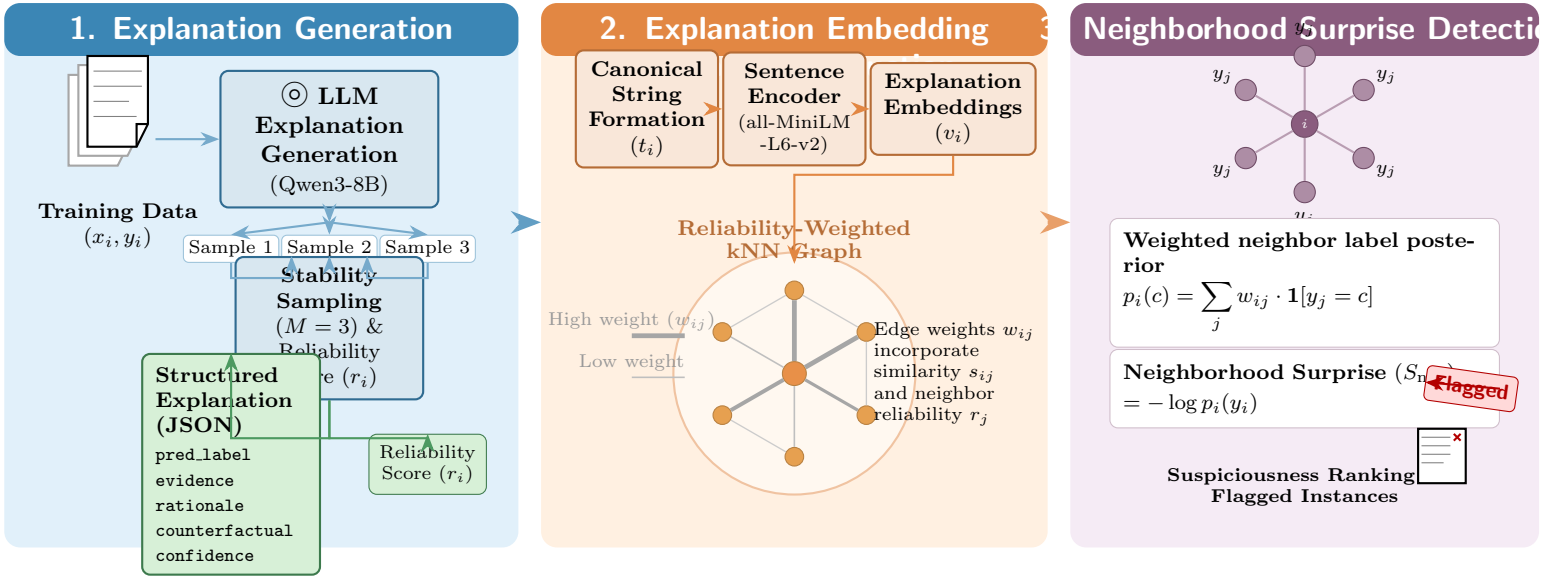
1. Explanation Generation
Training Data
(x_i, y_i)
LLM Explanation Generation
(Qwen3-8B)
Sample 1
Sample 2
Sample 3
Stability Sampling
$(M = 3)$ & Reliability
(r_i)
Structured Explanation (JSON)
pred_label
evidence
rationale
counterfactual
confidence
Reliability Score (r_i)
2. Explanation Embedding
Canonical String Formation
(t_i)
Sentence Encoder
(all-MiniLM -L6-v2)
Explanation Embeddings
(v_i)
Reliability-Weighted kNN Graph
High weight (w_{ij})
Low weight
Edge weights w_{ij} incorporate similarity s_{ij} and neighbor reliability r_j
Neighborhood Surprise Detecti
y_j
y_j
y_j
y_j
i
Weighted neighbor label posterior
$p_i(c) = \sum_j w_{ij} \cdot \mathbf{1}[y_j = c]$
Neighborhood Surprise $(S_n$
$= -\log p_i(y_i)$
Flagged
Suspiciousness Ranking
Flagged Instances