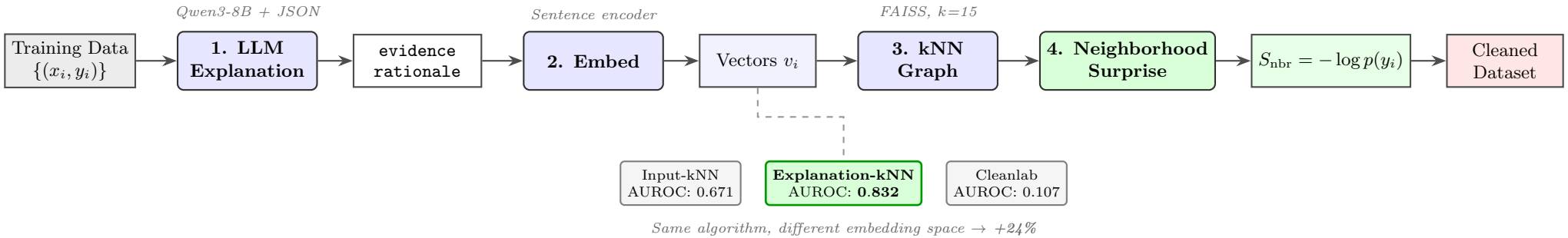
Qwen3-8B + JSON
Sentence encoder
FAISS, k=15
Training Data $\{(x_i, y_i)\}$
1. LLM Explanation
evidence rationale
2. Embed
Vectors v_i
3. kNN Graph
4. Neighborhood Surprise
$S_{\text{nbr}} = -\log p(y_i)$
Cleaned Dataset
Input-kNN AUROC: 0.671
Explanation-kNN AUROC: 0.832
Cleanlab AUROC: 0.107
Same algorithm, different embedding space → +24%