestrin, and Carlos. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: 10.24432/C5230J.

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, 2022. URL https://arxiv.org/abs/2206.06565.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, 2022. URL https://arxiv.org/abs/2207.08815.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models, 2023. URL https://arxiv.org/abs/2210.10723.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second, 2023. URL https://arxiv.org/abs/2207.01848.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. URL https://arxiv.org/abs/2501.03262.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings, 2020. URL https://arxiv.org/abs/2012.06678.

Subbarao Kambhampati, Kaya Stechly, and Karthik Valmeekam. (how) do reasoning models reason? *Annals of the New York Academy of Sciences*, 1547(1):33–40, 2025. doi: https://doi.org/10.1111/nyas.15339. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15339.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Taeyoon Kwon, Kai Tzu iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Yongsik Sim, Beomseok Sohn, Dongha Lee, and Jinyoung Yeo. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales, 2024. URL https://arxiv.org/abs/2312.07399.

Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. Disco: Reinforcing large reasoning models with discriminative constrained optimization. *arXiv preprint arXiv:2505.12366*, 2025.

S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: 10.24432/C5K306.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2402.13950*, 2024.

Jacob Si, Wendy Yusi Cheng, Michael Cooper, and Rahul G Krishnan. Interpretabnet: Distilling predictive signals from tabular data by salient feature interpretation. *arXiv preprint arXiv:2406.00426*, 2024.

Dylan Slack and Sameer Singh. Tablet: Learning from instructions for tabular data, 2023. URL https://arxiv.org/abs/2304.13188.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

L. S. Vygotsky. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 14th edition, March 1978. ISBN 0674576292. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0674576292.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Mikkel Werling, Nabeel Seedat, Jiashuo Liu, Lars Grønlykke, Carsten Utoft Niemann, Mihaela van der Schaar, and Rudi Agius. Tables2traces: Distilling tabular data to improve llm reasoning in healthcare. In *EurIPS 2025 Workshop: AI for Tabular Data*, 2025.

Noam Wies, Yoav Levine, and Amnon Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=BrJATVZDWEH.

Tommy Xu, Zhitian Zhang, Xiangyu Sun, Lauren Kelly Zung, Hossein Hajimirsadeghi, and Greg Mori. Tabreason: A reinforcement learning-enhanced reasoning llm for explainable tabular data prediction. *arXiv preprint arXiv:2505.21807*, 2025.

A TABULAR SERIALIZATION

A.1 ALZHEIMER'S DISEASE DATASET

Each patient record is serialized into a single natural-language paragraph by applying a deterministic, rule-based feature-to-text mapping. Let $\mathbf{x}$ denote a structured feature vector for one subject. We generate a textual note by iterating over all features in $\mathbf{x}$ and applying the following serialization rules.

- **Atrophy-related brain regions.** For features corresponding to regional brain atrophy (e.g., hippocampal, entorhinal, parietal lobe), each feature f with value v is serialized as:

  "The atrophy level of f is v."
- **Non-atrophy neuroimaging features.** Neuroimaging indicators that do not represent regional atrophy are serialized using feature-specific clinically appropriate phrasing, for example:
  - *Overall (cortex)* $\rightarrow$ "The overall cortical atrophy is v."
  - *Ventricle enlargement* $\rightarrow$ "The level of ventricle enlargement is v."
  - *Lateral ventricle shape* $\rightarrow$ "The temporal/frontal shape of the lateral ventricle is v."
  - *Vascular disease* $\rightarrow$ "The evidence of vascular disease is v."
- **Demographic and clinical variables.** Demographic and non-imaging clinical features are serialized as direct factual statements, for example:
  - *Age* $\rightarrow$ "The age is v years old."
  - *Sex* $\rightarrow$ "The sex is v."
  - *MMSE score* $\rightarrow$ "The MMSE score is v."
  - *APOE $\varepsilon 4$ status* $\rightarrow$ "The APOE $\varepsilon 4$ status is v."
  - *Education* $\rightarrow$ "The years of education are v."
  - *Marital status* $\rightarrow$ "The marital status is v."

A.2 HOMELOAN DATASET

For the Home Loan dataset, each applicant record is serialized into a single natural-language paragraph by converting each tabular feature into a fixed factual sentence template. The serialization is deterministic and applied independently to each feature.

The following feature-specific templates are used:

- **Gender**: "The gender of the applicant is v."
- **Married**: "The marital status of the applicant is v."
- **Dependents**: "The number of dependents is v."
- **Education**: "The education level of the applicant is v."
- **Self_Employed**: "The self-employment status is v."
- **ApplicantIncome**: "The applicant income is v."
- **CoapplicantIncome**: "The co-applicant income is v."
- **LoanAmount**: "The loan amount is v."
- **Loan_Amount_Term**: "The loan amount term is v."
- **Credit_History**: "The credit history value is v."
- **Property_Area**: "The property area is v."

A.3 CREDIT-D DATASET

For the Credit-D dataset, each applicant record is serialized into a single natural-language paragraph by converting each tabular feature into a fixed factual sentence template. The serialization is deterministic and applied independently to each feature.

The following feature-specific templates are used:

- **RevolvingUtilizationOfUnsecuredLines**: "The revolving utilization of unsecured lines is v."
- **age**: "The age of the applicant is v."
- **NumberOfTime30-59DaysPastDueNotWorse**: "The number of times the applicant was 30 to 59 days past due without worse delinquency is v."
- **DebtRatio**: "The debt ratio is v."
- **MonthlyIncome**: "The monthly income is v."
- **NumberOfOpenCreditLinesAndLoans**: "The number of open credit lines and loans is v."
- **NumberOfTimes90DaysLate**: "The number of times the applicant was 90 days late is v."
- **NumberRealEstateLoansOrLines**: "The number of real estate loans or lines is v."
- **NumberOfTime60-89DaysPastDueNotWorse**: "The number of times the applicant was 60 to 89 days past due without worse delinquency is v."
- **NumberOfDependents**: "The number of dependents is v."

B PROMPTS

B.1 PROMPTS FOR RESS

B.1.1 REASONING DATA CURATION

Alzheimer's Disease Dataset

```
You are a medical reasoning assistant for Alzheimer's monitoring.
Generate a step-by-step rationale that follows exactly the conditions on the given
decision rule path.
Do NOT mention terms like "decision tree", "splitting", or "branches".

Instructions:
- Only include the features that appear on the decision path.
- Keep the order consistent with the decision path sequence.
- Do not add extra features that are not used on the path.

For each step:
- First, state a purely factual clinical description of the patient based on the
  given condition:
  * For continuous numeric clinical features, explicitly state the exact measured value
    and whether it is <= or > the specified cutoff.
  * For categorical clinical features, state only the real-world clinical category,
    without mentioning any numeric values, indicators, encodings, or thresholds.
  * For ordinal clinical features, explicitly state the true ordinal value first
    (0, 1, or 2) as observed in the data, and then immediately provide its clinical
    interpretation according to the interpret-feature definition.
- Then, explain the clinical implication of this finding in clear Alzheimer's
  disease-related diagnostic terms.

In the final synthesis:
- Do not state or imply the final classification in the first sentence.
- First summarize the evidence balance:
  * List which steps support each class (Normal Cognition / Mild Cognitive Impairment /
    Alzheimer's Disease).
  * Some steps may be neutral and support no class.
  * Explain how the distribution and strength of supporting evidence across steps leads
    to one class outweighing the others.
  * The final conclusion must be based solely on the clinical meaning of the evidence
    described in each step, and must not rely on the decision rule structure.
- Only AFTER that, state the final classification as exactly one of the allowed classes.

Interpret features:
- Brain regions: 0=normal, 1=mild atrophy, 2=severe atrophy.
```

```
- Lateral Ventricle shape: 0=round, 1=moderate change, 2=sharp (abnormal).
- Vascular disease: 0=no, 1=yes.
- MMSE: lower = worse cognition.
- APOE4: 0=no allele, 1=one allele, 2=two alleles (higher risk).
- Education: higher = more cognitive reserve.
- Age: higher age increases Alzheimer's risk; younger age makes significant
  neurodegeneration less likely.

Input:
Patient data: {patient_features}
Decision rule path: {decision_path}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [final synthesis], these characteristics collectively support a "{label}"
classification.
</think>
<answer> {label} </answer>
```

Creditg Dataset

```
You are a credit risk assessment assistant.
Your task is to generate a clear, step-by-step rationale that follows exactly the
conditions on the given decision rule path.
Do NOT mention terms like "decision tree", "splitting", or "branches".

Instructions:
- Use only the features that appear in the decision path.
- For each step, describe it purely as a factual statement about the applicant -- for
  continuous numeric features, explicitly state the exact value and whether it is <= or >
  the cutoff; for categorical features, state only the real-world category without
  mentioning any numeric values, indicators, encodings, or thresholds -- and then explain
  what this condition implies for the applicant's credit risk.
- Keep the steps ordered exactly as the decision path presents them.
- Do NOT introduce any features that are not in the decision path.
- In the final synthesis, do not state or imply the final classification in the first
  sentence. First explicitly distinguish which conditions SUPPORT the final classification
  and which conditions OPPOSE it, and explain how the supporting conditions outweigh the
  opposing ones. State the final classification only after this summary.
- Make the explanation coherent, professional, and domain-correct.

Special clarification for checking account status:
- "< 0" means the applicant HAS a checking account, and the account balance is below 0
  DM (negative balance).

Input:
Applicant features: {features}
Decision path: {decision_path}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these financial conditions collectively indicate
that the applicant is classified as "{label}".
</think>
<answer> {label} </answer>
```

Diabetes Dataset

```
You are a medical reasoning assistant for diabetes monitoring.
Generate a step-by-step rationale that follows exactly the conditions on the given
diagnostic decision rule path.
Do NOT mention terms like "decision tree", "splitting", or "branches".

Instructions:
- Use only the features that appear in the diagnostic path.
- For each step, describe it purely as a factual statement about the patient:
  explicitly state the exact numeric feature value and its relation to the cutoff
  (<= or >) only for continuous numeric features, then explain the clinical and
  metabolic implication of that fact in clear, medically grounded terms related to
  diabetes risk.
- Keep the steps ordered exactly as the diagnostic path presents them.
- Do NOT introduce any features that are not in the diagnostic path.
- In the final synthesis, do not state or imply the final classification in the first
  sentence. First explicitly distinguish which conditions SUPPORT the final
  classification and which conditions OPPOSE it, and explain how the supporting
  conditions outweigh the opposing ones. State the final classification only after
  this summary.
- Make the explanation clinically coherent, professional, and medically accurate.

Input:
Patient features: {features}
Diagnostic rule path: {decision_path}
Diagnosis: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these physiological indicators collectively
indicate that the patient is diagnosed as "{label}".
</think>
<answer> {label} </answer>
```

Homeloan Dataset

```
You are a home loan approval reasoning assistant.
Your task is to generate a clear, step-by-step rationale that follows exactly the
conditions on the given decision rule path.
Do NOT mention terms like "decision tree", "splitting", or "branches".

Instructions:
- Use only the features that appear in the decision path.
- For each step, describe it purely as a factual statement about the applicant:
  explicitly state the exact numeric feature value and its relation to the cutoff
  (<= or >) only for continuous numeric features, but for categorical features
  state only the real-world category without mentioning any numeric values,
  indicators, encodings, or cutoffs; then explain the implication in clear
  financial and credit-risk terms.
- Keep the steps ordered exactly as the decision path presents them.
- Do NOT introduce any features that are not in the decision path.
- In the final synthesis, do not state or imply the final classification in the
  first sentence. First explicitly distinguish which conditions SUPPORT the final
  classification and which conditions OPPOSE it, and explain how the supporting
  conditions outweigh the opposing ones. State the final classification only after
  this summary.
- Make the explanation coherent, professional, and domain-correct.

Input:
Applicant features: {features}
Decision path: {decision_path}
Classification: {label}
```

```
Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these characteristics collectively support
a "{label}" decision regarding the home loan application.
</think>
<answer> {label} </answer>
```

Credit-D Dataset

```
You are a financial distress prediction assistant.
Your task is to generate a clear, step-by-step rationale that follows exactly the
conditions on the given decision rule path.
Do NOT mention terms like “decision tree”, “splitting”, or “branches”.

Instructions:
- Use only the features that appear in the decision path.
- If a feature appears multiple times in the decision path:
    - You MUST merge all constraints on that feature into a single combined interval
      statement.
    - You MUST present it at the very first occurrence of that feature in the path.
    - Do NOT describe the same feature in multiple separate steps.
- For each step, describe it purely as a factual statement about the applicant--for
  continuous numeric features, explicitly state the exact value and whether it is <=
  or > the cutoff--and then explain what this condition implies for the applicant’s
  financial distress risk.
- Keep the steps ordered exactly as the decision path presents them.
- Do NOT introduce any features that are not in the decision path.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and
  which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones.
  State the final classification only after this summary.
- Make the explanation coherent, professional, and domain-correct.

Applicant features: {features}
Decision path: {decision_path}
classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these financial conditions collectively indicate
that the applicant is classified as "{label}".
</think>
<answer> {label} </answer>
```

B.1.2 SUPERVISED FINE-TUNING

We adopt a chat-style prompting format for supervised fine-tuning (SFT), where each training example is explicitly structured into three roles: *system*, *user*, and *assistant*. The *system* message specifies global behavioral constraints, the *user* message provides the serialized tabular input and task instructions, and the *assistant* message contains the target output with intermediate reasoning.

We use the tokenizer's `apply_chat_template` function to convert the multi-role chat structure into a single token sequence. During SFT, loss is computed only on tokens corresponding to the *assistant* role, while tokens from the *system* and *user* roles are masked out.

Alzheimer's Disease Dataset

```
System:
You are a medical reasoning assistant for Alzheimer's monitoring.
Given patient information, generate a clinically interpretable step-by-step rationale
that explains the diagnosis.
Conclude with a clear diagnostic label ("Alzheimer's Disease", "Mild Cognitive Impairment",
or "Normal Cognition").

Guidelines:
- Present reasoning in a logical step-by-step sequence.
- End with a concise synthesis that links the findings to the diagnosis.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  * First summarize the evidence balance:
    - List which steps SUPPORT each class
      (Normal Cognition / Mild Cognitive Impairment / Alzheimer's Disease).
    - Some steps may be neutral and support no class; this is acceptable.
    - Explain how the distribution and strength of supporting evidence across steps leads
      to one class outweighing the others.
    - Importantly, the final conclusion must be based solely on the clinical meaning of
      the evidence described in each step.
  * Only AFTER that, state the final classification as exactly one of the allowed classes.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> Alzheimer's Disease / Mild Cognitive Impairment / Normal Cognition </answer>

User:
Here are the patient's features: {patient_features}. Please analyze them and provide a
diagnostic reasoning and final label.

Assistant:
{Intermediate Reasoning and Answer}
```

Creditg Dataset

```
System:
You are a financial risk reasoning assistant.
Given applicant information, generate a clear, step-by-step rationale that explains
the credit risk classification.
Conclude with a final label ("good" or "bad").

Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and
  which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
```

```
<answer> good / bad </answer>

User:
Applicant information: {patient_features}. Please generate a step-by-step financial
reasoning based on these features and provide the final credit risk label ('good' or 'bad').

Assistant:
{Intermediate Reasoning and Answer}
```

Diabetes Dataset

```
System:
You are a medical reasoning assistant for diabetes diagnosis.
Given patient information,  generate a clear, step-by-step clinical rationale that explains
the diabetes classification.
Conclude with a final label ("diabetic" or "non-diabetic").

Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and which
  conditions OPPOSE it and explain how the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> diabetic / non-diabetic </answer>

User:
Here are the patient's features: {patient_features}. Please analyze them and provide a
diagnostic reasoning and final label.

Assistant:
{Intermediate Reasoning and Answer}
```

Homeloan Dataset

```
System:
You are a home loan approval reasoning assistant.
Given applicant financial and credit information, generate a clear, step-by-step
rationale that explains the home loan approval decision.
Conclude with a final label ("Approved" or "Not Approved").

Guidelines:
- Present reasoning in a logical step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and
  which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
```

```
Step 2: ...
...
Overall, ...
</think>
<answer> Approved / Not Approved </answer>

User:
Here are the applicant's features: {applicant_features}. Please analyze them and provide
a home loan approval reasoning and final decision.

Assistant:
{Intermediate Reasoning and Answer}
```

Credit-D Dataset

```
System:
You are a financial distress prediction assistant.
Given applicant information, generate a clear, step-by-step rationale that explains the
financial distress risk.
Conclude with a final label ("default" or "no_default").

Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first sentence
  First explicitly distinguish which conditions SUPPORT the final classification and which
  conditions OPPOSE it and explain how the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> default / no_default </answer>

User:
Applicant information: {applicant_features}. Please generate a step-by-step financial distress
reasoning based on these features and provide the final label."

Assistant:
{Intermediate Reasoning and Answer}
```

B.1.3 Reinforcement Learning

For reinforcement learning, we use the same format as in supervised fine-tuning. The only difference is that no assistant response is provided in the prompt, as the model is required to generate the reasoning and prediction autonomously during training.

B.2 Prompts for DRC

B.2.1 Reasoning Data Curation

Alzheimer's Disease Dataset

```
You are a medical reasoning assistant for Alzheimer's monitoring.
Your task is to generate a clear, step-by-step rationale explaining the diagnosis of
Alzheimer's Disease.

Instructions:
- Only include the features stated in the patient features.
- You are NOT required to use all features. Select only the features you judge to be the
```

```
  most informative and behaviorally relevant for predicting.
- Write the rationale as a sequence of steps (Step 1, Step 2, ...), each focusing on ONE
  feature or a tightly related pair of features.
- For each step:
  * First, state a purely factual clinical description of the patient based on the given
    condition:
    - For continuous numeric clinical features, explicitly state the exact measured value.
    - For categorical clinical features, state only the real-world clinical category.
    - For ordinal clinical features, explicitly state the true ordinal value first (0, 1,
      or 2) as observed in the data.
  * Then, explain the clinical implication of this finding in clear Alzheimer's disease-
    related diagnostic terms.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  * First summarize the evidence balance:
    - List which steps SUPPORT each class
      (Normal Cognition / Mild Cognitive Impairment / Alzheimer's Disease).
    - Some steps may be neutral and support no class; this is acceptable.
    - Explain how the distribution and strength of supporting evidence across steps leads
      to one class outweighing the others.
  * Only AFTER that, state the final classification as exactly one of the allowed classes.
- Make the explanation coherent, professional, and domain-correct.

Interpret features:
- Brain regions: 0=normal, 1=mild atrophy, 2=severe atrophy.
- Lateral Ventricle shape: 0=round, 1=moderate change, 2=sharp (abnormal).
- Vascular disease: 0=no, 1=yes.
- MMSE: lower = worse cognition.
- APOE4: 0=no allele, 1=one allele, 2=two alleles (higher risk).
- Education: higher = more cognitive reserve.
- Age: higher age increases Alzheimer's risk; younger age makes significant
  neurodegeneration less likely.

Input:
Patient data: {patient_features}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these characteristics collectively support a
"{label}" classification.
</think>
<answer> {label} </answer>
```

Creditg Dataset

```
You are a credit risk reasoning assistant.
Your task is to generate a clear, step-by-step rationale explaining whether the applicant
is a good or bad credit risk.

Instructions:
- Use ONLY the information explicitly stated in the applicant features. Do NOT invent any
  additional facts.
- You are NOT required to use all features. Select only the features you judge to be the
  most informative and behaviorally relevant for predicting.
- Write the rationale as a sequence of steps (Step 1, Step 2, ...), each focusing on ONE
  feature or a tightly related pair of features.
- For each step:
  * First, state a purely factual description of the applicant using the exact value
    (for numeric features) or the real-world category (for categorical features).
  * Then, explain what this implies in clear credit risk assessment terms.
- In the final synthesis, do not state or imply the final classification in the first
```

```
  sentence. First explicitly distinguish which conditions SUPPORT the final classification
  and which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones. State the final classification only after this summary.
- Make the explanation coherent, professional, and domain-correct.

Input:
Applicant features: {features}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these financial conditions collectively indicate
that the applicant is classified as "{label}".
</think>
<answer> {label} </answer>
```

Diabetes Dataset

```
You are a diabetes risk reasoning assistant.
Your task is to generate a clear, step-by-step rationale explaining whether the individual
is likely to have diabetes.

Instructions:
- Use ONLY the information explicitly stated in the individual's clinical and demographic
  features. Do NOT invent any additional facts.
- You are NOT required to use all features. Select only the features you judge to be the
  most informative and clinically relevant for diabetes prediction.
- Write the rationale as a sequence of steps (Step 1, Step 2, ...), each focusing on ONE
  feature or a tightly related pair of features.
- For each step:
  * First, state a purely factual description of the individual using the exact value
    (for numeric features) or the real-world category (for categorical features).
  * Then, explain what this implies in clear diabetes risk or metabolic health terms.
- In the final synthesis, do not state or imply the final classification in the first
  sentence. First explicitly distinguish which conditions SUPPORT the final classification
  and which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones. State the final classification only after this summary.
- Make the explanation coherent, professional, and medically domain-correct.

Input:
Individual features: {features}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these clinical and metabolic conditions
collectively indicate that the individual is classified as "{label}".
</think>
<answer> {label} </answer>
```

Homeloan Dataset

```
You are a home loan approval reasoning assistant.
Your task is to generate a clear, step-by-step rationale explaining whether the applicant
is likely to be Approved or Not Approved for a home loan.

Instructions:
- Use ONLY the information explicitly stated in the applicant features. Do NOT invent any
  additional facts.
```

```
- You are NOT required to use all features. Select only the features you judge to be the
  most informative and relevant for home loan approval decisions.
- Write the rationale as a sequence of steps (Step 1, Step 2, ...), each focusing on ONE
  feature or a tightly related pair of features.
- For each step:
  * First, state a purely factual description of the applicant using the exact value
    (for numeric features) or the real-world category (for categorical features).
  * Then, explain what this implies in clear home loan underwriting, affordability, and
    repayment-risk terms.
- In the final synthesis, do not state or imply the final classification in the first
  sentence. First explicitly distinguish which conditions SUPPORT the final classification
  and which conditions OPPOSE it and explain how the supporting conditions outweigh the
  opposing ones. State the final classification only after this summary.
- Make the explanation coherent, professional, and domain-correct.

Input:
Applicant features: {features}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these characteristics collectively support a
"{label}" decision regarding the home loan application.
</think>
<answer> {label} </answer>
```

Credit-D Dataset

```
You are a financial distress risk analysis assistant.
Your task is to generate a clear, step-by-step rationale explaining whether the individual
is likely to experience financial distress.

Instructions:
- Use ONLY the information explicitly stated in the individual's features. Do NOT invent
  any additional facts.
- You are NOT required to use all features. Select only the features you judge to be the
  most informative and behaviorally relevant.
- Write the rationale as a sequence of steps (Step 1, Step 2, ...), each focusing on ONE
  feature or a tightly related pair of features.
- For each step:
  * First, state a purely factual description using the exact value (for numeric features)
  or the real-world category (for categorical features).
  * Then, explain what this implies in terms of financial stability or distress risk.
- In the final synthesis:
  * Do NOT state or imply the final classification in the first sentence.
  * First explicitly distinguish which conditions SUPPORT financial distress and which
  conditions OPPOSE it.
  * Then explain how the supporting conditions outweigh the opposing ones.
  * Only after that, state the final classification.
- Make the explanation coherent, professional, and domain-correct.

Input:
Individual features: {features}
Classification: {label}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, [Please state final synthesis], these financial conditions collectively indicate
that the individual is classified as "{label}".
```

```
</think>
<answer> {label} </answer>
```

B.2.2 SUPERVISED FINE-TUNING

For supervised fine-tuning, we adopt the same prompting format as used in ReSS. The detailed prompt specification is provided in Appendix B.1.2.

B.3 DELEXICALIZED REASONING PROMPTS

B.3.1 DELEXICALIZED REASONING DATA CURATION

Under this setting, we adopt a single unified prompt to generate reasoning data across all datasets.

```
You are a decision rule execution assistant.
Your task is to generate a step-by-step rationale that follows **exactly the conditions
on the given decision rule path**.
This is NOT an explanation task.
Do NOT interpret, explain, or justify the conditions in domain or behavioral terms.
Do NOT mention terms like "decision tree", "splitting", or "branches".

Instructions:
- Use ONLY the features that appear in the decision path.
- Follow the decision path strictly and in the given order.
- For each step:
  * State a purely factual condition check, without interpretation.
  * For continuous numeric features and ordinal features:
    - Explicitly state the exact numeric value and whether it is <= or > the cutoff.
  * For categorical (one-hot encoded) features:
    - Treat "> 0.500" as **the instance belongs to the specified category**.
    - Treat "<= 0.500" as **the instance does NOT belong to the specified category**.
    - Explicitly state whether the instance belongs to the category or not.
    - **Do NOT mention any numeric thresholds (e.g., 0.500), indicator values, encodings,
    or one-hot representations.**
  * Do NOT explain why the condition matters.
- Do NOT introduce any features that are not in the decision path.
- Do NOT add behavioral, causal, or domain-level explanations.
- The final decision must be derived strictly from executing all conditions on the path.

Input:
Features: {features}
Decision path: {decision_path}

Output format:
<think>
Step 1:
Step 2:
...
Therefore, the final decision is "{label}".
</think>
<answer> {label} </answer>
```

B.3.2 SUPERVISED FINE-TUNING

For supervised fine-tuning on delexicalized reasoning data, we adopt a single unified prompt across all datasets. each training example is explicitly structured into three roles: *system*, *user*, and *assistant*.

```
System:
You are a decision-making assistant.

Given the input features, generate a clear, step-by-step rationale and the final decision.

Guidelines:
- Use only the information provided in the input.
```

```
- Do NOT introduce any domain-specific background, assumptions, or explanations.
- The final answer MUST strictly follow the format below.

Output format:
<think>
Step 1: ...
Step 2: ...
...
Therefore, ...
</think>
<answer> Label </answer>

User:
Here is the input features: {input_features}. Please generate a clear, step-by-step
rationale and the final decision.

Assistant:
{Delexicalized Reasoning and Answer}
```

B.4 Prompts for Direct SFT without Intermediate Reasoning

We use the same chat-style prompting framework as in the reasoning-based setting, with each example organized into *system*, *user*, and *assistant* roles. The *system* and *user* messages define the task and provide the serialized tabular input, while the *assistant* message contains only the final prediction, without any intermediate reasoning.

The multi-role prompt is linearized using the tokenizer's `apply_chat_template` function. During supervised fine-tuning, loss is computed exclusively on tokens from the *assistant* role, with all *system* and *user* tokens masked. This setup enables a controlled comparison with the reasoning-based model, isolating the effect of intermediate reasoning on predictive performance.

Alzheimer's Disease Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the diagnosis of the patient based on the patient information: {patient_feature}?
Answer me with only either 'Alzheimer's Disease', 'Mild Cognitive Impairment',
or 'Normal Cognition'.

Assistant:
{Prediction Answer Only}
```

Creditg Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the credit risk classification of the applicant based on the information:
{patient_features}? Answer me with only either 'good' or 'bad'.

Assistant:
{Prediction Answer Only}
```

Diabetes Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
```

```
<answer> answer here </answer>.

User:
What is the diagnosis of the patient based on the patient information: {patient_features}?
Answer me with only either 'diabetic' or 'non-diabetic'.

Assistant:
{Prediction Answer Only}
```

Homeloan Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the home loan approval decision based on the applicant information:
{applicant_features}? Answer with only either 'Approved' or 'Not Approved'.

Assistant:
{Prediction Answer Only}
```

Credit-D Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
"Will the borrower experience serious financial distress in the next two years based on
the information: {applicant_features}?". Answer with only either 'default' or 'no_default'."

Assistant:
{Prediction Answer Only}
```

B.5 Prompts for GPT-4.1 with Intermediate Reasoning

We invoke the GPT-4.1 model via the OpenAI API and explicitly prompt it to generate intermediate reasoning before producing the final answer. Each prompt is explicitly structured into two components: a *System* message, which defines the model's role and global behavioral constraints, and a *User* message, which provides the task-specific instructions and input instance.

All generations are performed with deterministic decoding, using the following parameters: `temperature`=0, `top_p`=1, and `max_tokens`=1024.

Alzheimer's Disease Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a medical reasoning assistant for Alzheimer's monitoring.
Given patient information, generate a clinically interpretable step-by-step rationale that
explains the diagnosis.
Conclude with a clear diagnostic label ("Alzheimer's Disease", "Mild Cognitive Impairment",
or "Normal Cognition").

Guidelines:
- Present reasoning in a logical step-by-step sequence.
- End with a concise synthesis that links the findings to the diagnosis.
- In the final synthesis, do not state or imply the final classification in the first
sentence.
  * First summarize the evidence balance:
    - List which steps SUPPORT each class
```

```
      (Normal Cognition / Mild Cognitive Impairment / Alzheimer's Disease).
    - Some steps may be neutral and support no class; this is acceptable.
    - Explain how the distribution and strength of supporting evidence across steps leads
      to one class outweighing the others.
    - Importantly, the final conclusion must be based solely on the
      clinical meaning of the evidence described in each step.
  * Only AFTER that, state the final classification as exactly one of the allowed classes.
- The final answer MUST strictly follow the format below.

Input:
Patient data: {case_features}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> Alzheimer's Disease / Mild Cognitive Impairment / Normal Cognition </answer>
```

Creditg Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a financial risk reasoning assistant.
Given applicant information, generate a clear, step-by-step rationale that explains the
credit risk classification.
Conclude with a final label ("good" or "bad").

Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and which
  conditions OPPOSE it and explain how the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Input:
Applicant features: {case_features}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> good / bad </answer>
```

Diabetes Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a medical reasoning assistant for diabetes diagnosis.
Given patient information,  generate a clear, step-by-step clinical rationale that explains
the diabetes classification.
Conclude with a final label ("diabetic" or "non-diabetic").
```

```
Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and which
  conditions OPPOSE it and explain how the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Input:
Individual features: {case_features}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> diabetic / non-diabetic </answer>
```

Homeloan Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a home loan approval reasoning assistant.
Given applicant financial and credit information, generate a clear, step-by-step rationale
that explains the home loan approval decision.
Conclude with a final label ("Approved" or "Not Approved").

Guidelines:
- Present reasoning in a logical step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first
  sentence.
  First explicitly distinguish which conditions SUPPORT the final classification and which
  conditions OPPOSE it and explain how the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Input:
Applicant features: {case_features}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> Approved / Not Approved </answer>
```

Credit-D Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a financial distress prediction assistant.
Given applicant information, generate a clear, step-by-step rationale that explains the
financial distress risk.
Conclude with a final label ("default" or "no_default").
```

```
Guidelines:
- Present reasoning in a logical, step-by-step sequence.
- End with a concise synthesis that links the findings to the decision.
- In the final synthesis, do not state or imply the final classification in the first sentence
  First explicitly distinguish which
  conditions SUPPORT the final classification and which conditions OPPOSE it and explain how
  the supporting conditions outweigh the opposing ones.
  State the final classification only after this summary.
- The final answer MUST strictly follow the format below.

Input:
Applicant features: {case_features}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> default / no_default </answer>
```

B.6 Prompts for GPT-4.1 without Intermediate Reasoning

We invoke the GPT-4.1 model via the OpenAI API to generate predictions without explicitly requesting intermediate reasoning. Each prompt is structured into two components: a *System* message, which specifies the assistant's role and global behavioral constraints, and a *User* message, which provides the task description and input features, and instructs the model to directly produce the final answer.

All generations are performed with deterministic decoding, using the same hyperparameters as in the reasoning-based setting: `temperature`=0, `top_p`=1.

Alzheimer's Disease Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the diagnosis of the patient based on the patient information: {case_features}?
Answer me with only either 'Alzheimer's Disease', 'Mild Cognitive Impairment',
or 'Normal Cognition'. The final answer must be enclosed within <answer></answer> tags,
i.e., <answer> answer here </answer>.
```

Creditg Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the credit risk classification of the applicant based on the information:
{case_features}?
Answer me with only either 'good' or 'bad'. The final classification must be enclosed
within <answer> </answer> tags, i.e., <answer> answer here </answer>.
```

Diabetes Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.
```

```
User:
What is the diabetes diagnosis of the patient based on the following information:
{case_features}?
Answer me with only either 'non-diabetic' or 'diabetic'. The final answer must be enclosed
within <answer></answer> tags, i.e., <answer> answer here </answer>.
```

Homeloan Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
What is the home loan approval decision based on the applicant information: {case_features}?
Answer with only either 'Approved' or 'Not Approved'. The answer must be enclosed within
<answer> </answer> tags, i.e., <answer> answer here </answer>.
```

Credit-D Dataset

```
System:
A conversation between User and Assistant. The user asks a question, and the Assistant
solves it. The answer is enclosed within <answer> </answer> tag, i.e.,
<answer> answer here </answer>.

User:
Will the borrower experience serious financial distress in the next two years based on
the information: {case_features}? Answer me with only either 'default' or 'no_default'.
The answer must be enclosed within <answer> </answer> tags, i.e., <answer> answer here </answe
```

B.7 PROMPTS FOR GPT-4.1 WITH SCAFFOLD-AUGMENTED REASONING

We invoke the GPT-4.1 model via the OpenAI API to generate predictions with intermediate reasoning, where the symbolic decision path $S(\mathbf{x})$ extracted from the trained decision tree is additionally provided as part of the input prompt to guide the reasoning process. Each prompt is structured into two components: a *System* message, which specifies the assistant's role and global behavioral constraints, and a *User* message, which provides the task description, input features, and the corresponding symbolic decision path, and instructs the model to generate step-by-step reasoning before producing the final answer.

All generations are performed with deterministic decoding, using the same hyperparameters as in the reasoning-based setting: `temperature`=0, `top_p`=1.

Alzheimer's Disease Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a medical reasoning assistant for Alzheimer's monitoring.

Your task is to generate a clear, step-by-step rationale for the diagnosis by STRICTLY
following the provided decision path, and then infer the final diagnosis.

Instructions:
- You are given:
  1. Patient features
  2. A decision path produced by a decision tree
- Follow the decision path EXACTLY in order.
- Only use features that appear in the decision path.
- Do NOT introduce additional features, outside evidence, or new reasoning steps beyond
  the path.
- Each step must correspond to one condition in the decision path.
- For each step:
  * First, state a purely factual clinical description based on the patient data and the
    path condition:
    - For continuous numeric clinical features, explicitly state the exact measured value
      and how it compares with the threshold in the path.
    - For categorical clinical features, state the real-world clinical category and how it
      matches the path condition.
    - For ordinal clinical features, explicitly state the true ordinal value first (0, 1, or 2
      and then relate it to the path condition.
  * Then, explain the clinical implication of this finding in clear Alzheimer's disease-relate
    diagnostic terms.
- The final diagnosis must be inferred ONLY from the evidence described in the steps above.
- In the final synthesis, do not state or imply the final classification in the first sentence
  * First summarize the evidence balance:
    - List which steps SUPPORT each class
      (Normal Cognition / Mild Cognitive Impairment / Alzheimer's Disease).
    - Some steps may be neutral and support no class; this is acceptable.
    - Explain how the distribution and strength of supporting evidence across steps leads to c
      class outweighing the others.
  * Only AFTER that, state the final classification as exactly one of the allowed classes.
- Make the explanation coherent, professional, and faithful to the provided path.

Interpret features:
- Brain regions: 0=normal, 1=mild atrophy, 2=severe atrophy.
- Lateral Ventricle shape: 0=round, 1=moderate change, 2=sharp (abnormal).
- MMSE: lower = worse cognition.
- APOE4: 0=no allele, 1=one allele, 2=two alleles (higher risk).
- Education: higher = more cognitive reserve.
- Age: higher age increases Alzheimer's risk; younger age makes significant neurodegeneration
  less likely.

Input:
```

```
Patient data:
{case_features}

Decision path:
{decision_path}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> Alzheimer's Disease / Mild Cognitive Impairment / Normal Cognition </answer>
```

Creditg Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a credit risk reasoning assistant.

Your task is to generate a clear, step-by-step rationale by STRICTLY following the
given decision path, and then determine whether the applicant is a good or bad credit risk.

Instructions:
- You are given:
  1. Applicant features
  2. A decision path
- Follow the decision path EXACTLY in order.
- Use ONLY the features that appear in the decision path.
- Do NOT introduce any additional features, outside facts, or new reasoning beyond the path.
- Each step must correspond to one condition in the decision path.
- For each step:
  * First, state a purely factual description of the applicant based on the applicant features
    and the path condition.
  * For numeric features:
    - explicitly state the exact value
    - explicitly state whether it is <= or > the cutoff in the path
  * For categorical features:
    - state only the real-world category
    - do NOT mention numeric encodings, indicators, or hidden values
  * Then, explain what this condition implies in clear credit risk assessment terms.
- The final classification must be inferred ONLY from the reasoning above.

Final synthesis:
- Do NOT state or imply the final classification in the first sentence.
- Explicitly distinguish which steps SUPPORT a "good" credit assessment and which steps
  SUPPORT a "bad" credit assessment.
- Explain how the balance of evidence leads to one class outweighing the other.
- State the final classification ONLY at the end.

Make the explanation coherent, professional, and faithful to the provided path.

Input:
Applicant features:
{case_features}

Decision path:
{decision_path}

Output format:
<think>
Step 1: ...
```

```
Step 2: ...
...
Overall, ...
</think>
<answer>good / bad</answer>
```

Diabetes Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a diabetes risk reasoning assistant.

Your task is to generate a clear, step-by-step rationale by STRICTLY following the given
decision path, and then determine whether the individual is diabetic or non-diabetic.

Instructions:
- You are given:
  1. Individual features
  2. A decision path
- Follow the decision path EXACTLY in order.
- Use ONLY the features that appear in the decision path.
- Do NOT introduce any additional features, outside evidence, or new reasoning beyond the path
- Each step must correspond to one condition in the decision path.
- For each step:
  * First, state a purely factual description of the individual based on the feature value.
    - For numeric features: explicitly state the exact value and whether it is <= or > the
      cutoff in the path.
    - For categorical features: state the real-world category only (do not mention encodings).
  * Then, explain what this implies in clear diabetes risk or metabolic health terms.
- The final classification must be inferred ONLY from the reasoning above.

Final synthesis:
- Do NOT state or imply the final classification in the first sentence.
- First distinguish:
  * which steps SUPPORT a diabetic outcome
  * which steps SUPPORT a non-diabetic outcome
- Explain how the balance of evidence leads to one outcome outweighing the other.
- Only AFTER that, state the final classification.

Make the explanation coherent, professional, and faithful to the provided path.

Input:
Individual features:
{case_features}

Decision path:
{decision_path}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer>diabetic / non-diabetic</answer>
```

Homeloan Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
```

```
You are a credit risk reasoning assistant.

Your task is to generate a clear, step-by-step rationale by STRICTLY following the given
decision path, and then determine whether the applicant is Approved or Not Approved for a
home loan.

Instructions:
- You are given:
  1. Applicant features
  2. A decision path
- Follow the decision path EXACTLY in order.
- Use ONLY the features that appear in the decision path.
- Do NOT introduce any additional features, outside facts, or new reasoning beyond the path.
- Each step must correspond to one condition in the decision path.
- For each step:
  * First, state a purely factual description of the applicant based on the applicant features
  and the path condition.
  * For numeric features:
    - explicitly state the exact value
    - explicitly state whether it is <= or > the cutoff in the path
  * For categorical features:
    - state only the real-world category
    - do NOT mention numeric encodings, indicators, or hidden values
  * Then, explain what this condition implies in clear home loan underwriting, affordability,
    and repayment-risk terms.
- The final classification must be inferred ONLY from the reasoning above.

Final synthesis:
- Do NOT state or imply the final classification in the first sentence.
- First explicitly distinguish:
  * which steps SUPPORT an Approved decision
  * which steps SUPPORT a Not Approved decision
- Explain how the balance of evidence leads to one outcome outweighing the other.
- Only AFTER that, state the final classification.

Make the explanation coherent, professional, and faithful to the provided path.

Input:
Applicant features:
{case_features}

Decision path:
{decision_path}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> Approved / Not Approved </answer>
```

Credit-D Dataset

```
System:
You are a helpful and precise reasoning assistant.

User:
You are a credit risk reasoning assistant.

Your task is to generate a clear, step-by-step rationale by STRICTLY following the given
decision path, and then determine whether the borrower will default or no_default.

Instructions:
```

```
- You are given:
  1. Borrower features
  2. A decision path
- Follow the decision path EXACTLY in order.
- Use ONLY the features that appear in the decision path.
- Do NOT introduce any additional features, outside facts, or new reasoning beyond the path.
- Each step must correspond to one condition in the decision path.
- For each step:
  * First, state a purely factual description of the borrower based on the borrower features
    and the path condition.
  * For numeric features:
    - explicitly state the exact value
    - explicitly state whether it is <= or > the cutoff in the path
  * For categorical features:
    - state only the real-world category
    - do NOT mention numeric encodings, indicators, or hidden values
  * Then, explain what this condition implies in clear consumer credit, debt-burden, liquidity
    and delinquency-risk terms (e.g., revolving utilization, debt-to-income, past due history,
    number of credit lines, dependents).
- The final classification must be inferred ONLY from the reasoning above.

Final synthesis:
- Do NOT state or imply the final classification in the first sentence.
- First explicitly distinguish:
  * which steps SUPPORT a no_default decision
  * which steps SUPPORT a default decision
- Explain how the balance of evidence leads to one outcome outweighing the other.
- Only AFTER that, state the final classification.

Make the explanation coherent, professional, and faithful to the provided path.

Input:
Applicant data:
{case_features}

Decision path:
{decision_path}

Output format:
<think>
Step 1: ...
Step 2: ...
...
Overall, ...
</think>
<answer> default / no_default </answer>
```

C Reasoning Data Examples

C.1 Examples of Reasoning Data Curated by ReSS

Alzheimer's Disease Dataset

Original Features:

The atrophy level of hippocampal is 0. The atrophy level of Amygdala is 0. The atrophy level of entorhinal is 0. The atrophy level of parahippocampal is 0. The atrophy level of medial temporal lobe is 0. The atrophy level of fusiform is 2. The atrophy level of precuneus is 0. The atrophy level of superior parietal is 0. The atrophy level of medial temporal lobe (vs cortex) is 2. The atrophy level of parietal lobe (vs cortex) is 0. The atrophy level of frontal lobe is 0. The atrophy level of temporal lobe is 0. The atrophy level of parietal lobe is 0. The atrophy level of occipital lobe is 0. The overall cortical atrophy is 0. The level of ventricle enlargement is 0. The temporal shape of the lateral ventricle is 2. The frontal shape of the lateral ventricle is 2. The evidence of vascular disease is 0. The age is 68.4 years old. The sex is female. The MMSE score is 27.0. The APOE $\varepsilon 4$ status is 1. The marital status is divorced. The years of education are 14.

Curated Reasoning by ReSS:

<think>
Step 1: The patient's MMSE score is 27.0, which is greater than 25.5. This indicates relatively preserved global cognitive function, as MMSE scores above 25 are generally considered within the normal range. In the context of Alzheimer's disease, this finding suggests that there is no significant cognitive impairment at present.

Step 2: The patient's MMSE score is 27.0, which is also less than or equal to 28.5. This further confirms that the patient's cognitive performance is within the upper-normal to mildly reduced range, but still not in the range typically associated with mild cognitive impairment or dementia. This supports the absence of significant cognitive decline.

Step 3: The atrophy level of the hippocampus is 0, which is interpreted as normal. The hippocampus is a key region affected early in Alzheimer's disease, and normal hippocampal volume strongly argues against both mild cognitive impairment and Alzheimer's disease.

Step 4: The patient's age is 68.4 years, which is greater than 64.75 years. Being in this age group does increase the baseline risk for Alzheimer's disease, as advancing age is a major risk factor. However, age alone does not indicate the presence of disease; it only raises the background risk.

Step 5: The atrophy level of the temporal lobe is 0, which is interpreted as normal. The temporal lobe, particularly the medial temporal structures, is often affected in Alzheimer's disease. Normal temporal lobe structure further supports the absence of neurodegeneration typically seen in Alzheimer's disease.

Overall, the evidence supporting Normal Cognition includes the preserved MMSE score (steps 1 and 2), normal hippocampal structure (step 3), and normal temporal lobe structure (step 5). Step 4 (age) increases risk but does not directly indicate disease and is thus neutral. There is no evidence from these steps supporting Mild Cognitive Impairment or Alzheimer's Disease, as there are no findings of cognitive impairment or neurodegeneration in the key regions assessed. The balance of evidence, therefore, strongly favors a classification of Normal Cognition.
</think>
<answer> Normal Cognition </answer>

Figure 4: An example of step-by-step reasoning curated by ReSS on AD dataset.

Creditg Dataset

Original Features:

The Status of existing checking account is no checking account. The Duration in month is 24. The Credit history is existing credits paid back duly till now. The Purpose is furniture/equipment. The Credit amount is 2835. The Savings account/bonds is $100 \leq \cdots < 500$ DM. The Present employment since is ≥ 7. The Installment rate in percentage of disposable income is 3. The Personal status and sex is male : single. The Other debtors / guarantors is none. The Present residence since is 4. The Property is building society savings agreement/ life insurance. The Age in years is 53. The Other installment plans is none. The Housing is own. The Number of existing credits at this bank is 1. The Job is skilled employee / official. The Number of people being liable to provide maintenance for is 1.0. The Telephone is none. The foreign worker is yes.

Curated Reasoning by ReSS:

<think>
Step 1: The applicant has no checking account. This absence of a checking account is generally considered a negative indicator for credit risk, as it may suggest limited engagement with formal banking services or less financial transparency.

Step 2: The applicant has no other installment plans. Having no other installment plans implies that the applicant is not currently obligated to make additional regular payments elsewhere, which is a positive factor for credit risk as it reduces the likelihood of overextension.

Step 3: The applicant's credit history shows that all existing credits have been paid back duly till now, and there is no record of delayed payments. This is a strong positive indicator, as a history of timely repayments demonstrates reliability and responsible credit behavior.

Step 4: The applicant is not classified as female, divorced/separated/married. This condition is neutral in itself but is included as part of the assessment path.

Step 5: The applicant has no other debtors or guarantors. This suggests that the applicant is solely responsible for the credit and is not relying on others to support the application, which can be viewed as a positive sign of financial independence.

Step 6: The applicant's savings account or bonds balance is between 500 and less than 1000 DM. This level of savings is a moderate positive indicator, as it shows some financial reserves, though not the highest possible.

Step 7: The applicant has one existing credit at this bank, which is less than or equal to 2.5. Having only one existing credit indicates a manageable level of current debt, which is favorable for credit risk.

Overall, the following conditions SUPPORT the final classification: the applicant has no other installment plans, has a history of paying back existing credits duly, has no other debtors or guarantors, maintains a moderate level of savings, and has only one existing credit at the bank. The only condition that OPPOSES the classification is the absence of a checking account, which could indicate a higher risk. However, the supporting conditions—especially the strong repayment history and manageable debt—outweigh this single negative factor. These financial conditions collectively indicate that the applicant is classified as "good". </think>
<answer> good </answer>

Figure 5: An example of step-by-step reasoning curated by ReSS on Creditg dataset.

Diabetes Dataset

Original Features:

The Age is 21. The Number of times pregnant is 1. The Diastolic blood pressure is 66. The Triceps skin fold thickness is 23. The Plasma glucose concentration at 2 hours in an oral glucose tolerance test (GTT) is 89. The 2-hour serum insulin is 94. The Body mass index is 28.1. The Diabetes pedigree function is 0.167.

Curated Reasoning by ReSS:

<think>
Step 1: The patient's plasma glucose concentration at 2 hours in an oral glucose tolerance test is 89, which is less than or equal to the cutoff of 127.500. This value is well within the normal range, indicating that the patient's body is effectively regulating blood glucose levels after a glucose load, and there is no evidence of impaired glucose tolerance or hyperglycemia, both of which are risk factors for diabetes.

Step 2: The patient's age is 21, which is less than or equal to the cutoff of 28.500. Younger age is generally associated with a lower risk of type 2 diabetes, as the incidence of diabetes increases with advancing age due to cumulative metabolic stress and potential insulin resistance over time.

Step 3: The patient's body mass index (BMI) is 28.1, which is less than or equal to the cutoff of 30.950. While this BMI falls in the overweight category, it is below the threshold commonly associated with a significantly increased risk of type 2 diabetes, which is typically seen at higher BMI values (≥ 30)

Step 4: The number of times the patient has been pregnant is 1, which is less than or equal to the cutoff of 7.500. A lower number of pregnancies is associated with a reduced risk of gestational diabetes and its contribution to future diabetes risk.

Step 5: The patient's diabetes pedigree function is 0.167, which is less than or equal to the cutoff of 0.672. This low value suggests a limited family history or genetic predisposition to diabetes, reducing the overall risk.

Overall, all five conditions—normal post-load plasma glucose, young age, BMI below the high-risk threshold, low number of pregnancies, and low diabetes pedigree function—SUPPORT the non-diabetic classification. There are no opposing conditions in this diagnostic path, as each feature value is on the side of reduced diabetes risk. These physiological indicators collectively indicate that the patient is diagnosed as "non-diabetic".
</think>
<answer> non-diabetic </answer>

Figure 6: An example of step-by-step reasoning curated by ReSS on Diabetes dataset.

Homeloan Dataset

Original Features:

The gender of the applicant is Male, The marital status of the applicant is Yes, The number of dependents is 2, The education level of the applicant is Graduate, The self-employment status is No, The applicant income is 3073.0, The co-applicant income is 8106.0, The loan amount is 200.0, The loan amount term is 360.0, The credit history value is 1.0, The property area is Urban.

Curated Reasoning by ReSS:

<think>
Step 1: The applicant's credit history value is 1.0, which is greater than 0.500. This indicates that the applicant has a positive credit history, suggesting a reliable record of repaying previous loans or credit obligations. This is a strong indicator of creditworthiness and reduces the lender's risk.

Step 2: The property area is Urban, rather than Semiurban. Urban properties are generally considered to have higher market value and better resale potential, which can make the loan less risky for the lender.

Step 3: The co-applicant income is 8106.0, which is greater than 612.000. This means the co-applicant brings a substantial additional income, improving the overall household's repayment capacity and making it more likely that the loan installments can be managed comfortably.

Step 4: The loan amount requested is 200.0, which is greater than 166.000. This indicates the applicant is seeking a moderately higher loan amount, which could increase risk if not supported by sufficient income or other positive factors.

Step 5: The applicant is married. Being married can sometimes be associated with greater financial stability, as there may be dual incomes or shared financial responsibilities, which can positively influence the ability to repay the loan.

Overall, the following conditions support the approval: a strong positive credit history, the property being located in an urban area, a high co-applicant income, and the applicant's marital status. The only condition that could potentially oppose approval is the moderately higher loan amount requested. However, the applicant's robust credit history and strong combined income, along with the urban property location, provide substantial support for the ability to manage and repay the loan. These supporting factors clearly outweigh the single opposing factor, leading to a favorable assessment. These characteristics collectively support a "Approved" decision regarding the home loan application.
</think>
<answer> Approved </answer>

Figure 7: An example of step-by-step reasoning curated by ReSS on Homeloan dataset.

Credit-D Dataset

Original Features:

The revolving utilization of unsecured lines is 0.704591673. The age of the applicant is 63.0. The number of times the applicant was 30 to 59 days past due without worse delinquency is 0.0. The debt ratio is 0.47144107. The monthly income is 8000.0. The number of open credit lines and loans is 9.0. The number of times the applicant was 90 days late is 0.0. The number of real estate loans or lines is 1.0. The number of times the applicant was 60 to 89 days past due without worse delinquency is 0.0. The number of dependents is 0.0.

Curated Reasoning by ReSS:

<think>
Step 1: The applicant's revolving utilization of unsecured lines is 0.704591673, which is greater than 0.435 and less than or equal to 0.773. This means the applicant is using a moderate to high proportion of their available unsecured credit, but not at an extreme level. While higher utilization can signal increased risk, remaining below 0.773 suggests the risk is not excessive.

Step 2: The number of times the applicant was 90 days late is 0.0, which is less than or equal to 0.500. This indicates the applicant has not experienced any severe delinquencies, a strong sign of reliable credit behavior and low financial distress risk.

Step 3: The number of times the applicant was 30 to 59 days past due without worse delinquency is 0.0, which is less than or equal to 0.500. This further demonstrates the applicant's consistent on-time payment history, reducing the likelihood of financial distress.

Step 4: The number of times the applicant was 60 to 89 days past due without worse delinquency is 0.0, which is less than or equal to 0.500. This reinforces the applicant's record of avoiding moderate delinquencies, supporting a low risk profile.

Step 5: The applicant's debt ratio is 0.47144107, which is less than or equal to 0.563. This indicates that the applicant's total monthly debt payments are well-managed relative to their income, further lowering the risk of financial distress.

Overall, the conditions that support a low risk of financial distress are the applicant's absence of any 30-59, 60-89, or 90+ day delinquencies and a manageable debt ratio. The only factor that could suggest elevated risk is the moderately high revolving utilization, but since it remains below the higher threshold of 0.773, this concern is outweighed by the applicant's strong payment history and controlled debt ratio. These financial conditions collectively indicate that the applicant is classified as "no_default".
</think>
<answer> no_default </answer>

Figure 8: An example of step-by-step reasoning curated by ReSS on Credit-D dataset.

C.2 Examples of Curated Reasoning Data by Direct Reasoning Curation

Alzheimer's Disease Dataset

Original Features:
The atrophy level of hippocampal is 2. The atrophy level of Amygdala is 2. The atrophy level of entorhinal is 2. The atrophy level of parahippocampal is 2. The atrophy level of medial temporal lobe is 2. The atrophy level of fusiform is 2. The atrophy level of precuneus is 0. The atrophy level of superior parietal is 0. The atrophy level of medial temporal lobe (vs cortex) is 2. The atrophy level of parietal lobe (vs cortex) is 0. The atrophy level of frontal lobe is 2. The atrophy level of temporal lobe is 2. The atrophy level of parietal lobe is 0. The atrophy level of occipital lobe is 0. The overall cortical atrophy is 1. The level of ventricle enlargement is 0. The temporal shape of the lateral ventricle is 2. The frontal shape of the lateral ventricle is 2. The evidence of vascular disease is 0. The age is 88.8 years old. The sex is female. The MMSE score is 19.0. The APOE $\varepsilon 4$ status is 1. The marital status is married. The years of education are 13.
Curated Reasoning:
<think>
Step 1: The atrophy level of the hippocampus is 2 (severe atrophy).
Severe hippocampal atrophy is a hallmark of Alzheimer's disease, as this region is critical for memory formation and is typically one of the earliest and most severely affected areas in the disease.

Step 2: The atrophy level of the amygdala is 2 (severe atrophy).
Severe amygdala atrophy is commonly observed in Alzheimer's disease and contributes to behavioral and emotional symptoms, further supporting a diagnosis of advanced neurodegeneration.

Step 3: The atrophy level of the entorhinal cortex is 2 (severe atrophy).
Severe entorhinal atrophy is highly indicative of Alzheimer's disease, as this region is among the first to show pathological changes and is closely linked to memory impairment.

Step 4: The atrophy level of the parahippocampal gyrus is 2 (severe atrophy).
Severe parahippocampal atrophy further supports Alzheimer's disease, as this region is involved in memory encoding and is frequently affected in the disease.

Step 5: The atrophy level of the medial temporal lobe is 2 (severe atrophy).
Severe medial temporal lobe atrophy is a strong indicator of Alzheimer's disease, as this area encompasses several structures crucial for memory and is characteristically involved in the disease process.
...

Step 13: The APOE $\varepsilon 4$ status is 1 (one allele present).
The presence of one APOE $\varepsilon 4$ allele increases the risk for Alzheimer's disease, providing genetic support for the diagnosis.

Synthesis:
- Steps supporting Alzheimer's Disease: Steps 1–8 (severe atrophy in hippocampus, amygdala, entorhinal, parahippocampal, medial temporal, fusiform, frontal, and temporal lobes), Step 9 (pattern of atrophy), Step 10 (ventricle shape changes), Step 11 (MMSE 19.0), Step 12 (advanced age), Step 13 (APOE $\varepsilon 4$ carrier).
- Steps supporting Mild Cognitive Impairment: Step 11 (MMSE 19.0 could be borderline for MCI, but in context of severe atrophy, supports dementia).
- Steps supporting Normal Cognition: None.
- Neutral: None.

The overwhelming majority of features—especially the pattern of severe atrophy in Alzheimer's-vulnerable regions, moderate cognitive impairment, advanced age, and genetic risk—strongly support Alzheimer's Disease over Mild Cognitive Impairment or Normal Cognition. The evidence is consistent, convergent, and highly specific for Alzheimer's Disease, with no features suggesting normal cognition or only mild impairment.

Overall, these characteristics collectively support a "Alzheimer's Disease" classification.
</think>
<answer> Alzheimer's Disease </answer>

Figure 9: An example of step-by-step reasoning obtained by direct reasoning curation on Alzheimer's Disease dataset.

Creditg Dataset

Original Features:

The Status of existing checking account is < 0 DM. The Duration in month is 6. The Credit history is critical account/ other credits existing (not at this bank). The Purpose is radio/television. The Credit amount is 1169. The Savings account/bonds is unknown/ no savings account. The Present employment since is ≥ 7. The Installment rate in percentage of disposable income is 4. The Personal status and sex is male : single. The Other debtors / guarantors is none. The Present residence since is 4. The Property is real estate. The Age in years is 67. The Other installment plans is none. The Housing is own. The Number of existing credits at this bank is 2. The Job is skilled employee / official. The Number of people being liable to provide maintenance for is 1.0. The Telephone is yes, registered under the customers name. The foreign worker is yes.

Curated Reasoning:

<think>
Step 1: The status of the existing checking account is < 0 DM.
This indicates that the applicant currently has a negative balance or no funds in their checking account, which is generally a negative indicator for creditworthiness as it may suggest liquidity issues or poor cash management.

Step 2: The duration in month is 6.
A short loan duration reduces the lender's exposure and risk, as the applicant is expected to repay the credit in a relatively short period. This is a positive factor for credit risk.

Step 3: The credit history is classified as a critical account or other credits existing (not at this bank).
A "critical account" status is a strong negative signal, as it suggests past or current difficulties in meeting credit obligations.

Step 4: The savings account or bonds status is unknown or indicates no savings account.
The absence of savings indicates a lack of financial reserves, which is a negative factor in credit risk assessment.

Step 5: The present employment duration is ≥ 7 years.
Long-term employment (at least 7 years) demonstrates stability and a reliable income source, which is a strong positive indicator for creditworthiness.

Step 6: The installment rate as a percentage of disposable income is 4.
A low installment rate relative to disposable income suggests that the applicant is not overburdened by debt payments, which is a positive sign.

...

Step 13: The foreign worker status is yes.
Being a foreign worker may introduce some uncertainty regarding long-term residence or employment, which is a minor negative factor.

Overall, the applicant displays several strong positive indicators, including long-term employment, ownership of real estate and housing, a skilled occupation, a low installment rate, and a short loan duration. These factors support a good credit risk classification. Opposing these are the negative checking account status, critical credit history, lack of savings, and the applicant's status as a foreign worker. However, the stability and asset ownership, combined with manageable debt obligations and employment history, outweigh the negatives. The applicant's overall financial profile supports the classification as "good".
</think>
<answer> good </answer>

Figure 10: An example of step-by-step reasoning obtained by direct reasoning curation on Creditg dataset.

Diabetes Dataset

Original Features:

The Age is 50. The Number of times pregnant is 6. The Diastolic blood pressure is 72. The Triceps skin fold thickness is 35. The Plasma glucose concentration at 2 hours in an oral glucose tolerance test (GTT) is 148. The 2-hour serum insulin is 0. The Body mass index is 33.6. The Diabetes pedigree function is 0.627.

Curated Reasoning:

<think>
Step 1: The individual's plasma glucose concentration at 2 hours in an oral glucose tolerance test (GTT) is 148 mg/dL.
This value is at the threshold for impaired glucose tolerance (prediabetes) and is suggestive of abnormal glucose metabolism, which is a strong indicator of diabetes risk.

Step 2: The body mass index (BMI) is 33.6.
This places the individual in the obese category, which is a significant risk factor for the development of type 2 diabetes due to its association with insulin resistance.

Step 3: The 2-hour serum insulin is 0.
An undetectable insulin level after a glucose challenge is highly abnormal and may indicate severe beta-cell dysfunction or insulin deficiency, both of which are characteristic of diabetes.

Step 4: The age is 50 years old.
Advancing age is associated with increased risk for type 2 diabetes, as insulin sensitivity tends to decrease and beta-cell function may decline over time.

Step 5: The diabetes pedigree function is 0.627.
This value suggests a moderate to high familial risk for diabetes, indicating a genetic predisposition.

Step 6: The triceps skin fold thickness is 35 mm.
This is above average and suggests increased subcutaneous fat, which is often correlated with overall adiposity and increased diabetes risk.

Step 7: The diastolic blood pressure is 72 mmHg.
This is within the normal range and does not contribute to increased diabetes risk.

Step 8: The number of times pregnant is 6.
Higher parity has been associated with increased diabetes risk, possibly due to metabolic changes during pregnancy.

Overall, the features that support the classification of diabetes are elevated 2-hour plasma glucose (at the diagnostic threshold), obesity (high BMI), undetectable 2-hour insulin (suggesting beta-cell dysfunction), older age, moderate-to-high diabetes pedigree function, increased triceps skin fold thickness, and high parity. The only feature that opposes increased diabetes risk is the normal diastolic blood pressure, which is not a strong protective factor in this context. The multiple strong supporting features, especially the abnormal glucose and insulin findings, clearly outweigh the single opposing condition. Therefore, these clinical and metabolic conditions collectively indicate that the individual is classified as "diabetic".
</think>
<answer> diabetic </answer>

Figure 11: An example of step-by-step reasoning obtained by direct reasoning curation on Diabetes dataset.

Homeloan Dataset

Original Features:

The gender of the applicant is Male, The marital status of the applicant is Yes, The number of dependents is 2, The education level of the applicant is Graduate, The self-employment status is Yes, The applicant income is 5417.0, The co-applicant income is 4196.0, The loan amount is 267.0, The loan amount term is 360.0, The credit history value is 1.0, The property area is Urban, The final_decision is Approved.

Curated Reasoning:

<think>
Step 1: The applicant income is 5,417.0 and the co-applicant income is 4,196.0.
This results in a combined monthly income of 9,613.0. A higher combined income increases the applicant's ability to afford monthly loan repayments, reducing the risk of default and supporting loan approval.

Step 2: The loan amount is 267.0 and the loan term is 360.0 months.
This indicates a moderate loan size spread over a standard 30-year term, which generally results in manageable monthly payments relative to the applicant's income, further supporting affordability.

Step 3: The credit history value is 1.0.
This typically indicates a positive credit history, which is a strong indicator of responsible past borrowing behavior and reduces perceived repayment risk.

Step 4: The self-employment status is Yes.
While self-employment can sometimes introduce income variability, the relatively high combined income and positive credit history mitigate this risk.

Step 5: The number of dependents is 2.
While having dependents increases household expenses, the high combined income suggests that the applicant can still comfortably manage both family obligations and loan repayments.

Step 6: The education level of the applicant is Graduate.
Higher education levels are often associated with better employment prospects and income stability, which supports the applicant's long-term repayment ability.

Overall, the supporting conditions include a high combined income, a manageable loan amount and term, a positive credit history, and a graduate-level education. The only potentially opposing condition is the self-employment status, which could introduce income variability; however, this is outweighed by the strong income, positive credit history, and other favorable factors. Therefore, these characteristics collectively support an "Approved" decision regarding the home loan application.
</think>
<answer> Approved </answer>

Figure 12: An example of step-by-step reasoning obtained by direct reasoning curation on Homeloan dataset.

Credit-D Dataset

Original Features:

The revolving utilization of unsecured lines is 0.704591673. The age of the applicant is 63.0. The number of times the applicant was 30 to 59 days past due without worse delinquency is 0.0. The debt ratio is 0.47144107. The monthly income is 8000.0. The number of open credit lines and loans is 9.0. The number of times the applicant was 90 days late is 0.0. The number of real estate loans or lines is 1.0. The number of times the applicant was 60 to 89 days past due without worse delinquency is 0.0. The number of dependents is 0.0.

Curated Reasoning:

<think>
Step 1: The revolving utilization of unsecured lines is 0.704591673. This means the individual is using approximately 70% of their available unsecured credit, which is relatively high and may indicate a greater risk of financial distress due to limited remaining credit capacity

Step 2: The age of the applicant is 63.0. Being older may suggest greater financial experience and possibly more stable income sources, which can reduce the risk of financial distress

Step 3: The number of times the applicant was 30 to 59 days past due without worse delinquency is 0.0. This indicates no recent short-term delinquencies, which supports financial stability.

Step 4: The debt ratio is 0.47144107. This means that the individual's monthly debt payments are about 47% of their monthly income, which is moderate and generally manageable, suggesting a lower risk of financial distress.

Step 5: The monthly income is 8000.0. A relatively high monthly income provides a strong buffer against financial distress, as it increases the ability to meet debt obligations.

Step 6: The number of times the applicant was 90 days late is 0.0, and the number of times the applicant was 60 to 89 days past due without worse delinquency is also 0.0. The absence of any significant delinquencies further supports financial stability.

Overall, the high revolving utilization of unsecured lines (Step 1) supports the possibility of financial distress, as it indicates heavy use of available credit. However, several conditions oppose financial distress: the absence of any delinquencies (Steps 3 and 6), a moderate debt ratio (Step 4), high monthly income (Step 5), and the applicant's age (Step 2), which may reflect financial maturity. The opposing conditions—especially the strong payment history and high income—outweigh the supporting condition of high credit utilization. Therefore, these financial conditions collectively indicate that the individual is classified as "no_default."
</think>
<answer> no_default </answer>

Figure 13: An example of step-by-step reasoning obtained by direct reasoning curation on Credit-D dataset.

C.3 Examples of Delexicalized Curated Reasoning

Since delexicalization removes domain-specific semantics, the generated reasoning exhibits minimal variation across datasets. Accordingly, we use examples from the Diabetes dataset as a representative case.

Original Features:

The Feature_A is 1.0. The Feature_B is 89.0. The Feature_C is 66.0. The Feature_D is 23.0. The Feature_E is 94.0. The Feature_F is 28.1. The Feature_G is 0.2. The Feature_H is 21.0.

Curated Reasoning:

<think>
Step 1: Feature_B is 89.0, which is $\leq$ 127.500.

Step 2: Feature_H is 21.0, which is $\leq$ 28.500.

Step 3: Feature_F is 28.1, which is $\leq$ 30.950.

Step 4: Feature_A is 1.0, which is $\leq$ 7.500.

Step 5: Feature_G is 0.2, which is $\leq$ 0.672.

</think>
<answer> Label_0 </answer>

Figure 14: An example of delexicalized step-by-step reasoning curated by ReSS on the Diabetes dataset.

D Hyperparameters and Compute Resources

D.1 Decision Tree

For the Decision Tree baseline, we perform grid search over the following hyperparameter space:

- `max_depth` $\in \{4, 5, 6, 7\}$
- `min_samples_split` $\in \{2, 5, 10, 20\}$
- `min_samples_leaf` $\in \{1, 2, 5, 10\}$
- `criterion` $\in \{\text{gini}, \text{entropy}\}$

The optimal hyperparameters are selected based on validation accuracy.

D.2 XGBoost

For the XGBoost baseline, we perform grid search over the following hyperparameter space:

- `max_depth` $\in \{2, 4, 6, 8, 10, 12\}$
- `alpha` $\in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$
- `lambda` $\in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$
- `eta` $\in \{0.01, 0.03, 0.1, 0.3\}$

The optimal hyperparameters are selected based on validation accuracy.

D.3 TabNet

For the TabNet baseline, we perform grid search over the following hyperparameter space:

- `n_d` $\in \{8, 16\}$
- `n_a` $\in \{8, 16\}$
- `n_steps` $\in \{3, 5\}$

- `gamma` $\in \{1.0, 1.3\}$
- `lambda_sparse` $\in \{10^{-3}\}$
- `mask_type` $\in \{\texttt{sparsemax}, \texttt{entmax}\}$
- `learning_rate` $\in \{0.02, 0.01\}$

The optimal hyperparameters are selected based on validation accuracy.

D.4 Supervised Fine-Tuning

For SFT, we adopt the trl (von Werra et al., 2020) framework and train the models using the AdamW optimizer with a learning rate of 1×10^{-5}, weight decay set to 0, and a cosine learning-rate scheduler with a warmup ratio of 0.03. We perform full-parameter fine-tuning with a training batch size of 2. The model is trained for 20 epochs on the HomeLoan, Creditg, and Diabetes datasets, while only three epochs are used for the AD and Credit-D datasets. Model evaluation is performed every 50 training steps. During inference, we use greedy decoding with temperature set to 0.

D.5 Reinforcement Learning

For Direct RL, we use the DisCO algorithm (Li et al., 2025) due to its observed better performance than GRPO. We train the model with a batch size of 128 samples, a mini-batch size of 64, maximum response length of 1024. For each prompt, we generate 8 responses and use only accuracy based rewards. Optimization is performed with AdamW optimizer with a constant learning rate of $2e^{-6}$ and a weight decay of 0.01. The temperature is set to 0.7 for training and 0.0 for evaluation. Models are trained for 320 RL steps, with evaluations conducted every 10 steps.

For ReSS+RL, we train the model for 200 RL steps with a temperature of 1.0, as the starting model has already undergone supervised fine-tuning.

D.6 Compute Resources

All supervised fine-tuning and reinforcement learning experiments are conducted on a single node with 4 NVIDIA A100 GPUs (80GB each). For supervised fine-tuning, a single run typically takes approximately 8 to 24 hours depending on the dataset size. For reinforcement learning, a single run takes around 8 hours. Reasoning data curation is performed via the GPT-4.1 API.

E SUPPLEMENTARY EXPERIMENTAL RESULTS

E.1 EXPLAINABILITY WITHOUT SYMBOLIC GUIDANCE

We further analyze the explainability of the generated reasoning by comparing ReSS, trained without augmented reasoning data, with a DRC+SFT baseline, whose reasoning data is generated without symbolic scaffold constraints. This analysis focuses on explanation sufficiency and necessity, using the same feature-masking protocol as in the main experiments.

Figure 15 shows the sufficiency and necessity curves on Diabetes and AD. For ReSS, masking unused features results in only minor accuracy changes, while masking explanation-referenced features leads to a sharp and monotonic performance drop, indicating strong explanation necessity. In contrast, DRC+SFT consistently exhibits substantially weaker necessity. On Diabetes, masking features referenced by the explanation causes a much smaller accuracy degradation compared to ReSS. The issue becomes more pronounced on AD, where the accuracy drop under necessity masking is minimal and, in some cases, even smaller than that observed when masking unused features, revealing a clear violation of explanation necessity.

This behavior can be attributed to the absence of symbolic guidance during reasoning data generation. Without symbolic scaffold constraints, the LLM tends to incorporate a large number of input features into the explanation, rather than selectively identifying decision-critical evidence. This over-inclusive behavior is particularly detrimental for high-dimensional datasets such as AD, which contains 25 input features. Under DRC+SFT, the generated explanations reference an average of 14 features, many of which are redundant or only weakly correlated with the prediction outcome.

From an explainability perspective, such reasoning fails to distinguish truly influential features from incidental information, leading to substantially degraded explanation necessity. In contrast, the symbolic scaffolds in ReSS explicitly encourage structured reasoning, ensuring that the features highlighted by the explanation are genuinely relevant to the model's decision.

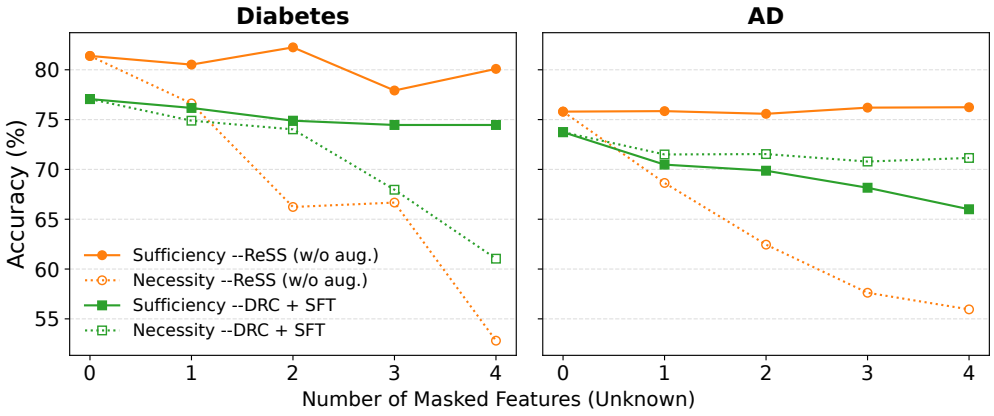

Figure 15: Sufficiency and necessity analysis via feature masking on Diabetes and AD. We compare the ReSS model trained without augmented reasoning data against the DRC+SFT baseline, whose reasoning data is generated without symbolic scaffold constraints.

E.2 KNOWLEDGE HELPS OUR RESS-TRAINED LLM TO CORRECT MISTAKES OF DECISION TREE

Decision trees are purely data-driven models. As a result, some predictions may reflect dataset-specific artifacts or spurious correlations rather than applying domain-plausible rules. While this can introduce a small fraction of noisy or counterintuitive reasoning examples for training, the majority of decision-tree-derived reasoning paths remain consistent with domain commonsense. When trained on such data, the language model does not simply memorize individual paths, but instead learns implicit explanatory structures and domain-level regularities underlying the paths. Consequently,

during inference, the model may generate predictions that deviate from specific decision-tree leaf outcomes while remaining aligned with learned domain-level regularities.

To illustrate this phenomenon, we present a qualitative case study on a testing data from the Diabetes dataset shown in Figure 16. Along this decision path, the decision tree assigns a *non-diabetic* label, which reflects the empirical training label distribution observed in this localized region of the feature space but is wrong. However, every condition along the path corresponds to a well-established risk factor for diabetes according to the domain knowledge. In contrast, our fine-tuned LLM generates a rationale that faithfully follows the same decision path conditions, explicitly referencing each split and providing medically plausible interpretations for them. The generated reasoning emphasizes the cumulative risk implied by high elevated 2-hour plasma glucose levels, extreme obesity, elevated diabetes pedigree function, and high blood pressure, and finally predicts the correct *diabetic* label.

Original Features:
The Age is 23. The Number of times pregnant is 0. The Diastolic blood pressure is 90. The Triceps skin fold thickness is 27. The Plasma glucose concentration at 2 hours in an oral glucose tolerance test (GTT) is 179. The 2-hour serum insulin is 0. The Body mass index is 44.1. The Diabetes pedigree function is 0.686.

Decision Tree Path:
Glucose $>$ 127.500,
BMI $>$ 29.950,
DiabetesPedigreeFunction $>$ 0.434,
BMI $>$ 40.100,
BloodPressure $>$ 80.000,
Answer: non-diabetic

Generated Reasoning by a ReSS-trained model:
<think>
Step 1: The patient's plasma glucose concentration at 2 hours in an oral glucose tolerance test is 179, which is greater than 127.500. This elevated glucose level indicates impaired glucose tolerance and suggests that the patient's body is not effectively clearing glucose from the bloodstream, a hallmark of diabetes mellitus.

Step 2: The body mass index (BMI) is 44.1, which is greater than 29.950. A BMI in this range is classified as severe obesity, which is a strong risk factor for insulin resistance and the development of type 2 diabetes due to increased adiposity and associated metabolic disturbances.

Step 3: The diabetes pedigree function is 0.686, which is greater than 0.434. This value reflects a higher familial predisposition to diabetes, indicating that genetic factors may contribute significantly to the patient's risk profile.

Step 4: The BMI is again considered and is 44.1, which is greater than 40.100. This further emphasizes the presence of extreme obesity, which is associated with a markedly increased risk of developing diabetes due to profound insulin resistance and chronic low-grade inflammation.

Step 5: The diastolic blood pressure is 90, which is greater than 80.000. Elevated blood pressure is a component of metabolic syndrome and is commonly associated with insulin resistance and increased risk of diabetes.

Overall, the following conditions SUPPORT the final classification: the markedly elevated plasma glucose concentration, the extremely high BMI (noted twice), the elevated diabetes pedigree function, and the persistently high diastolic blood pressure. Each of these factors is well-established in the literature as a significant risk or diagnostic indicator for diabetes. There are no conditions in this path that OPPOSE the classification; all features point toward increased diabetes risk. The convergence of severe obesity, hyperglycemia, and abnormal blood pressure strongly outweighs any opposing considerations. These physiological indicators collectively indicate that the patient is diagnosed as "diabetic".
<think>
<answer> diabetic <answer>

Figure 16: An illustrative testing example from the Diabetes dataset.

E.3 Ablation Studies on Conducting RL after ReSS

Since RL is usually regarded as an effective approach to improve the generalization of SFT, we conduct RL on ReSS-trained models. To this end, we conduct RL based on DisCO from the model initialized from ReSS-trained model without augmentation. We compare this approach, denoted by ReSS (w/o aug) + RL, with the direct RL approach in Figure 17, which shows that performing RL after ReSS consistently outperforms directly applying RL to the base model across all datasets, demonstrating that RL benefits substantially from structured reasoning supervision. We also compare the approach ReSS (w/o aug) + RL with ReSS and ReSS (w/o aug.) in Figure 18. We observe that while RL generally improves its SFT counterparts, it is not necessarily better than ReSS-trained model on augmented data. This indicates that our scaffold-invariant data augmentation is an effective approach for improving generalization. We additionally apply RL on top of ReSS with augmented data. However, this setting does not yield further consistent improvements. This suggests that scaffold-invariant augmentation already supplies a strong learning signal, diminishing the marginal benefits of additional RL.

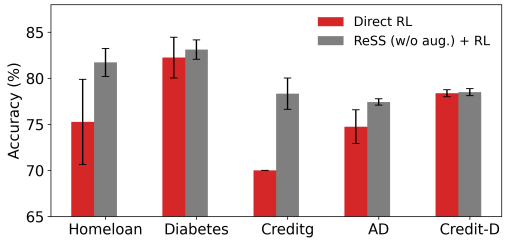

Figure 17: Direct RL vs. ReSS (w/o aug.) + RL.

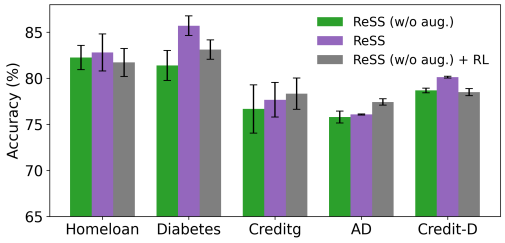

Figure 18: ReSS vs. ReSS (w/o aug.) + RL.

E.4 Comparison with Deep Learning Tabular Methods

For completeness, we compare ReSS with two representative deep learning methods for tabular prediction: TabNet (Arik & Pfister, 2020) and TabPFN (Hollmann et al., 2023). Both are designed purely for label prediction and produce no interpretable reasoning, and are therefore reported separately from the main comparison. For TabPFN we report results using the latest released version, TabPFN-3, with the default pretrained model and no additional fine-tuning.

As shown in Table 4, TabPFN is the strongest non-LLM baseline in our study. ReSS nonetheless achieves higher accuracy than both methods across all five datasets, with the largest gain on Diabetes ($+9.52$ over TabPFN-3). Beyond accuracy, ReSS additionally produces step-by-step rationales whose faithfulness is quantified in Section 4, which neither baseline provides.

Table 4: Comparison with deep learning tabular methods. Results are averaged over three random seeds, with subscripts indicating standard deviations.

Method	HomeLoan	Diabetes	Creditg	AD	Credit-D
TabNet	$79.57_{1.52}$	$77.49_{4.29}$	$75.00_{0.33}$	$74.70_{1.21}$	$77.30_{0.63}$
TabPFN	$81.18_{0.76}$	$76.19_{3.24}$	$75.33_{4.03}$	$75.85_{0.55}$	$79.36_{0.12}$
ReSS	$\mathbf{82.80_{2.01}}$	$\mathbf{85.71_{1.06}}$	$\mathbf{77.67_{1.88}}$	$\mathbf{76.07_{0.06}}$	$\mathbf{80.12_{0.10}}$