

MULEG: Multi-Level Explanation-Guided Feature Attribute Graph for Zero-Shot Tabular Classification

Jiayang Zhang^{1*}, Yunxiao Zhao^{1*†}, Xin Kang¹, Zhichao Yan¹, Boxiang Ma¹, Qinghua Chai¹, Xingtong Yu², and Ru Li^{1†}

1. School of Computer and Information Technology, Shanxi University, Taiyuan, China

2. The Chinese University of Hong Kong Hong Kong, Hong Kong SAR, China

zhangjiayang@sxu.edu.cn, zhaoyunxiao@sxu.edu.cn, kangxin@sxu.edu.cn,

{yanzhichao, maboxiang, charles}@sxu.edu.cn, xtyu@se.cuhk.edu.hk, liru@sxu.edu.cn

Abstract

Zero-shot classification for structured tabular data has attracted increasing attention in recent years. An important line of research is leveraging textual explanations as supervision to improve the modeling of tabular attribute features. Previous studies usually serialize tabular data into flattened explanation sequences, or employ logical rules to model directional associations between tabular attributes and labels. However, these methods often suffer from two major limitations: the former disrupts the inherent row-column topological structure of tables, while the latter mainly represents relations with fixed logical operators or aggregation weights, making it difficult to model learnable dependencies among features. To address these issues, we propose a novel method (MULEG), which harnesses multi-level textual explanations with a feature attribute graph to model both tabular topology and cell-element relations. Specifically, to preserve the intrinsic row-column topological structure of tables, we propose a multi-level textual explanation modeling with sample-level, attribute-level, and label-level perspectives. Moreover, to capture fine-grained feature dependencies, we propose a differentiable message-passing mechanism using a feature attribute graph that allows heterogeneous features to propagate through message passing. Finally, extensive experimental results on datasets covering sixteen tasks demonstrate the effectiveness of the proposed approach. Our code is available at [MULEG](#).

1 Introduction

Structured tabular data has attracted attention in the past few years due to its ubiquity in real-world applications, including classification and regression (Arik and Pfister, 2021), retrieval (Yu et al., 2025; Yan et al., 2026), question answering (Yan et al., 2025b; Xu et al., 2025; Tan et al., 2021)

*Equal contribution.

†Corresponding author.

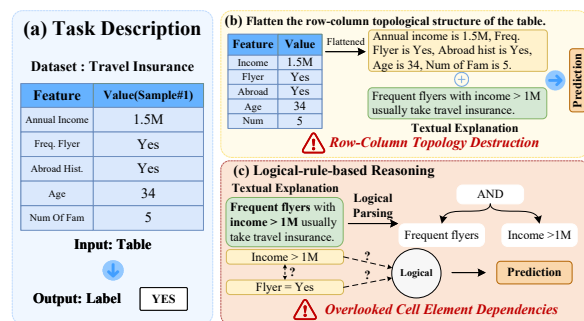


Figure 1: Illustration of zero-shot tabular classification, with (a) showing the task description, (b) showing flattened explanation-based prediction that disrupts row-column topology, and (c) showing logical-rule-based reasoning that overlooks cell-element dependencies.

and recommendation (Dinh et al., 2022; Guo et al., 2017). Tabular data exhibits a row-column topology: each row represents an individual sample, each column corresponds to a semantic attribute, and each feature serves as the basic unit of attribute-value pairs, with historical observed samples. However, real-world decision-making often relies not only on historical observations but also on high-level logical rules expressed in natural language (Chopra et al., 2019; Tomasello, 2009). As illustrated in Figure 1(a), an insurance agent may infer a customer’s potential interest from a rule such as “frequent travelers with high annual income are target customers,” without relying on extensive historical sales records. Inspired by this capability, zero-shot tabular classification (R. Menon et al., 2022) generalizes to unseen tasks using natural-language explanations, without labeled training examples. Given a tabular sample, models predict the target label by jointly reasoning over attribute-value pairs and provided tabular rules.

Previous studies follow two paradigms. The first paradigm serializes tabular data into flattened textual sequences as natural language explanations and combines them with conditional statements for zero-shot classification (R. Menon et al., 2022; Sanh et al., 2022; Team, 2024; Fang et al., 2024;

Ren et al., 2025; Fadlon and Bar, 2026). Such explanations are generated post-hoc (Zhao et al., 2025), independently of the prediction process. R. Menon et al. (2022) formulate the classification as textual entailment between flattened inputs and explanation-derived hypotheses, while Ghosh et al. (2023) model the strength of explanation rules via quantifier-aware reasoning. Recently, instruction-tuned large language models (LLMs) such as T0 (Sanh et al., 2022), FLAN-T5 (Chung et al., 2024), and LLaMA (Team, 2024; Liu et al., 2025) have adopted this research line, treating explanations as task instructions and converting tabular inputs into text before inference. However, as shown in Figure 1(b), serializing tabular features into flattened text sequences disrupts the row-column topological structure of tables. Such flattening weakens the alignment between rows, attributes, and features, making it difficult to preserve how individual features are organized within a structured sample.

The second paradigm introduces explicit reasoning with structured explanations, such as logical reasoning or explanation aggregation (Fan et al., 2026), to model relations between features and labels (Han et al., 2023; Wei et al., 2023; Carrasco et al., 2025; Lin et al., 2026). Han et al. (2023) convert explanations into logical forms for probabilistic inference, while Wei et al. (2023) aggregate pseudo-labels from multiple explanations for test-time adaptation. As shown in Figure 1(c), these methods organize features through logical conditions or aggregated predictions, where features mainly contribute to labels along rule-level paths. However, features are not always independent units: the predictive meaning of one feature may be conditioned on other related features, forming implicit dependency relations within the table. For example, *Annual Income* [1.5M] alone indicates purchasing ability, while *Frequent Flyer* [Yes] reflects frequent travel behavior and *Number of Family Members* [5] suggests a larger household. When combined, however, these features form an implicit dependency pattern: high income makes frequent travel more likely to translate into insurance affordability, and a larger family further shifts the implication toward family-level travel-insurance demand. Such cases suggest that the semantic contribution of a tabular feature can be modulated by its interactions with other features. Thus, fine-grained dependencies among features remain structurally implicit, motivating explicit graph-based modeling.

To this end, we propose MULEG, a novel zero-shot tabular classification method guided by a fea-

ture attribute graph with multi-level textual explanations. The core idea is to jointly construct row-column-aware multi-level textual explanations and an explanation-guided feature attribute graph that organizes attribute-level evidence for zero-shot reasoning. Specifically, we propose a multi-level textual explanation framework with sample-level, attribute-level, and label-level perspectives to preserve the row-column topological structure of tables. Then, we propose a differentiable message-passing mechanism using a feature attribute graph that allows heterogeneous features to capture dependency relations. Inspired by the Structural Causal Model (SCM), (Pearl, 2009; Schölkopf et al., 2012), we model the feature attribute graph by linking tabular features and target labels from a view of SCM. Finally, we conduct extensive experiments on datasets covering sixteen different tabular tasks, and the results show the effectiveness of the proposed approach.

Our contributions can be summarized as follows:

- (1) We propose MULEG, a novel zero-shot classification method that integrates multi-level textual explanations with a feature attribute graph to model the row-column organization of tabular data and semantic dependencies among feature attributes.
- (2) We conduct extensive experiments on sixteen tabular tasks, demonstrating that MULEG significantly outperforms state-of-the-art baselines across diverse task types, including semantic, mixed, numerical, ordinal, and symbolic domains.
- (3) In addition, we provide the comprehensive ablation, generalization and faithfulness analysis, confirming the effectiveness of MULEG.

2 Preliminaries

We first formalize the problem of zero-shot classification and then provide the learning objective. Let $\mathcal{A} = \{a_1, \dots, a_m\}$ denote the set of m tabular attributes, and let $x_i = \{(a_r, \xi_{ir})\}_{r=1}^m$ denote the i -th tabular instance, where ξ_{ir} is the observed value of attribute a_r in x_i .

Zero-Shot Classification. Given a source task $\mathcal{T}_{\text{seen}}$ and a target task $\mathcal{T}_{\text{unseen}}$, their label spaces $\mathcal{Y}_{\text{seen}}$ and $\mathcal{Y}_{\text{unseen}}$ satisfy $\mathcal{Y}_{\text{seen}} \cap \mathcal{Y}_{\text{unseen}} = \emptyset$. The overall candidate label space is defined as $\mathcal{Y} = \{y^{(1)}, \dots, y^{(C)}\} = \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$, where $C = |\mathcal{Y}|$. Specifically, the model is trained using labeled samples from the source task

$$\mathcal{T}_{\text{seen}} = \{(x_i, y_i) \mid y_i \in \mathcal{Y}_{\text{seen}}\}. \quad (1)$$

During testing, it predicts those samples from the

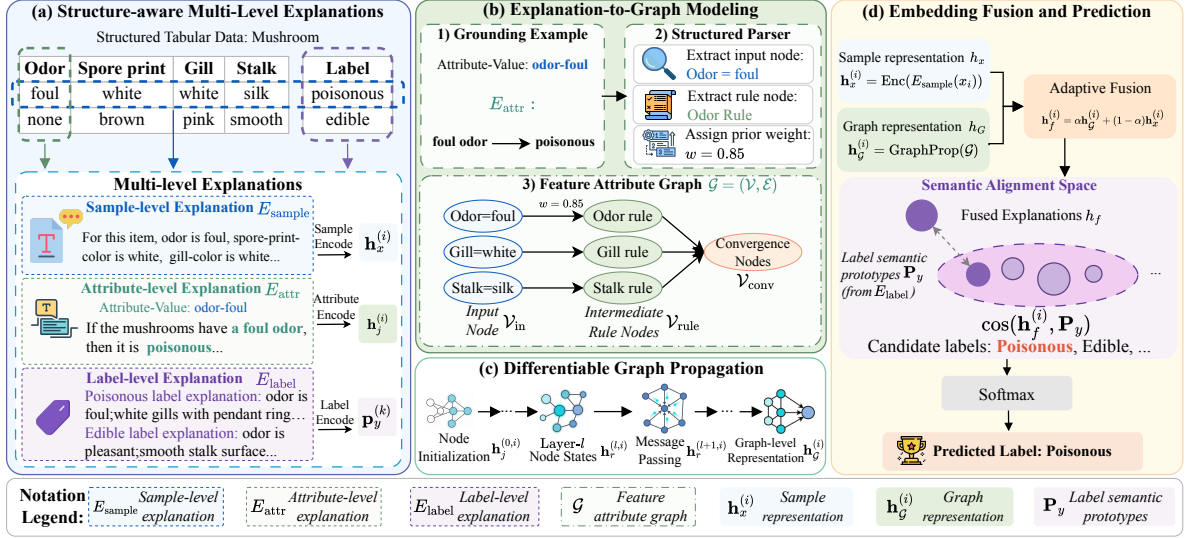


Figure 2: The overview of our proposed MULEG.

target task $\mathcal{T}_{\text{unseen}}$ using only the feature representation of unseen labels $\mathcal{Y}_{\text{unseen}}$.

Learning Objective. The objective is to learn a mapping function $\mathcal{F}_\theta$ that can generalize from the seen task $\mathcal{T}_{\text{seen}}$ to the unseen task $\mathcal{T}_{\text{unseen}}$. In this paper, the model takes a structured tabular sample x_i together with its auxiliary explanations E_i as input, and produces a probability distribution over the candidate label space:

$$\mathcal{F}_\theta : (x_i, E_i) \rightarrow P_\theta(Y_i | x_i, E_i), (x_i, y_i) \in \mathcal{T}_{\text{unseen}}, \quad (2)$$

where $P_\theta(Y_i = y_i | x_i, E_i)$ denotes the probability that sample x_i belongs to class y_i , conditioned on auxiliary explanations E_i . In particular, E_i represents task-level explanatory knowledge (R. Menon et al., 2022) that captures the mapping between input x_i and unseen labels $\mathcal{Y}_{\text{unseen}}$.

3 Our Proposed Approach

Figure 2 outlines the overview of MULEG, a **Multi-level Explanation-Guided Graph** modeling method that provides (a) structure-aware multi-level explanations, covering (b) the construction and propagation from explanations to graphs, as well as (c) the final fusion and prediction process. In the following sections, we elaborate on them in detail.

3.1 Structure-aware Multi-Level Explanations

To model the topology of tabular data, we construct multi-level explanatory information from three perspectives: sample, attribute, and label. Formally,

the auxiliary explanation E_i can be represented as

$$E_i = E(x_i, \mathcal{T}, \mathcal{Y}) = (E_{\text{sample}}(x_i), E_{\text{attr}}(\mathcal{T}), E_{\text{label}}(\mathcal{Y})), \quad (3)$$

where $E_{\text{sample}}(x_i)$ captures sample-level attribute-value explanations, $\mathcal{T}$ is the task-level attribute-value schema, $E_{\text{attr}}(\mathcal{T})$ describes attribute-level association rules, and $E_{\text{label}}(\mathcal{Y}) = \{E_{\text{label}}(y^{(k)})\}_{k=1}^C$ describes label semantics.

Multi-View Gated Sample Embedding. Following (R. Menon et al., 2022), we encode the sample-level explanation as $\mathbf{h}_e^{(i)} = \text{BERT}(E_{\text{sample}}(x_i))_{[\text{CLS}]} \in \mathbb{R}^d$. To capture heterogeneous tabular characteristics, we construct

$$\mathbf{h}_n^{(i)} = \text{LN}(\mathbf{W}_f[\mathbf{h}_b^{(i)}; \mathbf{h}_p^{(i)}; \mathbf{h}_c^{(i)}] + \mathbf{b}_f) \in \mathbb{R}^d,$$

where $\mathbf{h}_b^{(i)} \in \mathbb{R}^{d_b}$, $\mathbf{h}_p^{(i)} \in \mathbb{R}^{d_p}$, and $\mathbf{h}_c^{(i)} \in \mathbb{R}^{d_c}$ denote discretized-bin, high-order, and continuous feature views (Gorishniy et al., 2022); $\mathbf{W}_f \in \mathbb{R}^{d \times (d_b + d_p + d_c)}$ and $\mathbf{b}_f \in \mathbb{R}^d$. The gate is computed as $\mathbf{a}_n^{(i)} = \sigma(\mathbf{W}_g^n[\mathbf{h}_e^{(i)}; \mathbf{h}_n^{(i)}] + \mathbf{b}_g^n) \in (0, 1)^d$, where $\mathbf{W}_g^n \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_g^n \in \mathbb{R}^d$. The sample representation is

$$\mathbf{h}_x^{(i)} = (1 - \mathbf{a}_n^{(i)}) \odot \mathbf{h}_e^{(i)} + \mathbf{a}_n^{(i)} \odot \mathbf{h}_n^{(i)} \in \mathbb{R}^d, \quad (4)$$

where $\odot$ denotes element-wise multiplication.

Attribute-aware Embedding. We use $E_{\text{attr}}(\mathcal{T})$ as a structural prior to derive $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes directed structural dependencies. Let $\mathcal{V}_{\text{in}} \subseteq \mathcal{V}$ be the input-node subset grounded by observable attribute-value pairs. Each node $v_j \in \mathcal{V}$ has a textual description

E_{v_j} , encoded as $\mathbf{p}_j = \text{BERT}(E_{v_j})_{[\text{CLS}]} \in \mathbb{R}^d$. For $v_j \in \mathcal{V}_{\text{in}}$, its relevance to x_i is

$$\rho_{ij} = \frac{(\mathbf{h}_x^{(i)})^\top \mathbf{p}_j}{\|\mathbf{h}_x^{(i)}\|_2 \|\mathbf{p}_j\|_2}, \quad v_j \in \mathcal{V}_{\text{in}}. \quad (5)$$

For $v_j \in \mathcal{V} \setminus \mathcal{V}_{\text{in}}$, we set $\rho_{ij} = 0$. The relevance embedding is $\mathbf{r}_{ij} = \mathbf{W}_\rho \rho_{ij} + \mathbf{b}_\rho \in \mathbb{R}^d$, where $\mathbf{W}_\rho \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_\rho \in \mathbb{R}^d$. The attribute-aware node embedding is defined as

$$\mathbf{h}_j^{(i)} = \mathbf{W}_0[\mathbf{p}_j; \mathbf{h}_x^{(i)}; \mathbf{r}_{ij}] + \mathbf{b}_0 \in \mathbb{R}^d,$$

where $\mathbf{W}_0 \in \mathbb{R}^{d \times 3d}$ and $\mathbf{b}_0 \in \mathbb{R}^d$.

Semantic-aware Label Embedding. For each candidate label $y^{(k)} \in \mathcal{Y}$, we encode its label-level explanation as $\mathbf{p}_y^{(k)} = \text{BERT}(E_{\text{label}}(y^{(k)}))_{[\text{CLS}]} \in \mathbb{R}^d$, $k = 1, \dots, |\mathcal{Y}|$. The label prototype matrix is $\mathbf{P}_y = [\mathbf{p}_y^{(1)}, \mathbf{p}_y^{(2)}, \dots, \mathbf{p}_y^{(|\mathcal{Y}|)}] \in \mathbb{R}^{d \times |\mathcal{Y}|}$.

3.2 Explanation-to-Graph Modeling

To model dependencies among features, we construct a feature attribute graph (FAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from the above explanations. In zero-shot tabular classification, directly modeling

$$P_\theta(Y_i | X_i) \quad (6)$$

may overfit to observed task distributions and data formats, making generalization to unseen tasks difficult. This is because seen and unseen tabular instances often differ in feature dimensions, label spaces, and table types. Therefore, the key is to establish the relationship between table schemas and task labels. To this end, we first analyze the relationships among different conceptual components in zero-shot tabular classification from the perspective of a SCM (Pearl, 2009; Schölkopf et al., 2012), which further motivates an explanation-guided heterogeneous attribute graph. Specifically, we separate stable evidence from nuisance factors and organize explanation-supported evidence through graph structures, rather than relying on causal discovery or feature intervention (Peters et al., 2017). **SCM-inspired view for Tabular Classification.** We conceptualize the observed sample X_i as a mixture of decision-relevant evidence S_i and nuisance factors U_i . For a tabular instance $x_i = \{(a_r, \xi_{ir})\}_{r=1}^m$, the desired reasoning process is formulated as:

$$\begin{aligned} X_i &:= f_X(S_i, U_i, \epsilon_i^X), \quad S_i \perp\!\!\!\perp U_i \perp\!\!\!\perp \epsilon_i^X, \\ R_i &:= f_R(S_i; E_{\text{attr}}(\mathcal{T}), \epsilon_i^R), \\ Z_i &:= f_Z(R_i, \epsilon_i^Z), \\ Y_i &:= f_Y(Z_i; E_{\text{label}}(\mathcal{Y}), \epsilon_i^Y), \end{aligned} \quad (7)$$

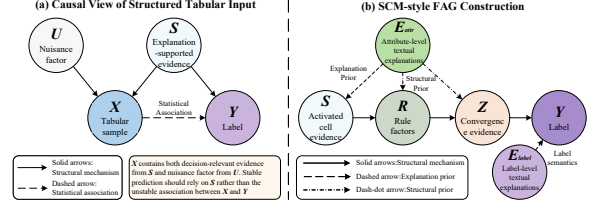


Figure 3: SCM view for zero-shot tabular classification. The observed sample X_i contains decision-relevant evidence S_i and nuisance factors U_i , while explanations guide the reasoning path $S_i \rightarrow R_i \rightarrow Z_i \rightarrow Y_i$.

where R_i denotes explanation-derived rule factors, Z_i represents aggregated decision evidence before prediction, and $\epsilon_i^X, \epsilon_i^R, \epsilon_i^Z, \epsilon_i^Y$ are exogenous noise variables. The semicolon indicates that $E_{\text{attr}}(\mathcal{T})$ and $E_{\text{label}}(\mathcal{Y})$ act as textual priors for the structural mechanisms.

This formulation specifies the reasoning path that MULEG aims to preserve. Instead of directly exploiting the relation $X_i \rightarrow Y_i$, inspired by the rationalization mechanism of (Zhao et al., 2026), attribute-level explanations guide the model to organize task-relevant evidence through:

$$S_i \rightarrow R_i \rightarrow Z_i \rightarrow Y_i, \quad (8)$$

in which, task-relevant evidence is transformed into explanation-derived rule factors, then aggregated into decision evidence for label prediction. Figure 3 illustrates this SCM-inspired view for zero-shot tabular classification. Further discussion is provided in Appendix B.

Feature Attribute Graph. FAG operationalizes the above reasoning path as a graph structure. Specifically, input nodes represent observable attribute-value evidence related to S_i , rule nodes represent explanation-derived rule factors R_i , and convergence nodes represent aggregated decision evidence Z_i . Textual explanations are not treated as observed causal variables; instead, they serve as structural priors for specifying which attributes, intermediate rules, and convergence factors should be connected. We use LLaMA-3-70B to parse task descriptions and attribute-level explanations into an explanation-supported graph structure. The node set in $\mathcal{G}$ is decomposed into three disjoint subsets:

$$\mathcal{V} = \mathcal{V}_{\text{in}} \dot{\cup} \mathcal{V}_{\text{rule}} \dot{\cup} \mathcal{V}_{\text{conv}}. \quad (9)$$

Here, $\mathcal{V}_{\text{in}}$ contains input nodes grounded by observable attribute-value pairs, $\mathcal{V}_{\text{rule}}$ contains intermediate rule nodes derived from textual explanations, and $\mathcal{V}_{\text{conv}}$ contains convergence nodes that aggregate multiple reasoning paths into task-level evidence. For an input instance $x_i = \{(a_r, \xi_{ir})\}_{r=1}^m$,

attribute-level explanations guide the composition of activated input evidence into higher-level rule representations and convergence evidence.

The edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains directed explanation-supported reasoning dependencies between nodes. Each edge $(v_p, v_q) \in \mathcal{E}$ is associated with a confidence weight $w_{pq} \in [0, 1]$, which serves as the structural prior in graph learning. Implementation details of FAG construction are provided in Appendix C.

3.3 Differentiable Graph Learning

Differentiable Graph Propagation. Before graph propagation, each node $v_j \in \mathcal{V}$ is initialized with the attribute-aware embedding $\mathbf{h}_j^{(0,i)} \in \mathbb{R}^d$. Given the feature attribute graph $\mathcal{G}$, MULEG performs iterative message passing to propagate explanation-supported evidence from input nodes through intermediate rule nodes to convergence nodes. Formally, given node representations $\{\mathbf{h}_r^{(l,i)}\}$ at layer l , the next-layer embedding of node v_r is defined as:

$$\mathbf{h}_r^{(l+1,i)} = \sigma \left(\sum_{j \in \mathcal{N}(r)} \lambda_{jr}^{(l,i)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l,i)} \right), \quad (10)$$

where $\mathcal{N}(r) = \{v_j \mid (v_j, v_r) \in \mathcal{E}\}$ denotes the set of predecessor nodes of node v_r , $\mathbf{W}^{(l)} \in \mathbb{R}^{d \times d}$ is a learnable transformation matrix, and $\sigma(\cdot)$ denotes a nonlinear activation function. To enable differentiable reasoning over the $\mathcal{G}$, discrete reasoning dependencies are relaxed into continuous propagation. Specifically, the propagation coefficient from node v_j to node v_r is computed as:

$$\lambda_{jr}^{(l,i)} = \text{softmax}_{j \in \mathcal{N}(r)} \left(\phi(\mathbf{h}_j^{(l,i)}, \mathbf{h}_r^{(l,i)}) \cdot w_{jr} \right), \quad (11)$$

where $\phi(\cdot, \cdot)$ is a learnable compatibility function and w_{jr} denotes the explanation-derived structural prior associated with edge $(v_j, v_r) \in \mathcal{E}$.

Salience-aware Graph Aggregation. After L propagation layers, we compute the node salience score as $s_r^{(L,i)} = \mathbf{w}_s^\top \mathbf{h}_r^{(L,i)}$, where $\mathbf{w}_s \in \mathbb{R}^d$, and perform Top- K pooling to extract the most informative reasoning nodes. The graph-level representation is then obtained by weighted aggregation over the selected node set $\mathcal{V}_K^{(i)}$:

$$\mathbf{h}_g^{(i)} = \sum_{r \in \mathcal{V}_K^{(i)}} \frac{\exp(s_r^{(L,i)})}{\sum_{j \in \mathcal{V}_K^{(i)}} \exp(s_j^{(L,i)})} \mathbf{h}_r^{(L,i)} \in \mathbb{R}^d, \quad (12)$$

$\mathcal{V}_K^{(i)}$ contains the nodes with the top- K salience scores for sample x_i .

3.4 Embedding Fusion and Prediction

Finally, the sample representation $\mathbf{h}_x^{(i)} \in \mathbb{R}^d$ and the graph representation $\mathbf{h}_g^{(i)} \in \mathbb{R}^d$ are fused into a final representation through a fusion coefficient $\alpha \in [0, 1]$:

$$\mathbf{h}_f^{(i)} = \alpha \mathbf{h}_g^{(i)} + (1 - \alpha) \mathbf{h}_x^{(i)} \in \mathbb{R}^d \quad (13)$$

where α is the learnable parameter, which controls the trade-off between the graph representation and the sample representation.

Furthermore, the prediction scores are determined by the cosine similarity between the fused representation $\mathbf{h}_f^{(i)}$ and the label semantic prototype set $\mathbf{P}_y = [\mathbf{p}_y^{(1)}, \dots, \mathbf{p}_y^{(|\mathcal{Y}|)}]$. Here, $\cos(\mathbf{h}_f^{(i)}, \mathbf{P}_y) \in \mathbb{R}^{|\mathcal{Y}|}$ denotes the vector of cosine similarities between $\mathbf{h}_f^{(i)}$ and all label prototype. The final prediction distribution is:

$$P_\theta(y \mid x_i) = \text{Softmax} \left(\cos(\mathbf{h}_f^{(i)}, \mathbf{P}_y) \right). \quad (14)$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We perform the MULEG on datasets from UCI (Dua and Graff, 2019), Kaggle, and Wikipedia, including 36 zero-shot tabular classification datasets such as Mushroom (Pinky et al., 2019), Banknote (Merkurjev et al., 2015), and Indian-Liver-Patient Dataset (Straw and Wu, 2022), etc. They cover diverse domains ranging from five categories: semantic, numerical, mixed, ordinal, and symbolic, and multi-class tasks (2-7 classes). Appendix A provides further dataset details. Following previous research (R. Menon et al., 2022; Ghosh et al., 2023), we utilize 20 of the 36 datasets as seen datasets for training and validation, and the remaining 16 for zero-shot testing. We use accuracy (Acc) to evaluate zero-shot classification performance on unseen tasks.

4.2 Baseline Models

We compare MULEG with following methods: (i) **No-Exp**, a method without explanation involvement; (ii) **Concat**, an explanation concatenation-based method; and (iii) **Sim**, a similarity-based method. All of them use RoBERTa (Liu et al., 2019) as the backbone model. Then, we include one explanation-enhanced method **ExEnt** (R. Menon et al., 2022) which frames prediction as textual entailment, one logical-based zero-shot classification method **CLORE** (Han et al., 2023)

Method	banknote (Num)	liver (Num)	car-eval (Ord)	contra. (Mix)	travel (Mix)	tic-tac (Sym)	Avg.
RoBERTa (No Exp)	55.32	32.56	22.37	43.45	35.81	32.26	36.96
RoBERTa (Concat)	59.72	37.60	23.11	45.64	40.52	30.60	39.53
RoBERTa (Sim)	56.73	67.83	19.08	23.39	34.17	32.29	38.91
T0-3B (No Exp)	55.27	32.17	20.81	46.44	57.04	67.71	46.57
T0-3B (Exp)	55.27	32.17	22.25	23.39	45.83	67.71	41.10
FLAN-T5 (No Exp)	55.27	31.30	22.25	43.39	65.83	67.19	47.54
FLAN-T5 (Exp)	29.09	32.17	22.25	43.39	65.83	67.71	43.41
LLaMA-3 (No Exp)	44.73	67.83	61.10	29.15	34.17	58.85	49.30
LLaMA-3 (Exp)	52.36	67.83	70.79	37.97	34.17	67.71	55.14
ExEnt	44.71	35.70	57.20	21.34	45.61	32.36	39.49
TALC	49.70	47.80	16.50	43.80	34.20	47.80	39.97
CLORE	53.82	49.57	66.47	38.98	46.81	67.71	53.89
MULEG (Ours)	65.82	68.26	71.68	46.78	62.31	67.71	63.76

Table 1: Zero-shot classification accuracy (%) on six **large-scale datasets** similar to (Wei et al., 2023).

Method	award (Sem)	coin-f (Sem)	coin-m (Sem)	elect. (Sem)	s-club (Sem)	s-lg (Sem)	hotel (Ord)	entr. (Mix)	driv. (Sym)	mani. (Sym)	Avg.
RoBERTa (NoExp)	66.72	50.00	50.00	50.00	40.00	60.00	50.00	43.64	37.61	50.00	49.80
RoBERTa (Concat)	70.70	50.00	48.34	50.00	60.00	40.00	50.00	49.30	37.78	50.00	50.61
RoBERTa (Sim)	22.22	50.00	50.00	50.00	60.00	40.00	50.00	47.73	62.50	50.00	48.24
T0-3B (No Exp)	72.22	25.00	50.00	12.50	60.00	60.00	25.00	52.27	12.50	50.00	41.95
T0-3B (Exp)	66.67	25.00	50.00	12.50	40.00	40.00	25.00	52.27	12.50	50.00	37.39
FLAN-T5 (No Exp)	61.11	50.00	50.00	50.00	40.00	60.00	25.00	47.73	62.50	50.00	49.63
FLAN-T5 (Exp)	55.56	50.00	50.00	50.00	40.00	60.00	50.00	52.27	62.50	50.00	52.03
LLaMA-3 (No Exp)	50.00	50.00	75.00	50.00	100.00	100.00	50.00	47.73	62.50	50.00	63.52
LLaMA-3 (Exp)	22.22	75.00	50.00	62.50	100.00	60.00	50.00	47.73	62.50	100.00	63.00
ExEnt	67.78	75.00	50.00	50.00	60.00	60.00	50.00	45.52	37.51	100.00	59.58
CLORE	22.22	50.00	50.00	50.00	60.00	40.00	50.00	45.45	37.50	100.00	50.52
MULEG (Ours)	77.78	75.00	75.00	75.00	100.00	60.00	83.33	50.00	95.83	100.00	79.19

Table 2: Zero-shot classification accuracy (%) on ten **small-scale datasets** similar to (R. Menon et al., 2022).

which performs logical reasoning over parsed explanations, and one chain-style model **TALC** (Wei et al., 2023). In addition, we also evaluate large instruction-tuned models similar to Ghosh et al. (2023) including **T0-3B** (Sanh et al., 2022), **FLAN-T5** (Chung et al., 2024), and **LLaMA 3.1-8B** (Team, 2024; Liu et al., 2025). For these LLMs, we report results under both explanation-free (**No Exp**) and explanation-enhanced (**Exp**) settings. The detailed prompts for these two settings are provided in Appendix D.

4.3 Main Results

(1) As illustrated in Table 1, MULEG generally outperforms all existing baseline methods on six widely used unseen tasks. Notably, MULEG achieves 68.26% and 71.68% on the *liver* and *car-eval* datasets, surpassing the best baselines, RoBERTa (Concat) and LLaMA-3 (Exp), by margins of 10.55% and 0.89% respectively. Overall, MULEG significantly outperforms previous methods involving different types of large and small models, demonstrating the efficacy of MULEG. (2) MULEG also demonstrates remarkable robust-

ness, maintaining superior performance despite the limited availability of test samples. As reported in Table 2, we observe that MULEG consistently outperforms previous methods in terms of average performance. In particular, even on datasets with extremely limited samples (e.g., *s-club* and *mani.*) or high task difficulty (e.g., *entr.*), MULEG still maintains competitive performance. (3) As demonstrated from Tables 1 and 2, MULEG exhibits robust capabilities across diverse dataset types. MULEG achieves clear advantages across four data types, namely Semantic (Sem), Numerical (Num), Ordinal (Ord), and Symbolic (Sym). Moreover, MULEG demonstrates competitive performance on the Mixed datasets.

4.4 Ablation Study

To verify component contributions (Figure 4), we evaluate variants across two categories: (i) **Structural Modeling** variants test input representations: *w/o* FAG (replaces feature attribute graph with raw text), *w/o* A-Exp (inputs values only), and *w/o* L-Exp (only uses atomic label names); (ii) **Architectural Components** variants assess specific module

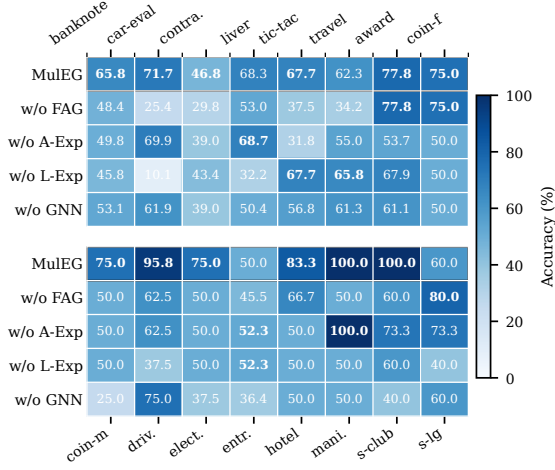


Figure 4: Ablation studies on sixteen datasets. Darker colors indicate higher classification accuracy.

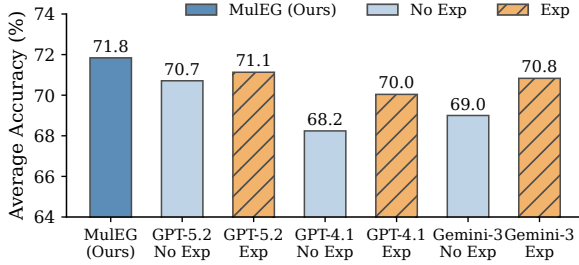


Figure 5: Results on recent LLMs.

removals: graph aggregation (w/o GNN).

Impact of Structural Modeling. Removing feature attribute graph structures (“w/o FAG”) or attribute names (“w/o A-Exp”) causes precipitous drops of 46.25% on *car-eval* and 58.33% on *driv.* respectively, which indicates that raw text and isolated label words lack the topological certainty required for complex reasoning. Conversely, the slight gain of 3.52% by “w/o L-Exp” on *travel* implies that elaborate label descriptions may occasionally introduce semantic noise, though the structural backbone remains vital.

Effectiveness of Architectural Components. Module ablations confirm the role of graph aggregation in structured reasoning. Removing graph aggregation (“w/o GNN”) prevents effective message passing over the feature attribute graph, indicating that the complete architecture ensures superior robustness across diverse domains.

4.5 Generalization Analysis

Results on Larger Tabular Datasets. To evaluate the generalization ability of MULEG, we conduct additional experiments on seven larger datasets following recent studies (Shi et al., 2025; Hegselmann et al., 2023). These datasets cover different scales and feature spaces, including Adult, Bank, Blood,

Method	Adult	Bank	Blood	Calif.	Heart	Airbn.	Merca.	Avg.
RoBERTa	65.32	67.89	52.34	48.23	62.78	46.12	45.37	55.44
RoBERTa+Exp	68.27	70.23	55.67	51.45	66.34	49.85	48.62	58.63
ExEnt	72.14	73.56	60.45	54.78	70.12	52.36	51.94	62.19
CLORE	70.56	69.78	58.23	42.34	68.45	55.71	54.28	59.91
T0-3B+Exp	66.71	68.12	53.78	59.12	64.56	57.63	56.89	60.97
FLAN-T5+Exp	64.35	66.90	51.89	57.33	63.22	56.17	55.42	59.33
LLaMA-3+Exp	77.12	73.56	66.90	62.45	78.67	63.28	61.75	69.10
MULEG	89.56	88.25	74.11	76.54	84.31	77.21	75.86	80.83

Table 3: Results on seven larger tabular datasets (Shi et al., 2025; Hegselmann et al., 2023). Avg. denotes the average accuracy across the seven datasets.

California, Heart, Airbnb, and Mercari. As shown in Table 3, MULEG achieves the best average accuracy of 80.83%, outperforming the strongest baseline by 11.73 percentage points, indicating that the proposed approach can generalize to larger and higher-dimensional structured datasets.

Results on Recent LLMs. We further compare MULEG with recent LLMs, including GPT-5.2, GPT-4.1, and Gemini-3, under both explanation-free and explanation-enhanced settings. Figure 5 reports the average accuracy over 16 unseen datasets. The results show that directly leveraging explanations into LLMs does not always yield stable improvements; on structurally demanding tasks such as car-evaluation and tic-tac-toe, explanations may even degrade performance. In contrast, MULEG provides an explicit feature attribute graph to model dependency relations among features, achieving more structurally grounded zero-shot prediction.

4.6 Faithfulness Analysis

Following established protocols (Wang et al., 2022; DeYoung et al., 2020), we evaluate faithfulness by measuring the performance degradation caused by masking the top- K nodes versus a random baseline (Appendix E provides details). As shown in Figure 7, masking nodes identified by MULEG causes a larger accuracy drop than random masking. The widening gap as K increases indicates that the model relies on joint attribute dependencies rather than isolated features. Moreover, the moderate degradation at lower K values suggests that MULEG avoids over-reliance on fragile single-feature shortcuts by aggregating evidence across multiple supporting factors.

4.7 Sensitivity Analysis

In MULEG, α in Eq.(13) controls the trade-off between the tabular sample representation alone ($\alpha = 0.0$) and the multi-level explanation propagated graph representation ($\alpha = 1.0$). Figure 8 summarizes the overall accuracy distribution across

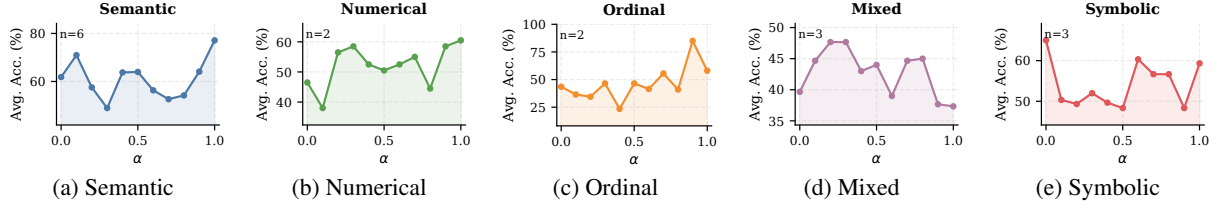


Figure 6: Sensitivity analysis of the fusion coefficient α across task types. Each subplot reports the average classification accuracy over datasets of the same type as α varies.

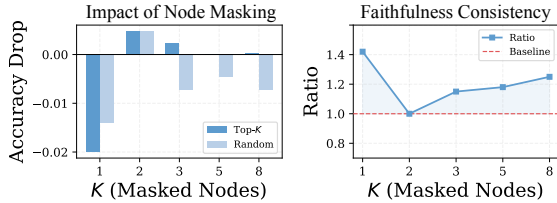


Figure 7: Faithfulness analysis. Masking top- K high-importance nodes consistently causes larger accuracy drops than random masking (left), and yields faithfulness ratios above 1.0 across different values of K (right).

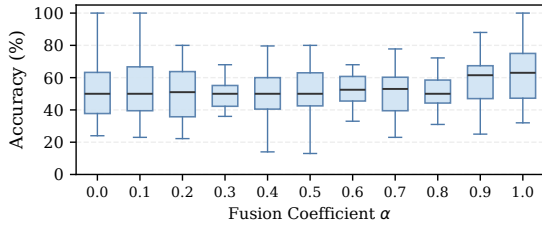


Figure 8: Sensitivity analysis of the fusion coefficient α . Each box summarizes the accuracy across all datasets.

all datasets under different α values. We observe that when α approaches 1.0, the performance reaches its peak, highlighting the critical role of our graph representation in MULEG. Figure 6 further shows that graph representations provide clear gains on Semantic, Numerical, Ordinal, and Symbolic tasks, while improvements on Mixed-type tasks are weaker due to heterogeneous node interference during graph propagation. Additional analyses are provided in Appendix F.

4.8 Qualitative Case Study

Finally, we provide additional case studies as shown in Figure 9. More Analysis can be seen in Appendix G. Appendix H also investigate the impact of graph propagation depth.

5 Related Work

Classification and Zero-shot Classification. Classification learns a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from inputs to labels (Murphy, 2012). Prior work improves generalization by incorporating auxiliary signals such

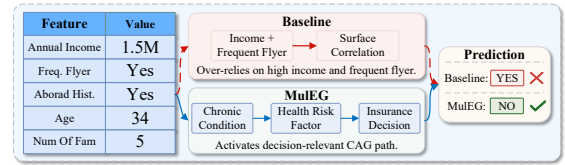


Figure 9: Qualitative Case Study.

as semantic labels, metadata, descriptions, prompts, and explanations (Hancock et al., 2018; Kuhn and Johnson, 2019; Srivastava et al., 2018; He et al., 2024), often with multi-task learning (Suresh et al., 2017), self-supervised/transfer learning (Dosovitskiy et al., 2014; Chen et al., 2020), or prompt-based generalization (Ye et al., 2021; Mishra et al., 2022; Chen et al., 2022; Sanh et al., 2022). Zero-shot classification predicts unseen labels without labeled training examples (Pourpanah et al., 2022). Our setting further requires modeling tabular row-column topology and feature dependencies.

Tabular Classification with Explanations.

Explanation-based zero-shot tabular classification mainly follows two research lines. One line serializes tables into text for entailment- or instruction-based prediction (Yin et al., 2019; R. Menon et al., 2022; Sanh et al., 2022; Ghosh et al., 2023; Team, 2024; Liu et al., 2025; Fang et al., 2024; Ren et al., 2025; Fadlon and Bar, 2026), but weakens the row-column structure of tabular data. The other line converts explanations into logical forms, task descriptions, or executable reasoning structures (Han et al., 2023; Wei et al., 2023; Carrasco et al., 2025; Lin et al., 2026), but does not sufficiently model semantic dependencies among feature attributes. MULEG differs from these approaches by combining multi-level textual explanations with a feature attribute graph.

6 Conclusion

In this paper, we propose a novel zero-shot classification framework termed MULEG, which integrates a feature attribute graph with multi-level explanations to capture both the row-column structure

of tabular data and semantic dependencies among feature attributes. We build the feature attribute graphs with multi-level explanations and then introduce a differentiable message-passing mechanism that allows feature attributes from heterogeneous representation to propagate through the graph. We conduct extensive experiments on zero-shot classification datasets covering sixteen unseen tasks, and the results validate the effectiveness of MULEG.

Limitations

Despite its overall effectiveness, MULEG may encounter challenges on mixed-type tabular tasks. In these tasks, heterogeneous node types may introduce interference during graph aggregation and representation propagation, potentially affecting attribute graph updates and limiting performance. Addressing these interactions and improving instance-specific reasoning over heterogeneous node types are important directions for future work.

Ethics Statement

This paper does not involve the presentation of a new dataset and the utilization of demographic or identity characteristics information.

Acknowledgements

We thank all the anonymous reviewers for their valuable and constructive feedback. This work was supported by the National Natural Science Foundation of China (No.62376144), the China Scholarship Council, the Natural Language Processing Innovation Team (Sanjin Talents) Project of Shanxi Province, and the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016).

References

2019. [Internet Firewall Data](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5131M>.
2020. [Shill Bidding Dataset](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5Z611>.
- Neda Abdelhamid, Aladdin Ayesah, and Fadi Thabtah. 2014. [Phishing detection based associative classification data mining](#). *Expert Systems with Applications*, 41(13):5948–5959.
- Stefan Aeberhard, Danny Coomans, and Olivier De Vel. 1994. [Comparative analysis of statistical pattern recognition methods in high dimensional settings](#). *Pattern Recognition*, 27(8):1065–1077.
- Muhammad Amin and Amir Ali. 2018. [Caesarian Section Classification Dataset](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5N59X>.
- Sercan Ö Arik and Tomas Pfister. 2021. [Tabnet: Attentive interpretable tabular learning](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687.
- Marko Bohanec and Vladislav Rajkovic. 1988. [Knowledge acquisition and explanation for multi-attribute decision making](#). In *8th intl workshop on expert systems and their applications*, pages 59–78. Avignon France.
- Luis M Candanedo and Véronique Feldheim. 2016. [Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models](#). *Energy and buildings*, 112:28–39.
- Lucas Carrasco, Felipe Urrutia, and Andrés Abeliuk. 2025. [Zero-shot decision tree construction via large language models](#). *Preprint*, arXiv:2501.16247.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. [Meta-learning via language model in-context tuning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics.
- Sahil Chopra, Michael Henry Tessler, and Noah D Goodman. 2019. [The first crank of the cultural ratchet: Learning and transmitting concepts through language](#). In *Proceedings of the annual meeting of the cognitive science society*, volume 41.
- Hyung Won Chung, Le Hou, Shayne Longpre, et al. 2024. [Scaling instruction-finetuned language models](#). *J. Mach. Learn. Res.*, 25:70:1–70:53.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for*

- Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Tuan Dinh, Yuchen Zeng, Ruisi Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. 2022. [Lift: Language-interfaced fine-tuning for non-language machine learning tasks](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 766–774.
- Dheeru Dua and Casey Graff. 2019. [UCI machine learning repository](#).
- Matthias Elter, Rüdiger Schulz-Wendland, and Thomas Wittenberg. 2007. [The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process](#). *Medical physics*, 34(11):4164–4172.
- Daniel Fadlon and Kfir Bar. 2026. [How much pretraining does structured data need?](#) In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3352–3365.
- Yue Fan, Hu Zhang, Yunxiao Zhao, Guangjun Zhang, Hao Zhan, Ru Li, Hongye Tan, and Yuanlong Wang. 2026. [Leibniz: Theory-of-mind driven neuro-symbolic logical reasoning via multi-agent collaboration](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20186–20210, San Diego, California, United States. Association for Computational Linguistics.
- Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. [Large language models \(llms\) on tabular data: Prediction, generation, and understanding—a survey](#). *arXiv preprint arXiv:2402.17944*.
- Sayan Ghosh, Rakesh R. Menon, and Shashank Srivastava. 2023. [LaSQuE: Improved zero-shot classification from explanations through quantifier modeling and curriculum learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7403–7419, Toronto, Canada. Association for Computational Linguistics.
- Yury Gorishniy, Ivan Rubachev, and Artem Babenko. 2022. [On embeddings for numerical features in tabular deep learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24991–25004.
- Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. [Deepfm: a factorization-machine based neural network for ctr prediction](#). In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17*, page 1725–1731. AAAI Press.
- Chi Han, Hengzhi Pei, Xinya Du, and Heng Ji. 2023. [Zero-shot classification by logical reasoning on natural language explanations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8967–8981, Toronto, Canada. Association for Computational Linguistics.
- Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. [Training classifiers with natural language explanations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895. Association for Computational Linguistics.
- Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. [Harnessing explanations: Llm-to-llm interpreter for enhanced text-attributed graph representation learning](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. [Tabllm: Few-shot classification of tabular data with large language models](#). In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 5549–5581. PMLR.
- Waldemar Koczkodaj. 2017. [Somerville Happiness Survey](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5PW36>.
- Murat Koklu and Ilker Ali Ozkan. 2020. [Multiclass classification of dry beans using computer vision and machine learning techniques](#). *Computers and Electronics in Agriculture*, 174:105507.
- Max Kuhn and Kjell Johnson. 2019. [Feature engineering and selection: A practical approach for predictive models](#). Chapman and Hall/CRC.
- Juncai Li, Ru Li, Xiaoli Li, Qinghua Chai, and Jeff Z. Pan. 2024. Inference helps PLMs’ conceptual understanding: Improving the abstract inference ability with hierarchical conceptual entailment graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22088–22104, Miami, Florida, USA. Association for Computational Linguistics.
- Yupian Lin, Guangya Yu, Cheng Yuan, Huan Du, Hui Luo, Yang Bian, Jingping Liu, Zhidong He, Wen Du, and Tong Ruan. 2026. [Logtop: Logic tree-of-program with table instruction-tuned llms for controlled logical table-to-text generation](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5291–5303.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*.
- Wei-Yin Loh. 1997. [Teaching Assistant Evaluation](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C55P6M>.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. 2019. [Data poisoning against differentially-private learners: attacks and defenses](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4732–4738.
- Ekaterina Merkurjev, Egil Bae, Andrea L Bertozzi, and Xue-Cheng Tai. 2015. [Global binary optimization on graphs for classification of high-dimensional data](#). *Journal of Mathematical Imaging and Vision*, 52(3):414–435.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.
- Kevin P Murphy. 2012. [Machine Learning: A Probabilistic Perspective](#). MIT Press.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. [Elements of Causal Inference: Foundations and Learning Algorithms](#). MIT Press.
- Nusrat Jahan Pinky, SM Islam, and R Sharmin Alice. 2019. [Edibility detection of mushroom using ensemble methods](#). *International Journal of Image, Graphics and Signal Processing*, 11(4):55–62.
- Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. [A review of generalized zero-shot learning methods](#). *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070.
- Rakesh R. Menon, Sayan Ghosh, and Shashank Srivastava. 2022. [CLUES: A benchmark for learning classifiers using natural language explanations](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6523–6546, Dublin, Ireland. Association for Computational Linguistics.
- Md Anisur Rahman, Md Zahidul Islam, and Terry Bossomaier. 2015. [Modex and seed-detective: Two novel techniques for high quality clustering by using good initial seeds in k-means](#). *Journal of King Saud University-Computer and Information Sciences*, 27(2):113–128.
- Hans Reichenbach. 1956. *The Direction of Time*. Dover Publications, Mineola, N.Y.
- Kevin Ren, Santiago Cortes-Gomez, Carlos Miguel Patiño, Ananya Joshi, Ruiqi Lyu, Jingjing Tang, Alistair Turcan, Khurram Yamin, Steven Wu, and Bryan Wilder. 2025. [Predicting language models’ success at zero-shot probabilistic prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18337–18363.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2022. [Multitask prompted training enables zero-shot task generalization](#). In *International Conference on Learning Representations*.
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Moyschidou. 2012. [On causal and anticausal learning](#). In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262.
- Ruxue Shi, Hengrui Gu, Xu Shen, and Xin Wang. 2025. [Harnessing llms explanations to boost surrogate models in tabular data classification](#). In *International Conference on Database Systems for Advanced Applications*, pages 71–87. Springer.
- Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. [Zero-shot learning of classifiers from natural language quantification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316. Association for Computational Linguistics.
- Isabel Straw and Honghan Wu. 2022. [Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction](#). *BMJ health & care informatics*, 29(1):e100457.

- Ya Su, Hu Zhang, Dan Qiao, YuJie Wang, Yunxiao Zhao, Yue Fan, Shike Li, Ru Li, and Hongye Tan. 2026. Suggest-verify-revise: A three-stage document-level event causality identification with narrative consistency. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19067–19091. Association for Computational Linguistics.
- Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. [Clinical intervention prediction and understanding using deep networks](#). *arXiv preprint arXiv:1705.08498*.
- Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. [GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online. Association for Computational Linguistics.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard university press.
- Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. [A fine-grained interpretability evaluation benchmark for neural NLP](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–84, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kangda Wei, Sayan Ghosh, Rakesh Menon, and Shashank Srivastava. 2023. [Leveraging multiple teachers for test-time adaptation of language-guided classifiers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7068–7088. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Hao Xu, Yunxiao Zhao, Jiayang Zhang, Zhiqiang Wang, and Ru Li. 2025. [LOG: A local-to-global optimization approach for retrieval-based explainable multi-hop question answering](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9085–9095, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zhichao Yan, Jiaoyan Chen, Jiapu Wang, Xiaoli Li, Ru Li, and Jeff Z Pan. 2025a. [Decomposing and revising what language models generate](#). *arXiv preprint arXiv:2509.00765*.
- Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Jiye Liang, Ru Li, and Jeff Z. Pan. 2025b. [Atomic fact decomposition helps attributed question answering](#). *IEEE Transactions on Knowledge and Data Engineering*, 37(12):6959–6972.
- Zhichao Yan, Yunxiao Zhao, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Ru Li, and Jeff Z Pan. 2026. [Beyond factual accuracy: Evaluating global reasoning integrity in rag systems with logicscore](#). *arXiv preprint arXiv:2601.15050*.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. [CrossFit: A few-shot learning challenge for cross-task generalization in NLP](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. [Gnnexplainer: Generating explanations for graph neural networks](#). In *NeurIPS*, volume 32, pages 9240–9251.
- Xiaohan Yu, Pu Jian, and Chong Chen. 2025. [TableRAG: A retrieval augmented generation framework for heterogeneous document reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14063–14082, Suzhou, China. Association for Computational Linguistics.
- Yunxiao Zhao, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2024. [AGR: Reinforced causal agent-guided self-explaining rationalization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–518, Bangkok, Thailand. Association for Computational Linguistics.
- Yunxiao Zhao, Zhiqiang Wang, Xingtong Yu, Xiaoli Li, Jiye Liang, and Ru Li. 2026. [Learnable game-theoretic policy optimization for data-centric self-explanation rationalization](#). *IEEE Transactions on Knowledge and Data Engineering*, 38(2):1159–1173.
- Yunxiao Zhao, Hao Xu, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2025. [Explaining black-box language models with knowledge probing systems: A post-hoc explanation perspective](#). In *International Conference on Database Systems for Advanced Applications*, pages 220–235. Springer.

Appendix

A Dataset Description

We present detailed statistics and brief descriptions of all datasets used in our experiments in this appendix. In total, the benchmark comprises 36 real-world tabular classification tasks spanning five distinct task types. Specific details and statistics for each dataset are summarized in Table 4. We classify these tasks based on their feature semantics and reasoning requirements:

Task Types. *Semantic tasks* involve rich textual or categorical metadata that require semantic understanding of attribute labels. *Mixed tasks* contain heterogeneous numerical, categorical, and textual features. *Numerical tasks* are primarily composed of continuous measurements requiring distributional modeling. *Ordinal tasks* involve labels or features with an intrinsic ordering, such as ratings or levels. *Symbolic tasks* involve logical rules, board states, or mathematical properties that require symbolic reasoning.

A.1 Seen Tasks (20 Datasets)

Semantic tasks.

- website-phishing: Classifies websites as legitimate, suspicious, or phishing based on hand-crafted URL and webpage indicators.
- internet-firewall: Predicts firewall actions from network-flow metadata, including ports, bytes, packets, and duration.
- mushroom: Predicts whether a mushroom is edible or poisonous from categorical morphological attributes such as cap shape, gill color, and odor.
- shill-bidding: Detects shill bidding behavior from auction-level and bidder-level behavioral statistics.
- water-potability: Predicts whether water is potable from physicochemical measurements such as pH, solids, and turbidity.
- color-luminance: Predicts whether luminance exceeds a predefined threshold from color identifiers.

Numerical tasks.

- dry-bean: Predicts bean variety from numerical shape descriptors.

- occupancy-detection: Predicts room occupancy from continuous environmental sensor readings.
- vertebral-column: Predicts orthopedic assessment categories from numerical spinal measurements.
- mammographic-mass: Predicts mass severity from numerical imaging assessment features.

Mixed tasks.

- caesarian-section: Predicts C-section versus non-C-section deliveries from maternal and clinical indicators.
- stroke-prediction: Predicts stroke occurrence from medical history, demographics, and lifestyle features.
- job-change: Predicts whether an individual is likely to change jobs from demographic and experience-related features.
- campus-placement: Predicts student placement outcomes from academic records and profile features.
- engineering-placement: Predicts placement outcomes from mixed academic and experiential indicators.

Ordinal tasks.

- wine: Predicts wine category from chemical analysis features.
- student-performance: Predicts academic performance levels from demographic and schooling features.
- teaching-assistant-evaluation: Predicts teaching assistant performance on a three-level ordinal scale.
- somerville-happiness: Predicts resident happiness from survey-based quality-of-life indicators.

Symbolic tasks.

- **proteinogenic-acid**: Predicts whether an amino acid is proteinogenic from symbolic biochemical indicators.

A.2 Unseen Tasks (16 Datasets)

Semantic tasks.

- award-nomination-result: Predicts whether an entry is awarded based on award-related metadata.

Split	Dataset Name	Samples	Feats.	Classes	Task Type	Source
Train / Dev (Seen Tasks)	internet-firewall (int, 2019)	65,531	5	4	Semantic	UCI
	mushroom (Pinky et al., 2019)	8,124	5	2	Semantic	UCI
	website-phishing (Abdelhamid et al., 2014)	1,353	5	3	Semantic	UCI
	shill-bidding (shi, 2020)	6,319	5	2	Semantic	UCI
	water-potability	2,010	5	2	Semantic	Kaggle
	color-luminance	23	3	2	Semantic	Wikipedia
	dry-bean (Koklu and Ozkan, 2020)	13,611	5	7	Numerical	UCI
	occupancy-detection (Candanedo and Feldheim, 2016)	8,141	5	2	Numerical	UCI
	mammographic-mass (Elter et al., 2007)	830	5	2	Numerical	UCI
	vertebral-column (Ma et al., 2019)	310	5	3	Numerical	UCI
	caesarian-section (Amin and Ali, 2018)	80	5	2	Mixed	UCI
	job-change	8,954	5	2	Mixed	Kaggle
	stroke-prediction	4,918	5	2	Mixed	Kaggle
	engineering-placement	2,965	5	2	Mixed	Kaggle
	campus-placement	214	5	2	Mixed	Kaggle
	teaching-assistant (Loh, 1997)	151	5	3	Ordinal	UCI
	student-performance	666	5	4	Ordinal	UCI
	somerville-happiness (Koczkodaj, 2017)	143	5	2	Ordinal	UCI
	wine (Aeberhard et al., 1994)	178	5	3	Ordinal	UCI
	proteinogenic-acid	41	3	2	Symbolic	Wikipedia
Test (Unseen Tasks)	award-nomination-result	54	3	2	Semantic	Wikipedia
	coin-face-value	21	3	2	Semantic	Wikipedia
	coin-metal	24	4	3	Semantic	Wikipedia
	election-outcome	24	5	2	Semantic	Wikipedia
	soccer-club-region	15	5	2	Semantic	Wikipedia
	soccer-league-type	15	5	2	Semantic	Wikipedia
	banknote-auth (Merkurjev et al., 2015)	1,373	4	2	Numerical	UCI
	indian-liver-patient (ILPD) (Straw and Wu, 2022)	575	5	2	Numerical	UCI
	contraceptive-method (Rahman et al., 2015)	1,473	5	3	Mixed	UCI
	entrepreneur	218	5	2	Mixed	Kaggle
	travel-insurance	1,986	5	2	Mixed	Kaggle
	car-evaluation(Bohanec and Rajkovic, 1988)	1,728	5	4	Ordinal	UCI
	hotel-rating	16	4	2	Ordinal	Wikipedia
	tic-tac-toe-endgame	958	5	2	Symbolic	UCI
	driving-points	21	5	2	Symbolic	Wikipedia
	manifold-orientability	18	3	2	Symbolic	Wikipedia

Table 4: Dataset Splits and Statistics. The datasets are grouped by the data split (Train/Dev vs. Test) and their specific Task Type. Sources without formal citations include clickable links.

- coin-face-value: Predicts whether a coin’s face value exceeds one unit from numismatic attributes.
- coin-metal: Predicts the type of precious metal used in a coin, such as gold, silver, or none.
- election-outcome: Predicts whether a candidate is elected or defeated from election-level metadata.
- soccer-club-region: Predicts the geographic region of a soccer club from club and league descriptors.
- soccer-league-type: Predicts whether a club belongs to a men’s or women’s league from metadata attributes.

Numerical tasks.

- banknote-authentication: Predicts whether a banknote is genuine or counterfeit from numerical image-statistics features.
- indian-liver-patient: Predicts liver disease status from numerical laboratory test results.

Mixed tasks.

- contraceptive-method-choice: Predicts contraceptive method categories, including no-use, short-term, and long-term use, from socio-demographic features.
- entrepreneur-competency: Predicts entrepreneurial competency from personal profile attributes and psychometric ratings.
- travel-insurance: Predicts whether travel insurance is purchased from demographic and travel-history attributes.

Ordinal tasks.

- car-evaluation: Predicts car acceptability on a four-level ordinal scale.
- hotel-rating: Predicts whether a hotel rating exceeds a predefined threshold.

Symbolic tasks.

- tic-tac-toe-endgame: Predicts game outcomes from symbolic board configurations.
- driving-championship-points: Predicts whether season points exceed a predefined threshold from structured racing metadata.
- manifold-orientability: Predicts whether a manifold is orientable from symbolic topological invariants.

B Structural Causal Models Foundation

Structural Causal Models (SCMs) provide a principled language for describing how variables are generated through structural mechanisms, where directed edges represent assumed structural dependencies rather than symmetric statistical correlations (Pearl, 2009; Peters et al., 2017). Following the SCM view in Section 3.2, we regard the observed tabular sample X_i as being generated from decision-relevant evidence S_i and nuisance factors U_i . Since directly modeling $P_\theta(Y_i | X_i)$ may exploit unstable shortcuts induced by U_i , FAG organizes the explanation-supported reasoning path

$$S_i \rightarrow R_i \rightarrow Z_i \rightarrow Y_i.$$

Here, attribute-level explanations $E_{\text{attr}}(\mathcal{T})$ guide the construction of rule factors R_i , while label-level explanations $E_{\text{label}}(\mathcal{Y})$ guide the alignment between aggregated evidence Z_i and the target label Y_i . Therefore, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an SCM-inspired structural prior graph: edges in $\mathcal{E}$ denote explanation-supported reasoning dependencies, rather than statistically identified causal effects.

Principle 1: Explanation-supported Common Cause Principle.

Reichenbach’s Common Cause Principle states that statistical dependence among variables can be explained by a shared cause or common mechanism (Reichenbach, 1956). Inspired by this principle, when multiple observable attribute-value elements provide evidence for the same decision-relevant factor in textual explanations, FAG introduces an intermediate rule node in $\mathcal{V}_{\text{rule}}$ to summarize their shared contribution. In the

SCM view, this corresponds to using explanations to organize the relevant part of S_i into rule factors R_i , rather than directly connecting all observed input nodes in $\mathcal{V}_{\text{in}}$ to the prediction target. Thus, the common cause principle is used as an explanation-level structural prior: it motivates grouping related evidence through shared rule factors, but does not imply that the introduced rule nodes are statistically verified latent causes.

Principle 2: Independent Causal Mechanisms Principle.

The Independent Causal Mechanisms principle suggests that different causal mechanisms should be modeled as autonomous and modular components whenever possible (Schölkopf et al., 2012; Peters et al., 2017). Following this principle, FAG represents different explanation-supported reasoning factors as separate rule nodes in $\mathcal{V}_{\text{rule}}$, and connects them to convergence nodes in $\mathcal{V}_{\text{conv}}$ only when such relations are supported by $E_{\text{attr}}(\mathcal{T})$ or the provided textual explanations. This modular structure avoids entangling unrelated decision mechanisms in a single opaque representation (Zhao et al., 2024), while supporting reasoning from S_i to R_i and ultimately to Z_i . Accordingly, FAG provides an SCM-inspired organization of explanation-supported structural mechanisms.

C Implementation Details of Feature Attribute Graph Construction

This appendix provides implementation details for constructing the feature attribute graph. While Section 3.2 introduces the explanation-to-graph modeling process, this appendix specifies the graph construction procedure, node types, edge definitions, structural prior weights, validation rules, and the LLM prompt template used for reproducibility. The FAG is constructed from the task schema and attribute-level explanations $E_{\text{attr}}(\mathcal{T})$. It serves as an explanation-supported structural prior for differentiable reasoning, rather than a statistically discovered causal graph.

C.1 LLM-based FAG Construction Procedure

Given a task schema and dataset-provided attribute-level explanations, we construct a task-level graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Building on the insight that inference facilitates conceptual understanding (Li et al., 2024) and the potential of LLM-based decomposition and prior knowledge (Yan et al., 2025a; Su et al., 2026), we use LLaMA-3-70B to derive structural priors from textual explanations. The LLM receives two inputs: (1) observable attributes and

possible values, and (2) attribute-level explanations describing decision logic. It does not predict labels but parses explanations into graph components.

The construction process contains three steps. First, *semantic node extraction* identifies graph nodes corresponding to observable attributes, explanation-derived rule factors, and convergence factors. Second, *relation specification* extracts directed explanation-supported relations between nodes and assigns structural prior weights. Third, *structural refinement* removes unsupported or ill-formed nodes and edges, and enforces a valid reasoning topology. The constructed FAG is shared across samples within the same task, while node states are initialized in a sample-conditioned manner during graph-based reasoning.

C.2 Node Types

The node set $\mathcal{V}$ is decomposed into three disjoint subsets:

$$\mathcal{V} = \mathcal{V}_{\text{in}} \dot{\cup} \mathcal{V}_{\text{rule}} \dot{\cup} \mathcal{V}_{\text{conv}}. \quad (15)$$

Input nodes. Input nodes $\mathcal{V}_{\text{in}}$ correspond to observable attributes in the task schema, such as *age*, *annual income*, *odor*, *blood pressure*, or *education level*. Each input node must be grounded in an explicit dataset attribute. For each sample x_i , its relevance is computed using the corresponding sample representation $\mathbf{h}_x^{(i)}$ and the node prototype.

Rule nodes. Rule nodes $\mathcal{V}_{\text{rule}}$ represent intermediate decision factors derived from attribute-level explanations $E_{\text{attr}}(\mathcal{T})$. They summarize higher-level semantic concepts or reasoning rules supported by multiple attributes, such as *travel risk*, *health condition*, *financial capability*, or *eligibility*. These nodes bridge raw attribute evidence and task-level decision evidence.

Convergence nodes. Convergence nodes $\mathcal{V}_{\text{conv}}$ aggregate multiple reasoning paths before prediction. They summarize global evidence produced by rule nodes and provide graph-level evidence for subsequent label alignment. Candidate labels are not represented as FAG nodes; instead, they are encoded separately as label-level semantic prototypes from $E_{\text{label}}(\mathcal{Y})$.

C.3 Edge Construction and Structural Prior Weights

The edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains directed explanation-supported reasoning dependencies. An edge $(v_p, v_q) \in \mathcal{E}$ is added only when the provided explanations explicitly state or strongly imply that

source node v_p contributes to target node v_q . The edge direction follows the reasoning flow from input evidence to rule factors and then to convergence evidence:

$$\mathcal{V}_{\text{in}} \rightarrow \mathcal{V}_{\text{rule}} \rightarrow \mathcal{V}_{\text{conv}}. \quad (16)$$

For example, an explanation such as “*frequent flyers with annual income over 1M usually take travel insurance*” supports directed edges from *annual income* and *frequent flyer* to an intermediate rule node such as *high-income frequent flyer*, which can further connect to a convergence node summarizing travel-insurance evidence.

Each edge $(v_p, v_q) \in \mathcal{E}$ is assigned a structural prior weight $w_{pq} \in [0, 1]$ according to the strength of textual support:

$$w_{pq} = \begin{cases} 0.85-0.95, & \text{explicit relation,} \\ 0.65-0.80, & \text{strong prior,} \\ 0.50-0.65, & \text{weak support.} \end{cases} \quad (17)$$

If the LLM outputs a confidence score within the corresponding interval, we directly use the score. Otherwise, we use the midpoint of the interval, namely 0.90, 0.72, and 0.57 for explicit, strong-prior, and weak-support relations, respectively. These weights serve as structural priors in graph propagation and are incorporated into the propagation coefficient $\lambda_{jr}^{(l,i)}$ in Section 3.3.

C.4 Structural Validation

To reduce unsupported or noisy graph structures, we apply deterministic validation after LLM generation. First, *schema consistency* requires every input node to correspond to an attribute in the task schema. Second, *explanation support checking* removes edges that cannot be traced to the provided textual explanations. Third, *direction validation* enforces the intended reasoning order from input nodes to rule nodes and then to convergence nodes. Direct input-to-convergence edges are allowed only when explicitly supported by the explanations. Fourth, *acyclicity checking* ensures that $\mathcal{G}$ is a directed acyclic graph; when a cycle is detected, the lowest-confidence edge in the cycle is removed. Finally, near-duplicate nodes are merged before graph propagation.

These validation steps ensure that the generated FAG is faithful to the provided explanations, structurally well-formed, and suitable for differentiable message passing.

C.5 Prompt Template

We use a standardized prompt to guide the LLM to construct FAGs from the task schema and dataset-provided attribute-level explanations. The prompt follows the three-stage procedure described above: semantic node extraction, relation specification, and structural refinement. The full prompt is shown in Figure 10.

D Prompt for Explanation Usage

This appendix presents the prompt templates used in our experiments for two settings: inference without dataset-provided explanations and inference with dataset-provided explanations. As shown in Figure 11, the two prompts share the same input data, candidate labels, and output constraints, while differing only in whether domain knowledge explanations are provided to the model. This design ensures a fair and reproducible comparison between explanation-free inference and explanation-enhanced inference.

E Faithfulness Evaluation

We elaborate on the technical implementation of the perturbation-based protocol utilized to quantify the faithfulness of the learned explanations.

Evaluation Protocol. To quantify whether the learned importance scores reflect the true decision-making process, we perform a deletion test (also known as comprehensiveness) (Wang et al., 2022; DeYoung et al., 2020; Ying et al., 2019) by progressively masking input features.

Masking Strategy. Given an input graph $\mathcal{G}$ and the learned importance scores $\mathcal{S}$, we employ two masking policies for comparison:

- **Top- K Strategy:** We rank all nodes based on $\mathcal{S}$ and mask the features of the top- K nodes. To minimize the impact on graph connectivity while removing semantic information, we replace the feature vectors of selected nodes with the mean feature vector of the dataset (Mean-Masking).
- **Random Baseline:** We randomly select K nodes from the input to mask. To ensure statistical reliability, this random selection is repeated 5 times with different random seeds, and the average performance is reported.

F Detailed Sensitivity Analysis of Fusion Coefficient α

Figure 12 provides the complete sensitivity analysis of the fusion coefficient α across representative unseen tasks, where $\alpha = 0$ denotes purely semantic prediction and $\alpha = 1$ denotes purely graph-based reasoning.

Symbolic and Ordinal Tasks. Datasets in these categories, such as *car-evaluation* and *tic-tac-toe*, exhibit a distinct preference for higher α values. For instance, on the *car-evaluation* task, accuracy surges from 37% (at $\alpha = 0$) to 70% (at $\alpha = 0.9$). This trend suggests that purely semantic encoders struggle to infer strict logical rules (e.g., hierarchical buying criteria) from flattened text. In these cases, the GNN component is critical for propagating structural constraints, justifying a high fusion weight.

Numerical Tasks. Similarly, numerical tasks show substantial gains from structural integration. A prime example is *banknote-authentication*, where performance jumps from 29% with pure text to 56% with graph integration ($\alpha = 0.7$). This indicates that while PLMs are robust in processing text, they often fail to capture the distribution nuances of continuous numerical features.

Semantic and Mixed Tasks. In contrast, semantic tasks (e.g., *award-nomination*) and mixed-type tasks (e.g., *travel-insurance*) often achieve peak performance at lower to mid-range α values (typically $\alpha \approx 0.2$ – 0.4). For these domains, the rich semantic priors from the pre-trained encoder are the primary signal. Excessive structural injection ($\alpha \rightarrow 1$) may introduce noise or over-smooth the specific semantic context. MULEG’s adaptive mechanism successfully identifies this need, assigning lower α values to preserve semantic specificity while using the graph to resolve ambiguities.

G Qualitative Case Study

To qualitatively validate the reasoning mechanism of MULEG, we analyze representative test instances from the *Travel Insurance* task where surface-level correlations mislead standard baselines. As illustrated in Figure 14, **Case A** presents a high-income applicant (1.7M), a feature statistically correlated with purchasing insurance. Consequently, the baseline falls into this shortcut trap (red dashed path) and predicts a false positive. In contrast, MULEG correctly predicts *No* by activat-

ing a structured causal chain (blue solid path): it first aggregates demographics into a *Travel Profile* ($w = 0.29$) and subsequently identifies a latent risk factor *Chronic Condition* ($w = 0.08$) which decisively overrides the income bias. Similarly, in **Case B**, MULEG successfully discerns the high cost associated with a large family size (7 members), avoiding the baseline’s error. These cases demonstrate that MULEG’s performance gains stem from interpreting underlying domain logic via intermediate reasoning steps rather than fitting spurious correlations.

H Effect of GNN Depth on Performance

We investigate the influence of GNN depth by varying the number of message-passing layers from 1 to 5. Figure 13 summarizes the test accuracy across datasets using a heatmap to facilitate the identification of layer-wise trends.

Empirical results indicate that GNNs with moderate depth (2–4 layers) consistently achieve optimal or near-optimal performance. Conversely, shallow (1-layer) or excessively deep (5-layer) configurations often lead to performance degradation, reflecting a trade-off between relational modeling capacity and the risk of over-smoothing. Specifically, a 2-layer GNN suffices for tasks with simpler dependencies (e.g., banknote-auth, election), whereas tasks involving complex compositional patterns (e.g., car-eval, soccer-club) benefit from deeper architectures (3–4 layers). The consistent underperformance of 5-layer models suggests that excessive message passing can lead to noise accumulation, diluting discriminative signals.

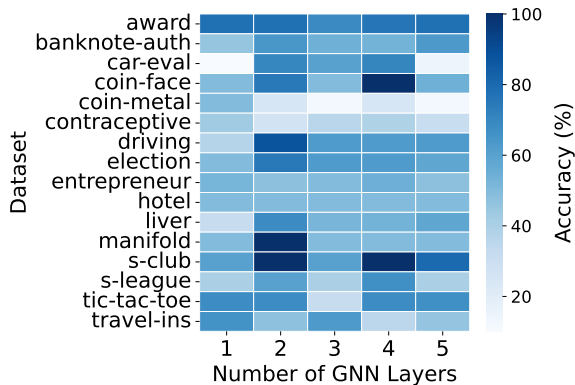


Figure 13: Accuracy under different GNN depths.

I Implementation Details

We provide the complete implementation details to facilitate reproducibility.

Hyperparameter	Value
<i>Optimization</i>	
Backbone Learning Rate	8×10^{-6}
Head Learning Rate	4×10^{-5}
Weight Decay	0.03
Effective Batch Size	128
Max Epochs	10
Gradient Accumulation	4
<i>Loss Function</i>	
Focal Loss α	0.25
Focal Loss γ	2.0
Ordinal Loss α	0.5
Ordinal Loss β	0.2
<i>Task-Adaptive Ranges</i>	
Top-K Nodes (K)	[1, 12]
Temperature (τ)	[0.07, 1.0]
Fusion Coefficient (α)	[0.0, 1.0]
<i>Evaluation</i>	
Random Seeds	6
Seed Offset	42
Computing Infrastructure	NVIDIA V100 (32GB)

Table 5: Detailed Hyperparameter Settings for MULEG.

I.1 Model Architecture

Our framework is implemented using PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020). We utilize bert-base-uncased (Devlin et al., 2019) as the text encoder backbone. For tasks containing primarily numerical features (more than 5 numeric columns), we employ a hybrid encoder with a cross-attention module to fuse numerical embeddings with textual representations. The causal reasoning component is instantiated as a 2-layer graph neural network initialized with BERT embeddings of attribute explanations.

I.2 Training Protocol

We train the model for 10 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear learning rate scheduler (no warmup). To balance the stability of the pre-trained backbone and the learning capacity of new modules, we adopt a differential learning rate strategy:

- Backbone learning rate: 8×10^{-6} (scaled by 0.2 from the base rate).
- Head & GNN learning rate: 4×10^{-5} (base rate).

We use a batch size of 32 per device with gradient accumulation steps of 4, resulting in an effective batch size of 128. To address class imbalance and the ordinal nature of specific tasks, we integrate

Dataset	Samples	Feat.	Task
Bank	45,211	16	Binary (2)
Adult Income	48,842	14	Binary (2)
California	20,640	9	Multi-class (5)
Heart	1,025	13	Binary (2)
Blood	748	4	Binary (2)
Airbnb Listing	103,000	26	Multi-class (3)
Mercari Price	148,200	8	Multi-class (11)

Table 6: Statistics of the datasets used in the supplementary experiments.

Focal Loss ($\gamma = 2.0$) and a specialized Ordinal Loss into the training objective.

I.3 Task-Adaptive Hyperparameters

A key feature of our method is the task-adaptive mechanism, where hyperparameters are dynamically adjusted based on task characteristics (e.g., task difficulty and data type). Table 5 summarizes the specific settings. Crucially, the fusion coefficient α , temperature τ , and retrieval count K vary across tasks to optimize performance.

I.4 Computing Infrastructure

All experiments are conducted on a single NVIDIA Tesla V100 GPU (32GB). Under these settings, training on the full benchmark typically completes within 4 hours. We report the mean performance over six independent runs to ensure statistical significance. Tables 7 and 8 present the corresponding results on the six large-scale and ten small-scale unseen datasets, respectively.

J Supplementary Experiments on Larger and Multiclass Datasets

Referring to the dataset selection of (Shi et al., 2025) and (Hegselmann et al., 2023), we conduct supplementary zero-shot experiments on seven widely adopted benchmark datasets with varying scales, feature dimensions, and class structures. In addition to binary classification datasets, we include three multiclass datasets to further evaluate the generalization ability of MULEG under more complex label spaces. Specifically, the datasets include Bank (45K samples, 16 features, 2 classes), Adult Income (48K samples, 14 features, 2 classes), California (20K samples, 9 features, 5 classes), Heart (1K samples, 13 features, 2 classes), Blood (748 samples, 4 features, 2 classes), Airbnb Listing (103K samples, 26 features, 3 classes), and Mercari Price Suggestion (148.2K samples, 8 features, 11 classes). Detailed statistics are reported in Table 6.

K Use of AI Assistants

We use ChatGPT to improve the presentations of this paper¹.

¹<https://chatgpt.com/>

Method	banknote (Num)	liver (Num)	car-eval (Ord)	contra. (Mix)	travel (Mix)	tic-tac (Sym)	Avg.
RoBERTa (No Exp)	55.3 $\pm$ 1.2	32.6 $\pm$ 2.1	22.4 $\pm$ 1.8	43.5 $\pm$ 1.5	35.8 $\pm$ 2.4	32.3 $\pm$ 1.7	37.0 $\pm$ 0.8
RoBERTa (Concat)	59.7 $\pm$ 1.4	37.6 $\pm$ 1.9	23.1 $\pm$ 2.0	45.6 $\pm$ 1.7	40.5 $\pm$ 2.1	30.6 $\pm$ 1.5	39.5 $\pm$ 0.7
RoBERTa (Sim)	56.7 $\pm$ 1.1	67.8 $\pm$ 1.5	19.1 $\pm$ 1.3	23.4 $\pm$ 2.2	34.2 $\pm$ 1.8	32.3 $\pm$ 1.9	38.9 $\pm$ 0.7
T0-3B (No Exp)	55.3 $\pm$ 2.3	32.2 $\pm$ 2.0	20.8 $\pm$ 1.6	46.4 $\pm$ 1.9	57.0 $\pm$ 2.5	67.7 $\pm$ 0.0	46.6 $\pm$ 0.8
T0-3B (Exp)	55.3 $\pm$ 2.1	32.2 $\pm$ 1.8	22.3 $\pm$ 2.1	23.4 $\pm$ 2.6	45.8 $\pm$ 2.2	67.7 $\pm$ 0.0	41.1 $\pm$ 0.8
FLAN-T5 (No Exp)	55.3 $\pm$ 1.8	31.3 $\pm$ 2.4	22.3 $\pm$ 1.9	43.4 $\pm$ 1.5	65.8 $\pm$ 1.2	67.2 $\pm$ 1.1	47.5 $\pm$ 0.7
FLAN-T5 (Exp)	29.1 $\pm$ 3.1	32.2 $\pm$ 2.2	22.3 $\pm$ 2.5	43.4 $\pm$ 1.8	65.8 $\pm$ 1.6	67.7 $\pm$ 0.0	43.4 $\pm$ 0.9
LLaMA-3 (No Exp)	44.7 $\pm$ 2.6	67.8 $\pm$ 1.4	61.1 $\pm$ 2.1	29.2 $\pm$ 2.8	34.2 $\pm$ 2.4	58.9 $\pm$ 1.9	49.3 $\pm$ 0.9
LLaMA-3 (Exp)	52.4 $\pm$ 2.4	67.8 $\pm$ 1.5	70.8 $\pm$ 1.7	38.0 $\pm$ 2.2	34.2 $\pm$ 2.6	67.7 $\pm$ 0.0	55.1 $\pm$ 0.8
ExEnt	44.7 $\pm$ 1.7	35.7 $\pm$ 2.1	57.2 $\pm$ 1.8	21.3 $\pm$ 2.5	45.6 $\pm$ 1.9	32.4 $\pm$ 2.2	39.5 $\pm$ 0.9
TALC	49.7 $\pm$ 1.5	47.8 $\pm$ 1.6	16.5 $\pm$ 1.4	43.8 $\pm$ 1.7	34.2 $\pm$ 2.0	47.8 $\pm$ 1.8	40.0 $\pm$ 0.7
CLORE	53.8 $\pm$ 1.4	49.6 $\pm$ 1.9	66.5 $\pm$ 1.5	39.0 $\pm$ 2.1	46.8 $\pm$ 1.8	67.7 $\pm$ 0.0	53.9 $\pm$ 0.7
MULEG (Ours)	65.8 $\pm$ 0.8	68.3 $\pm$ 1.1	71.7 $\pm$ 1.0	46.8 $\pm$ 1.2	62.3 $\pm$ 1.4	67.7 $\pm$ 0.0	63.8 $\pm$ 0.4

Table 7: Zero-shot classification accuracy (%) on six large-scale unseen datasets. Results are reported as mean $\pm$ standard deviation over six independent runs.

Method	award (Sem)	coin-f (Sem)	coin-m (Sem)	elect. (Sem)	s-club (Sem)	s-lg (Sem)	hotel (Ord)	entr. (Mix)	driv. (Sym)	mani. (Sym)	Avg.
RoBERTa (No Exp)	66.7 $\pm$ 3.1	50.0 $\pm$ 4.2	50.0 $\pm$ 3.5	50.0 $\pm$ 4.0	40.0 $\pm$ 3.8	60.0 $\pm$ 2.9	50.0 $\pm$ 3.1	43.6 $\pm$ 2.8	37.6 $\pm$ 3.6	50.0 $\pm$ 4.1	49.8 $\pm$ 1.1
RoBERTa (Concat)	70.7 $\pm$ 2.8	50.0 $\pm$ 3.9	48.3 $\pm$ 4.1	50.0 $\pm$ 3.5	60.0 $\pm$ 3.2	40.0 $\pm$ 3.6	50.0 $\pm$ 2.8	49.3 $\pm$ 2.4	37.8 $\pm$ 3.9	50.0 $\pm$ 3.8	50.6 $\pm$ 1.0
RoBERTa (Sim)	22.2 $\pm$ 4.5	50.0 $\pm$ 4.1	50.0 $\pm$ 3.8	50.0 $\pm$ 3.6	60.0 $\pm$ 4.0	40.0 $\pm$ 3.9	50.0 $\pm$ 3.2	47.7 $\pm$ 2.5	62.5 $\pm$ 3.5	50.0 $\pm$ 4.0	48.2 $\pm$ 1.2
T0-3B (No Exp)	72.2 $\pm$ 3.5	25.0 $\pm$ 5.2	50.0 $\pm$ 4.5	12.5 $\pm$ 4.8	60.0 $\pm$ 3.9	60.0 $\pm$ 3.5	25.0 $\pm$ 4.1	52.3 $\pm$ 3.1	12.5 $\pm$ 4.8	50.0 $\pm$ 4.5	42.0 $\pm$ 1.3
T0-3B (Exp)	66.7 $\pm$ 3.8	25.0 $\pm$ 4.9	50.0 $\pm$ 4.2	12.5 $\pm$ 4.5	40.0 $\pm$ 4.2	40.0 $\pm$ 4.5	25.0 $\pm$ 4.6	52.3 $\pm$ 2.9	12.5 $\pm$ 4.5	50.0 $\pm$ 4.8	37.4 $\pm$ 1.3
FLAN-T5 (No Exp)	61.1 $\pm$ 4.1	50.0 $\pm$ 4.5	50.0 $\pm$ 3.9	50.0 $\pm$ 4.2	40.0 $\pm$ 4.1	60.0 $\pm$ 3.8	25.0 $\pm$ 5.1	47.7 $\pm$ 3.5	62.5 $\pm$ 3.8	50.0 $\pm$ 4.2	49.6 $\pm$ 1.3
FLAN-T5 (Exp)	55.6 $\pm$ 4.5	50.0 $\pm$ 4.8	50.0 $\pm$ 4.4	50.0 $\pm$ 4.5	40.0 $\pm$ 4.6	60.0 $\pm$ 4.2	50.0 $\pm$ 3.9	52.3 $\pm$ 3.4	62.5 $\pm$ 4.1	50.0 $\pm$ 4.6	52.0 $\pm$ 1.3
LLaMA-3 (No Exp)	50.0 $\pm$ 3.9	50.0 $\pm$ 4.1	75.0 $\pm$ 3.2	50.0 $\pm$ 3.8	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	50.0 $\pm$ 4.5	47.7 $\pm$ 3.2	62.5 $\pm$ 3.5	50.0 $\pm$ 3.9	63.5 $\pm$ 1.1
LLaMA-3 (Exp)	22.2 $\pm$ 5.1	75.0 $\pm$ 3.6	50.0 $\pm$ 4.2	62.5 $\pm$ 4.1	100.0 $\pm$ 0.0	60.0 $\pm$ 4.5	50.0 $\pm$ 3.8	47.7 $\pm$ 3.5	62.5 $\pm$ 4.0	100.0 $\pm$ 0.0	63.0 $\pm$ 1.2
ExEnt	67.8 $\pm$ 2.8	75.0 $\pm$ 3.1	50.0 $\pm$ 3.8	50.0 $\pm$ 3.5	60.0 $\pm$ 3.4	60.0 $\pm$ 3.2	50.0 $\pm$ 2.9	45.5 $\pm$ 2.4	37.5 $\pm$ 3.5	100.0 $\pm$ 0.0	59.6 $\pm$ 1.0
CLORE	22.2 $\pm$ 3.5	50.0 $\pm$ 4.5	50.0 $\pm$ 4.2	50.0 $\pm$ 4.1	60.0 $\pm$ 3.6	40.0 $\pm$ 4.2	50.0 $\pm$ 3.5	45.5 $\pm$ 2.8	37.5 $\pm$ 4.1	100.0 $\pm$ 0.0	50.5 $\pm$ 1.1
MULEG (Ours)	77.8 $\pm$ 1.5	75.0 $\pm$ 2.1	75.0 $\pm$ 2.0	75.0 $\pm$ 1.8	100.0 $\pm$ 0.0	60.0 $\pm$ 2.5	83.3 $\pm$ 1.6	50.0 $\pm$ 1.9	95.8 $\pm$ 1.1	100.0 $\pm$ 0.0	79.2 $\pm$ 0.6

Table 8: Zero-shot classification accuracy (%) on ten small-scale unseen datasets. Results are reported as mean $\pm$ standard deviation over six independent runs.

Prompt: Feature Attribute Graph Generation with Explanations

Task Definition: You are given a structured-data classification task with: (1) a set of observable attributes with names and possible values; (2) natural-language explanations describing how decisions are made; (3) candidate labels for the classification task.

Your goal is to construct a task-level Feature Attribute Graph (FAG) that formalizes the explanation-derived reasoning process.

Important:

- Use only the provided attributes and explanations.
- Do not introduce unsupported external information.
- Do not predict labels.
- Return only a JSON object.

Step 1: Semantic Node Extraction Identify a set of semantic nodes V . Each node must represent a single, interpretable factor. Categorize each node into one of the following types:

- **Input Node:** Directly corresponds to an observable attribute in the task schema. Use the original attribute name whenever possible.
- **Intermediate Rule Node:** Represents an abstract rule, logical implication, or intermediate decision factor inferred from the explanations.
- **Convergence Node:** Aggregates multiple reasoning paths before prediction. It should summarize decision-relevant evidence, but it should not directly duplicate a class label.

Each node must include:

- id: a unique identifier;
- name: a concise node name;
- type: input, intermediate_rule, or convergence;
- text: a brief natural-language description of its meaning.

Step 2: Causal Relation Specification Construct a set of directed edges B . Add an edge only if the provided explanations support a relation from the source node to the target node. An edge can be added when:

- an explanation explicitly states that one factor affects or determines another;
- an attribute contributes to an intermediate rule or decision factor;
- multiple attributes jointly support a higher-level conclusion.

Assign a confidence weight w in $[0,1]$:

- **0.85–0.95:** explicit causal or logical relation;
- **0.65–0.80:** strong domain-prior relation;
- **0.50–0.65:** weak or indirect supporting relation.

Each edge must include:

- src: source node id;
- dst: target node id;
- w: confidence weight;
- rationale: a short explanation of why the edge is supported.

Step 3: Structural Refinement and Validation Ensure that:

- all input nodes correspond to provided attributes;
- edges are supported by the provided explanations;
- edges flow from input nodes to intermediate rule nodes to convergence nodes;
- the graph is acyclic;
- redundant nodes are merged;
- no unsupported external information is introduced.

Output Format: Return the FAG in JSON format:

```
{
  "nodes": [ {"id": "...", "name": "...", "type": "...", "text": "..."} ],
  "edges": [ {"src": "...", "dst": "...", "w": ..., "rationale": "..."} ]
}
```

Figure 10: Prompt template for constructing the Feature Attribute Graph from dataset-provided explanations.

Prompt : LLMs No Explanation Tests

System: You are an expert classifier for tabular data.
Task: Your task is to classify the given data instance into exactly one of the provided candidate labels.
Input Data: f"{data_str}"
Candidate Labels: f"{labels_str}"
Instructions:
1. Analyze the input data features.
2. Select the most appropriate label from the "Candidate Labels" list.
3. Output ONLY the label name. Do not output any explanation, reasoning, or punctuation.
Prediction:

Prompt : LLMs Use Explanation Tests

System: You are an expert classifier for tabular data.
Task: Your task is to classify the given data instance into exactly one of the provided candidate labels, guided by the provided domain explanations.
Domain Knowledge & Explanations: f"{explanation_text}"
Input Data: f"{data_str}"
Candidate Labels: f"{labels_str}"
Instructions:
1. Read the "Domain Knowledge" to understand the decision logic.
2. Analyze the "Input Data" features against these rules.
3. Select the most appropriate label from the "Candidate Labels" list.
4. Output ONLY the label name. Do not output any explanation, reasoning, or punctuation.
Prediction:

Figure 11: Prompt templates for LLM-based zero-shot classification under explanation-free and explanation-enhanced settings.

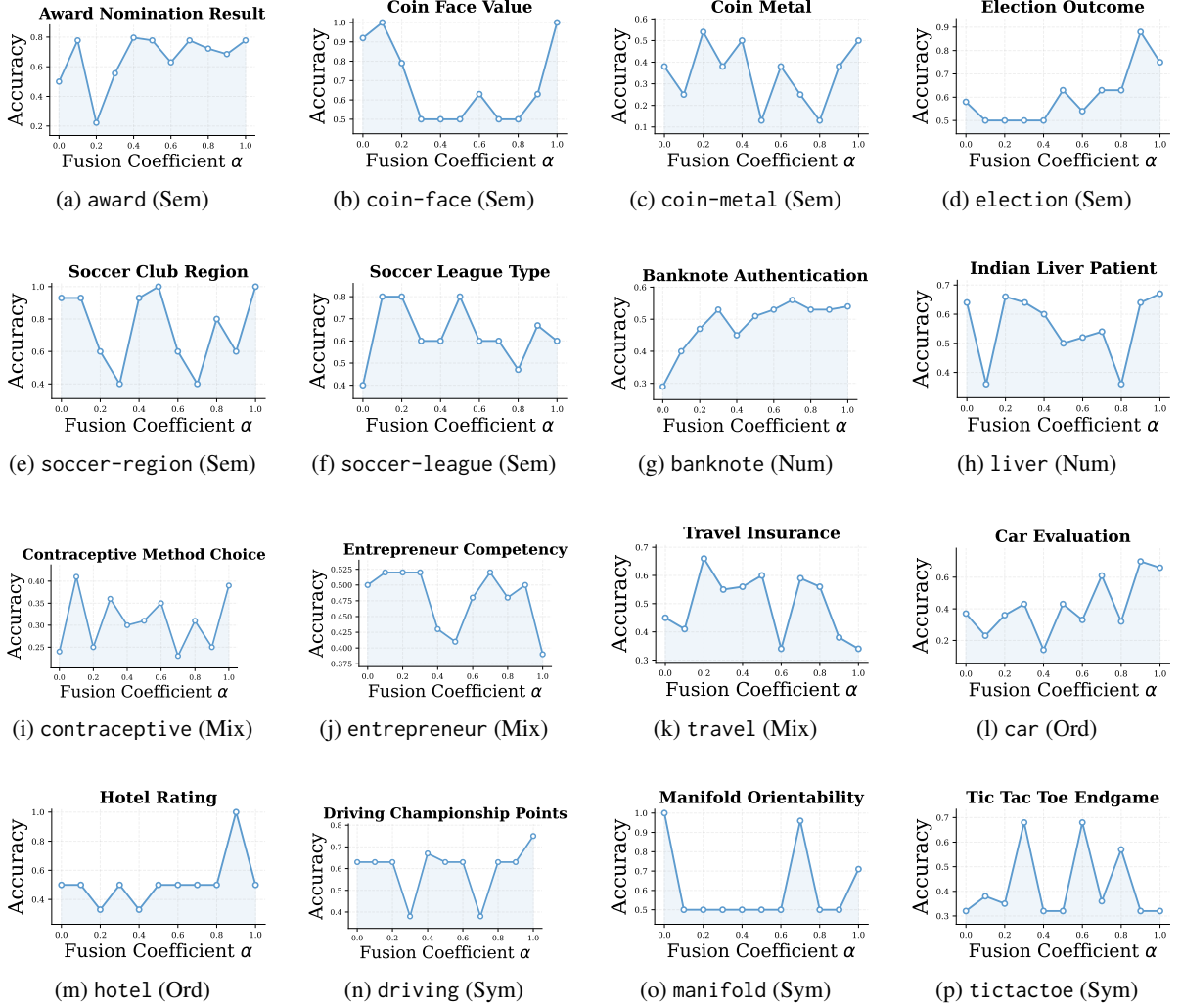


Figure 12: Sensitivity analysis of the fusion coefficient α on representative unseen tasks, grouped by task type. Each subplot reports classification accuracy as a function of α , where $\alpha = 0$ relies solely on textual semantics and $\alpha = 1$ relies solely on graph-based reasoning.

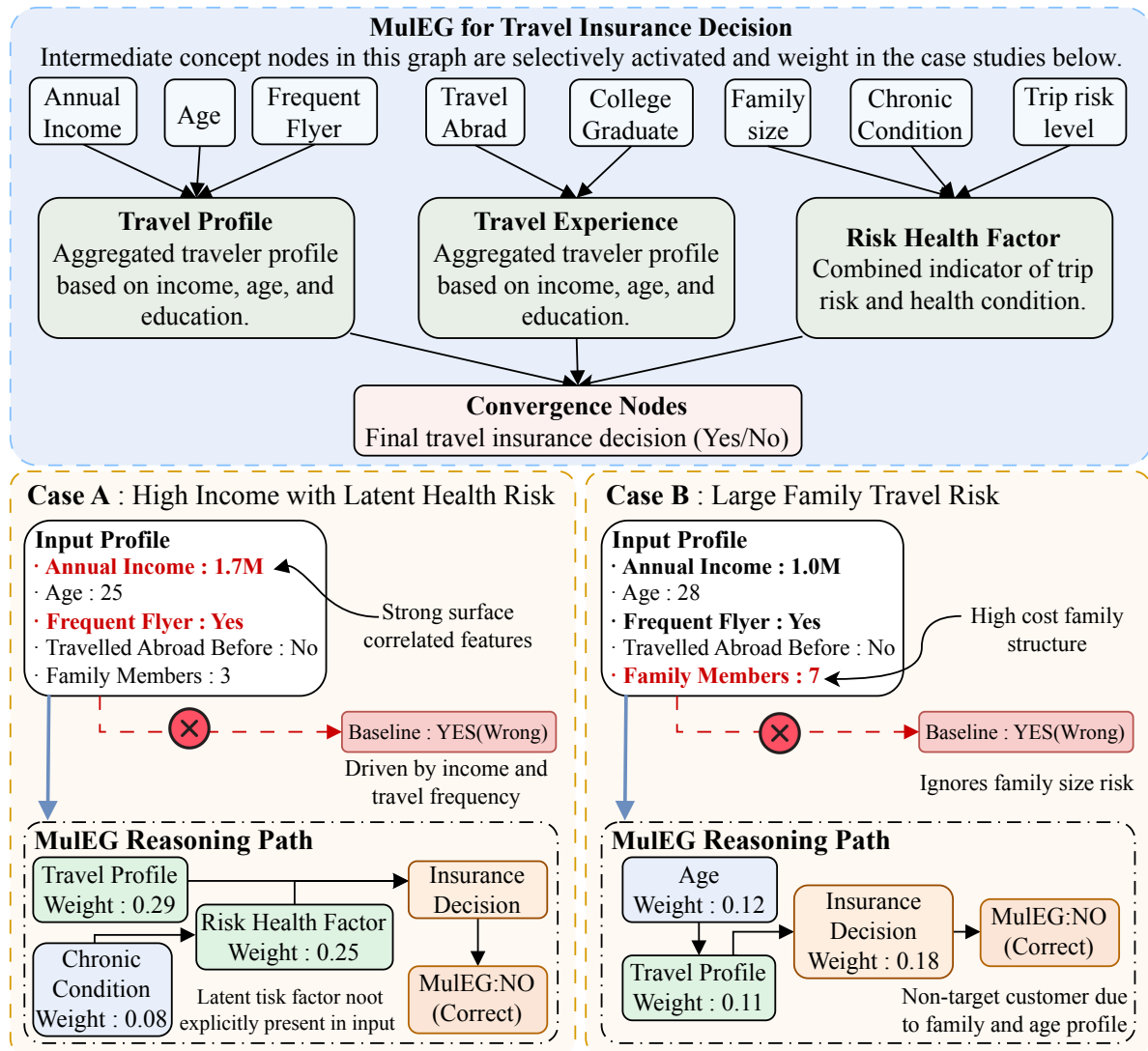


Figure 14: Qualitative comparison of reasoning paths on the Travel Insurance task. While the baseline model relies on spurious surface correlations (e.g., high income) to incorrectly predict *Yes*, MULEG activates intermediate causal concepts such as identifying latent *Chronic Conditions* (Case A) or *High-Cost Family Structure* (Case B) to correctly derive the *No* decision. This validates that MULEG prioritizes structured causal logic over statistical shortcuts.