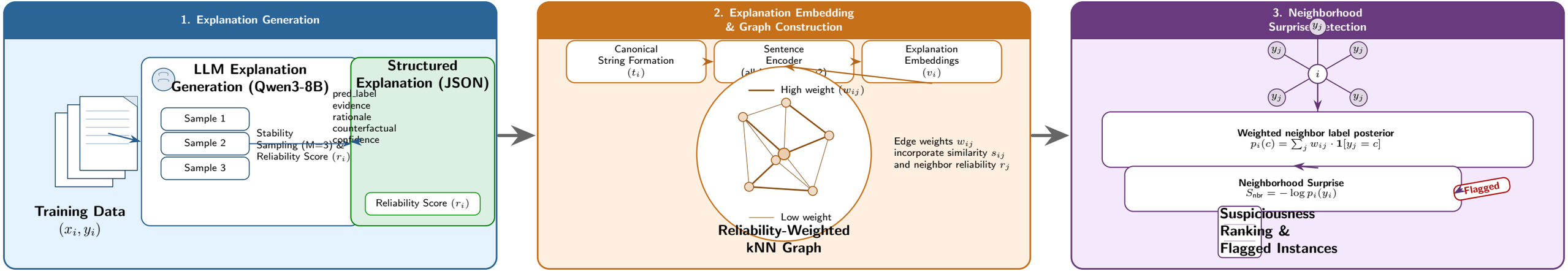
1. Explanation Generation
LLM Explanation Generation (Qwen3-8B)
Sample 1
Sample 2
Sample 3
Stability Sampling (M=3) & Reliability Score (r_i)
Structured Explanation (JSON)
pred_label
evidence
rationale
counterfactual
confidence
Reliability Score (r_i)
Training Data
(x_i, y_i)
2. Explanation Embedding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