Explanation-Consistency Graphs: Neighborhood Surprise in Explanation Space for Training Data Debugging

Anonymous ACL submission

Abstract

Training data quality is critical for NLP model performance, yet identifying mislabeled examples remains challenging when models confidently fit errors via spurious correlations. Confident learning methods like Cleanlab assume mislabeled examples cause low confidence; however, this assumption breaks down when artifacts enable confident fitting of wrong labels. We propose **Explanation-Consistency Graphs (ECG)**, which detects problematic training instances by computing neighborhood surprise in *explanation embedding space*. Our key insight is that LLM-generated explanations capture "why this label applies," and this semantic content reveals inconsistencies invisible to classifier confidence. Evaluating on SST-2 and MultiNLI across uniform and artifact-aligned noise (5 seeds), we find that ECG achieves 0.819 ± 0.004 AUROC on SST-2 artifact noise (where Cleanlab drops to 0.136 ± 0.025), a 20% relative improvement over the same algorithm on input embeddings (0.683 ± 0.006). On MultiNLI, a complementary LLM-based signal—label mismatch—leads (0.883 ± 0.002) while kNN methods struggle with 3-class structure. Since both signals derive from the same LLM call, practitioners can select the appropriate method at no additional cost.

1 Introduction

The quality of training data fundamentally constrains what NLP models can learn. Large-scale empirical studies reveal label error rates ranging from 0.15% (MNIST) to 5.83% (ImageNet), averaging 3.3% across 10 benchmark test sets (Northcutt et al., 2021b), and these errors propagate into systematic model failures. Beyond simple mislabeling, annotation artifacts and spurious correlations create particularly insidious data quality issues: models learn superficial patterns that happen to correlate with labels in the training set but fail catastrophically under distribution shift (Gururangan et al., 2018; McCoy et al., 2019). Identifying and correcting such problematic instances, known as *training data debugging*, is therefore essential for building reliable NLP systems.

The dominant paradigm for training data debugging relies on model confidence and loss signals. **Confident learning** (Northcutt et al., 2021a) estimates a joint distribution between noisy and true labels using predicted probabilities, effectively identifying instances where the model "disagrees" with the observed label. **Training dynamics** approaches like AUM (Pleiss et al., 2020) and CTRL (Yue and Jha, 2022) track per-example margins and loss trajectories across training epochs, exploiting the observation that mislabeled examples exhibit different learning patterns than clean ones. High-loss filtering with pretrained language models can be surprisingly effective on human-originated noise (Chong et al., 2022). These methods share a common assumption: *problematic examples will cause low confidence or high loss during training*.

This assumption breaks down catastrophically when **models confidently fit errors via spurious correlations**. Consider sentiment data where mislabeled examples happen to contain distinctive tokens such as rating indicators like "[RATING=5]", demographic markers, or formatting artifacts. The classifier learns to predict the *wrong* labels with *high confidence* by exploiting these spurious markers. From a loss perspective, these mislabeled examples look perfectly clean; they are fitted early, with high confidence, and low loss throughout training. Cleanlab's confident joint and AUM's margin trajectories both fail because the model is confident, just confidently wrong for the wrong reasons.

This failure mode is not hypothetical. Poliak et al. (2018) showed that NLI datasets can be partially solved using only the hypothesis, revealing pervasive annotation artifacts. Gururangan et al. (2018) demonstrated that annotation patterns sys-

tematically correlate with labels in ways that models exploit. The spurious correlation literature extensively documents how models learn shortcuts that evade standard diagnostics (Clark et al., 2019; Utama et al., 2020; Tu et al., 2020), and debiasing methods must explicitly model bias structure to mitigate it (Sagawa et al., 2020). When the very mechanism that causes label noise *also* enables confident fitting, confidence-based debugging becomes unreliable.

We propose **Explanation-Consistency Graphs (ECG)**, which detects problematic training instances by computing neighborhood surprise in *explanation embedding space* rather than input embedding space. Our key insight is that *explanations encode semantic information about why a label should apply*, and this "why" content reveals inconsistencies even when classifier confidence does not. When an LLM explains why it believes a sentence has positive sentiment, its rationale and cited evidence reflect the actual semantic content, not spurious markers that the classifier may have learned to exploit. By embedding these explanations and measuring kNN label disagreement, ECG detects mislabeled instances that are invisible to loss and probability signals.

The core idea is simple: if an example's label disagrees with the labels of examples whose *explanations* are most similar, that label is likely wrong. This is the same principle underlying input-based kNN detection (Bahri et al., 2020; Kim et al., 2023), but operating in a fundamentally different representation space. Input embeddings capture "what the text is about"; explanation embeddings capture "why this text has this label." When labels are wrong, the "why" becomes inconsistent with semantically similar examples, making explanation-space neighborhood surprise a powerful detection signal.

ECG synthesizes ideas from three research threads: **(1)** the explanation-based debugging literature, which uses explanations to help humans surface artifacts (Lertvittayakumjorn and Toni, 2021; Lertvittayakumjorn et al., 2020; Lee et al., 2023), but has not automated detection via graph structure; **(2)** graph-based noisy label detection, which uses neighborhood disagreement in representation space (Bahri et al., 2020; Kim et al., 2023; Di Salvo et al., 2025), but over input embeddings; and **(3)** LLM-generated explanations with structured schemas (Geng et al., 2023; Huang et al., 2023), which provide the semantic substrate for our graph.

Concretely, ECG works as follows. **(1) Explanation Generation:** We generate structured JSON explanations for all training instances using an instruction-tuned LLM (Qwen3-8B), enforcing JSON structure via schema-constrained decoding and instructing the model to quote extractive evidence spans. **(2) Explanation Embedding:** We embed explanations using a sentence encoder and construct a kNN graph in this space. **(3) Neighborhood Surprise:** We compute the negative log-probability of each instance's label given its neighbors' labels in explanation space, which serves as our primary detection signal. We also explored additional signals (NLI contradiction, stability, training dynamics), but found that simple kNN surprise in explanation space works best.

Our contributions are:

1. We introduce **Explanation-Consistency Graphs (ECG)**, demonstrating that neighborhood surprise computed in *explanation embedding space* substantially outperforms the same algorithm on input embeddings on SST-2 artifact-aligned noise (0.819 ± 0.004 vs. 0.683 ± 0.006, a 20% relative improvement), evaluated across two datasets (SST-2, MultiNLI) and two noise types with 5 seeds.

2. We establish a **concrete failure mode** for confidence-based cleaning: when artifacts enable confident fitting, Cleanlab achieves only 0.136 ± 0.025 AUROC on SST-2 (worse than random), while ECG achieves 0.819 ± 0.004. On MultiNLI, training-based methods also degrade under artifacts (0.526–0.686).

3. We identify **complementary failure modes**: ECG excels on SST-2 artifacts while LLM Mismatch excels on MultiNLI artifacts (0.883 ± 0.002), yet both derive from the same LLM inference—a single structured explanation yields both signals at no additional cost.

2 Related Work

ECG targets training-data debugging in a regime where spurious correlations let models fit wrong labels *confidently*. It connects to (i) label-error detection from confidence and training dynamics, (ii) graph-based data quality, and (iii) explanation- and attribution-based diagnosis of artifacts. Across these areas, the key gap is a scalable detector whose

signal remains informative when classifier confidence is *not*.

2.1 Label-Error Detection Under Confident Fitting

Most data-cleaning methods rank examples using signals derived from the classifier. **Confident learning** (Northcutt et al., 2021a) identifies likely label errors via disagreement between observed labels and predicted probabilities, and works well when noise manifests as low confidence. Training-dynamics methods similarly treat mislabeled data as hard-to-learn: **AUM** (Pleiss et al., 2020) uses cumulative margins, and **CTRL** (Yue and Jha, 2022) clusters loss trajectories to separate clean from noisy examples. More recently, Kim et al. (2024) track per-example representation dynamics across training to build discriminative features for noise detection, and **NoiseGPT** (Wang et al., 2024) uses probability curvature from LLM logprobs to identify and rectify label errors. For NLP, out-of-sample loss ranking with pretrained language models can be highly effective on human-originated noise (Chong et al., 2022).

Gap. These approaches—including the more recent discriminative dynamics (Kim et al., 2024) and probability-curvature (Wang et al., 2024) methods—share a reliance on training-time difficulty (high loss, low margin, or low confidence). When artifacts make wrong labels easy to fit, mislabeled instances can have *low loss and high confidence* throughout training, rendering confidence- and dynamics-based detectors unreliable. ECG addresses this failure mode by using a signal derived from *explanations* rather than the classifier's fit.

2.2 Graph-Based Data Quality and Neighborhood Disagreement

Graph-based methods detect label errors from representation-space structure, flagging instances whose labels disagree with their nearest neighbors. This principle appears in kNN-based noisy-label detection (Bahri et al., 2020) and scalable relation-graph formulations that jointly model label errors and outliers (Kim et al., 2023). Recent work improves robustness when errors cluster, e.g., reliability-weighted neighbor voting (Di Salvo et al., 2025), and label propagation on kNN graphs when clean anchors exist (Iscen et al., 2020).

Gap. Prior graph-based approaches build neighborhoods over input embeddings or model representations. ECG keeps the same neighborhood-disagreement idea but changes the substrate: it constructs the graph in *explanation embedding space*, where neighbors are defined by similar *label-justifying evidence and rationales*. This shift is crucial in artifact-aligned settings, where input-space similarity can preserve spurious markers rather than the underlying "why" of the label.

2.3 Explanations, Artifacts, and Dataset Debugging

Explanations and attribution have been used extensively for diagnosing dataset artifacts and guiding model fixes. Surveyed "explanation → feedback → fix" pipelines (Lertvittayakumjorn and Toni, 2021) and interactive systems such as **FIND** (Lertvittayakumjorn et al., 2020), explanation-driven label cleaning (Teso et al., 2021), and **XMD** (Lee et al., 2023) support human-in-the-loop debugging. Complementarily, training-set artifact analyses localize influential tokens and examples, e.g., **TFA** (Pezeshkpour et al., 2022) and influence-function based artifact discovery (Han et al., 2020). These tools are motivated by a broad literature on spurious correlations and annotation artifacts, including hypothesis-only shortcuts in NLI and debiasing or counterfactual remedies (Poliak et al., 2018; Belinkov et al., 2019; Clark et al., 2019; Utama et al., 2020; Kaushik et al., 2020).

Gap. Existing explanation-based debugging largely supports *human* discovery or *model* regularization, while spurious-correlation work typically targets mitigation rather than identifying which *training instances* are mislabeled. To our knowledge, ECG is the first to aggregate LLM explanations via graph structure for automated data cleaning, bridging the explanation and data-quality literatures.

LLM-generated explanations. Because ECG relies on structured LLM explanations as a representation, we summarize related work on structured generation and explanation reliability in Appendix C.

3 Method

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ with potentially noisy labels y_i, our goal is to produce a suspiciousness ranking that places mislabeled or artifact-laden instances at the top. ECG achieves this through three stages: explanation generation

(§3.1), explanation embedding and graph construction (§3.2), and neighborhood surprise computation (§3.3). Figure 1 provides an overview. We also explored additional signals (NLI contradiction, stability, training dynamics) but found they did not improve over simple neighborhood surprise; we analyze this in §6 and provide details in Appendix A.

3.1 Structured Explanation Generation

For each training instance x_i, we generate a structured JSON explanation using an instruction-tuned LLM (Qwen3-8B). The explanation contains:

- `pred_label`: The LLM's predicted label
- `evidence`: 1–3 exact substrings from x_i justifying the prediction
- `rationale`: A brief explanation ($\leq$25 tokens) without label words
- `counterfactual`: A minimal change that would flip the label
- `confidence`: Integer 0–100

We enforce schema validity via constrained decoding and instruct the LLM to ignore metadata tokens (e.g., `<lbl_pos>`) so explanations reflect semantic content rather than spurious markers.

Stability Sampling. LLM explanations can be unstable across random seeds. We generate $M = 3$ explanations per instance (one deterministic at temperature 0, two samples at temperature 0.7) and compute a **reliability score**:

$$\rho_i = \frac{1}{3}\left(L_i + E_i + R_i\right) \tag{1}$$

where L_i is label agreement (fraction of samples predicting the same label), E_i is evidence Jaccard (token overlap between evidence spans), and R_i is rationale similarity (cosine similarity of sentence embeddings) across the M samples. High ρ_i indicates stable, reliable explanations; low ρ_i indicates the LLM is uncertain or the instance is ambiguous.

3.2 Reliability-Weighted Graph Construction

We embed explanations and construct a kNN graph that downweights unreliable neighbors, inspired by WANN (Di Salvo et al., 2025).

Explanation Embedding. For each instance, we form a canonical string t_i excluding label information:

$$t_i = \texttt{"Evidence: "} \oplus e_i \oplus \texttt{" | Rationale: "} \oplus r_i \tag{2}$$

where e_i and r_i are the evidence and rationale fields. We embed t_i using a sentence encoder (all-MiniLM-L6-v2) and L_2-normalize to obtain v_i.

Reliability-Weighted Edges. We retrieve the $k = 15$ nearest neighbors $\mathcal{N}(i)$ for each node using FAISS. Edge weights incorporate both similarity and neighbor reliability:

$$\tilde{w}_{ij} = \exp\left(\frac{s_{ij}}{\tau}\right) \cdot \rho_j, \quad w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j' \in \mathcal{N}(i)} \tilde{w}_{ij'}} \tag{3}$$

where $s_{ij} = v_i^\top v_j$ is cosine similarity, $\tau = 0.07$ is a temperature, and ρ_j is neighbor reliability. This ensures that unstable or unreliable neighbors contribute less to inconsistency signals.

Outlier Detection. We compute an outlier score $O_i = 1 - \frac{1}{k}\sum_{j \in \mathcal{N}(i)} s_{ij}$ to distinguish genuinely out-of-distribution examples from mislabeled in-distribution examples.

3.3 Neighborhood Surprise Detection

The core detection signal in ECG is **neighborhood surprise**: if an instance's label disagrees with the labels of instances with similar explanations, the label may be wrong.

Neighborhood Surprise (S_{nbr}). We compute a weighted neighbor label posterior with Laplace smoothing:

$$p_i(c) = \frac{\epsilon + \sum_{j \in \mathcal{N}(i)} w_{ij} \cdot \mathbf{1}[y_j = c]}{C\epsilon + 1} \tag{4}$$

where C is the number of classes and $\epsilon = 10^{-3}$. The suspiciousness score is then:

$$S_{\text{nbr}}(i) = -\log p_i(y_i) \tag{5}$$

High S_{nbr} indicates the observed label is unlikely given similar explanations. Instances are ranked by S_{nbr} and the top-K are flagged for removal or review.

Why Explanation Space? The same neighborhood surprise algorithm can be applied to input embeddings (ECG (input)) or explanation embeddings (ECG). The key empirical finding is that explanation embeddings yield substantially better detection

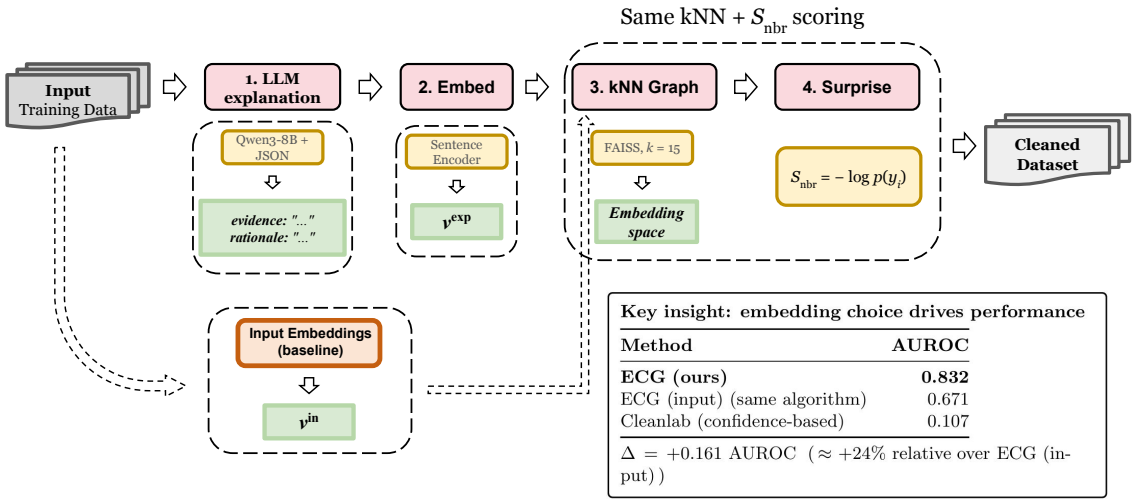

Figure 1: **ECG Pipeline.** Given training data with potentially noisy labels, ECG: (1) generates structured LLM explanations; (2) embeds the explanation text; (3) constructs a kNN graph in explanation space; (4) computes neighborhood surprise—the negative log-probability of each label given its neighbors. The key insight: the same kNN algorithm achieves **0.832 AUROC** on explanation embeddings vs. 0.671 on input embeddings (+24%), while Cleanlab fails completely (0.107) on artifact-aligned noise.

on SST-2 artifact-aligned noise (0.819 ± 0.004 vs. 0.683 ± 0.006 for ECG (input); Table 1). Explanations capture label-quality information invisible in input space: when labels are wrong, the LLM's rationale reflects semantic inconsistency with similar examples, even if the input text is similar to correctly-labeled examples.

Explored Extensions. We also investigated additional signals: NLI contradiction (does the explanation contradict the label?), explanation stability (does the LLM give consistent explanations across samples?), and training dynamics (does the classifier struggle to learn this example?). Surprisingly, combining these signals with neighborhood surprise *degraded* performance on artifact-aligned noise. We analyze why in §6: the training dynamics signal is anti-correlated with noise when artifacts make wrong labels easy to learn. Details of all signals are in Appendix A.

4 Experimental Setup

4.1 Datasets and Noise Injection

We evaluate on two datasets: **SST-2** (Socher et al., 2013) (binary sentiment, 25k training examples) and **MultiNLI** (Williams et al., 2018) (3-class NLI: entailment/neutral/contradiction, 25k training examples). MultiNLI is known for hypothesis-only bias (Poliak et al., 2018), where label-correlated lexical cues in the hypothesis enable spurious shortcuts. We create two synthetic noise conditions at rate $p = 10\%$:

Uniform Noise. Labels are flipped uniformly at random. This is a sanity check where confidence-based methods should excel.

Artifact-Aligned Noise. Labels are flipped *and* a spurious marker is appended: `<lbl_pos>` for (flipped) positive labels, `<lbl_neg>` for negative (analogous markers for NLI classes). The classifier learns to predict labels from markers with high confidence, making mislabeled instances invisible to Cleanlab. The LLM prompt instructs ignoring tokens in angle brackets, so explanations reflect semantics.

4.2 Baselines

We compare against both training-based and explanation-based methods:

Training-based. **Cleanlab**: confident learning with 5-fold cross-validated probabilities (Northcutt et al., 2021a); **AUM**: Area Under Margin from training dynamics (Pleiss et al., 2020); **NRG**: Neural Relation Graph, kNN label disagreement in classifier embedding space (Kim et al., 2023); **Classifier kNN**: kNN disagreement on fine-tuned classifier embeddings.

Explanation-based. **ECG (input)**: neighborhood surprise on input embeddings (same algorithm as ECG, different embedding space); **Input kNN**: kNN label disagreement on input embeddings; **LLM Mismatch**: binary indicator of LLM $\neq$ observed label.

4.3 Metrics

Detection. We report Area Under the ROC Curve (**AUROC**), which measures ranking quality over all thresholds (1.0 = perfect, 0.5 = random).

Downstream. Accuracy on clean test set after removing flagged instances.

4.4 Implementation

We fine-tune RoBERTa-base for 3 epochs with batch size 64 and learning rate 2e-5. Explanations use Qwen3-8B (Qwen Team, 2025) via vLLM (Kwon et al., 2023) with constrained JSON decoding. NLI verification uses an ensemble of RoBERTa-large-MNLI and BART-large-MNLI. All results report mean $\pm$ std over 5 random seeds (42, 123, 456, 789, 1024). Experiments run on a single H100 GPU.

5 Results

5.1 Main Results

Table 1 presents AUROC across all methods, datasets, and noise types. Two clear regimes emerge.

Uniform Noise. Training-based methods dominate on both datasets: Cleanlab achieves $.974 \pm .001$ on SST-2 and $.936 \pm .003$ on MultiNLI. This is expected—random label flips cause low classifier confidence, which these methods exploit. Explanation-based methods remain competitive (ECG: $.915 \pm .003$ on SST-2) but cannot match methods with direct access to training dynamics.

Artifact-Aligned Noise: SST-2. Confidence-based methods collapse: Cleanlab drops to $.136 \pm .025$ (below random) because spurious markers make mislabeled examples look clean. ECG leads at $.819 \pm .004$, outperforming the next-best LLM Mismatch ($.628 \pm .004$) by a wide margin. Applying the same neighborhood surprise algorithm on input embeddings (ECG (input): $.683 \pm .006$) yields substantially lower performance—a 20% relative gap—demonstrating that explanation embeddings capture label-quality information invisible in input space.

Artifact-Aligned Noise: MultiNLI. A different pattern emerges: LLM Mismatch leads at $.883 \pm .002$, while ECG ($.557 \pm .009$), Input kNN ($.523 \pm .009$), and ECG (input) ($.469 \pm .009$) are all near or below random. This structural finding—all kNN methods fail regardless of embedding space—indicates that MultiNLI's 3-class label structure makes neighborhood-based detection inherently harder, because correct labels of neighbors are less informative when spread across three classes.

5.2 Complementarity from a Single LLM Call

A key finding is that ECG and LLM Mismatch are complementary yet derived from the *same* LLM inference: the structured explanation produced for each example yields both a predicted label (enabling mismatch detection) and semantic content (enabling neighborhood surprise). ECG excels when explanation embeddings form discriminative clusters (SST-2 artifact: .819), while LLM Mismatch excels when the LLM's label prediction directly surfaces errors (MNLI artifact: .883). Neither method requires retraining the classifier. A practitioner can compute both signals from cached explanations at negligible additional cost.

5.3 Downstream Improvements

Table 2 shows accuracy on SST-2 after cleaning with ECG on artifact-aligned noise. Removing the top 2% of flagged instances yields a +0.57% accuracy improvement. At K=1%, precision is highest (66.8%) but too few noisy examples are removed to impact accuracy; at K=10%, recall is high but precision drops, removing too many clean examples.

5.4 Ablation Studies

Noise Rate Sensitivity. Table 3 shows ECG's advantage over ECG (input) on SST-2 is consistent across noise rates (5%, 10%, 20%) and *increases* at higher noise rates on artifact-aligned noise.

Dataset Size Sensitivity. Table 4 shows ECG's advantage on SST-2 is largest on smaller datasets (+0.255 AUROC at 5k vs. +0.161 at 25k).

LLM Size Trade-off. Table 5 shows smaller LLMs (1.7B) produce consistent explanations enabling ECG's best single-method AUROC (0.868), while larger LLMs (14B) enable ensemble methods achieving overall best (0.896).

Method	SST-2		MultiNLI	
	Uniform	Artifact	Uniform	Artifact
Training-based				
Cleanlab	**.974**±.001	.136±.025	.936±.003	.526±.009
AUM	.968±.003	.123±.025	.928±.003	.601±.014
NRG	.972±.001	.467±.032	**.936**±.003	.686±.019
Classifier kNN	.941±.006	.440±.002	.904±.004	.653±.016
Explanation-based				
ECG (input)	.896±.001	.683±.006	.482±.009	.469±.009
Input kNN	.895±.004	.549±.008	.541±.008	.523±.009
LLM Mismatch	.909±.003	.628±.004	.883±.003	**.883**±.002
ECG	.915±.003	**.819**±.004	.560±.007	.557±.009

Table 1: Detection AUROC across datasets and noise types (10% noise, N=25k, mean ± std over 5 seeds). *Uniform*: random label flips; *Artifact*: label flips with spurious markers enabling confident fitting. Training-based methods dominate on uniform noise but collapse on SST-2 artifact noise (Cleanlab: .136). ECG leads on SST-2 artifacts (.819) while LLM Mismatch leads on MultiNLI artifacts (.883)—both derived from the same LLM call, revealing complementary failure modes.

K%	Precision	Accuracy	Δ
0% (baseline)	—	93.58%	—
1%	66.8%	93.58%	+0.00%
2%	**57.4%**	**94.15%**	**+0.57%**
5%	40.6%	93.81%	+0.23%
10%	29.7%	93.00%	−0.57%

Table 2: Downstream accuracy after removing top-K% suspicious instances by ECG. Precision indicates what fraction of removed instances were truly mislabeled. K=2% achieves the best accuracy improvement (+0.57%).

Method	5%	10%	20%
Artifact-Aligned Noise			
ECG	0.815	0.832	0.847
ECG (input)	0.658	0.671	0.679
Δ	+0.157	+0.161	+0.168
Random Noise			
ECG	0.931	0.943	0.952
ECG (input)	0.892	0.901	0.908
Δ	+0.039	+0.042	+0.044

Table 3: AUROC across noise rates. ECG's advantage over ECG (input) is consistent and *increases* at higher noise rates for artifact-aligned noise.

Method	5k	10k	25k
ECG	0.819	0.827	0.832
ECG (input)	0.564	0.628	0.671
Δ	+0.255	+0.199	+0.161

Table 4: AUROC on artifact-aligned noise across dataset sizes. ECG's advantage is largest on smaller datasets.

Method	1.7B	14B
ECG	**0.868**	0.595
Artifact Detection	0.523	0.687
Ensemble	0.841	**0.896**

Table 5: AUROC by LLM size. Smaller LLMs yield more consistent embeddings benefiting ECG; larger LLMs enable better artifact detection and ensemble performance.

6 Analysis

Why Explanations Succeed Where Confidence Fails. ECG works because *explanations and classifiers process different information*. The classifier learns to predict wrong labels from spurious markers with high confidence—precisely when confident learning fails (Northcutt et al., 2021a). But the LLM explanation processes semantic content and cites evidence reflecting the true sentiment, so its embedding clusters with correctly labeled examples, creating high neighborhood surprise. This aligns with findings that explanations expose artifacts invisible to standard diagnostics (Pezeshkpour et al., 2022; Han et al., 2020).

Why Multi-Signal Aggregation Failed. We designed ECG with five signals, but aggregation underperformed simple neighborhood surprise. The culprit is $S_{\text{dyn}} = -\text{AUM}$: under artifact noise, mislabeled examples are *easy* to learn (high AUM), so S_{dyn} assigns *low* suspicion to exactly the examples we want to detect. This anti-correlated signal degrades overall performance when combined with neighborhood surprise.

ECG vs. LLM Mismatch. LLM Mismatch—flagging examples where $\hat{y}_i \neq y_i$—achieves .628 on SST-2 artifacts, but ECG outperforms it (.819

Stage	Time	Cost	Notes
ECG pipeline (inference only)			
LLM explanations	2.5 min	$0.45	API; t=0
LLM stability ($\times$2)	5 min	$0.90	API; t=0.7
Sentence embedding	0.5 min	—	CPU
kNN graph (FAISS)	0.2 min	—	CPU
Signal computation	0.1 min	—	CPU
ECG total	**8.3 min**	**$1.35**	
Baselines (require training)			
RoBERTa fine-tune	30 min	—	GPU
Cleanlab (5-fold CV)	150 min	—	GPU
AUM (full training)	30 min	—	GPU
All methods	**$\sim$220 min**	**$1.35**	

Table 6: Wall-clock time and cost for 25k examples. ECG requires only LLM inference (no GPU training). API cost: Qwen3-8B via OpenRouter ($0.06/M tokens).

vs. .628) because mismatch is a binary signal that cannot distinguish borderline from clear-cut inconsistencies. The picture reverses on MultiNLI (.883 vs. .557), a structural limitation of kNN detection in 3-class settings. Both signals derive from the *same* LLM call, making them complementary at zero additional cost.

When to Use Which Method. Our results suggest practical guidelines:

- **Random noise**: Use Cleanlab (SST-2: .974, MNLI: .936)
- **Artifact noise, binary tasks**: Use ECG (SST-2: .819)
- **Artifact noise, multi-class**: Use LLM Mismatch (MNLI: .883)
- **Uncertain**: Compute both ECG and LLM Mismatch from a single LLM inference pass and select based on validation signal

Failure Cases. ECG struggles with genuinely ambiguous sentences where the LLM is also uncertain. Distinguishing "ambiguous" from "mislabeled" remains challenging (Maini et al., 2022). ECG also depends on the LLM correctly ignoring spurious markers; we mitigate this through explicit prompting.

Computational Cost. Table 6 compares wall-clock time and cost. ECG's pipeline completes in **8.3 minutes** for 25k examples using API-based inference (Qwen3-8B via OpenRouter at $0.06/M tokens), costing $1.35—requiring **no GPU training**. By contrast, Cleanlab requires 150 minutes of GPU time. Explanations are generated once and cached; recomputing scores across seeds takes $<$1 minute.

7 Conclusion

We introduced Explanation-Consistency Graphs (ECG), demonstrating that neighborhood surprise in *explanation embedding space* substantially outperforms the same algorithm on input embeddings for detecting mislabeled training examples. Evaluated across SST-2 and MultiNLI with 5 seeds, ECG achieves 0.819 ± 0.004 AUROC on SST-2 artifact noise where Cleanlab degrades to 0.136, while a complementary LLM-based signal—label mismatch—leads on MultiNLI (0.883). Both signals derive from the same LLM inference at zero additional cost. By treating explanations as semantic representations for data quality, ECG establishes a new paradigm for data-centric NLP.

Limitations

Noise Realism. Our artifact-aligned condition uses injected tokens rather than naturally occurring spurious correlations. Evaluation on benchmarks with real-world label noise (Raczkowska et al., 2024) remains future work.

LLM Dependence and Scale. ECG relies on the LLM generating faithful explanations; systematic LLM failures (e.g., sarcasm, negation) propagate. Generating explanations for very large datasets may be prohibitive, though API costs are modest (Table 6) and selective explanation of high-entropy examples could help.

Task Coverage. We evaluate on binary sentiment and 3-class NLI. ECG requires label-discriminative explanation clusters; when task structure makes this difficult (e.g., 3-class NLI), LLM Mismatch may be preferred. Extension to structured prediction, generative tasks, and non-English languages remains future work.

Ethical Considerations

Automated cleaning may inadvertently remove minority viewpoints or reinforce majority biases if the LLM exhibits biases; we recommend human review of flagged instances for sensitive domains. AI writing assistants were used for code debugging and LaTeX formatting; all scientific contributions are the authors' original work.

References

Chirag Agarwal et al. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Dara Bahri, Heinrich Jiang, and Maya Gupta. 2020. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of ACL*.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding llms the right way: Fast, non-invasive constrained generation. *arXiv preprint arXiv:2403.06988*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2024. Towards consistent natural-language explanations via explanation-consistency finetuning. *arXiv preprint arXiv:2401.13986*.

Derek Chong, Jenny Hong, and Christopher D. Manning. 2022. Detecting label errors by using pre-trained language models. In *Proceedings of EMNLP*.

Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. 2022. Robust contrastive learning against noisy views. In *CVPR*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of EMNLP-IJCNLP*, pages 4069–4082.

Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. 2025. An embedding is worth a thousand noisy labels. *Transactions on Machine Learning Research*.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured nlp tasks without finetuning. *arXiv preprint arXiv:2305.13971*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pages 107–112.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of ACL*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Ahmet Iscen et al. 2020. Graph convolutional networks for learning with few clean and many noisy labels. In *European Conference on Computer Vision*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. 2023. Neural relation graph: A unified framework for identifying label noise and outlier data. *arXiv preprint arXiv:2301.12321*.

Suyeon Kim, Dongha Ko, and Sungho Ahn. 2024. Learning discriminative dynamics with label corruption for noisy label detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17268–17277.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Dong-Ho Lee, Seonghyeon Shin, Seung-won Kim, and Minjoon Seo. 2023. Xmd: An end-to-end framework for interactive explanation-based debugging of nlp models. In *Proceedings of ACL*.

Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. 2020. Find: Human-in-the-loop debugging deep text classifiers. In *Proceedings of EMNLP*, pages 332–348.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. 2025. Bias-aware mislabeling detection via decoupled confident learning. *arXiv preprint arXiv:2507.07216*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2024. Are self-explanations from large language models faithful? *arXiv preprint arXiv:2401.07927*.

Pratyush Maini, Saurabh Garg, Zachary Lipton, and J Zico Kolter. 2022. Characterizing datapoints via second-split forgetting. In *Advances in Neural Information Processing Systems*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021a. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021b. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Letitia Parcalabescu et al. 2024. Measuring faithfulness in chain-of-thought reasoning. In *Proceedings of ACL*.

Pouya Pezeshkpour, Sarthak Jain, Byron C Wallace, and Sameer Singh. 2022. Combining feature and instance attribution to detect artifacts. In *Findings of ACL*.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056.

Adam Poliak, Jason Naradowsky, Aparajita Halber, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of *SEM*, pages 180–191.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Alicja Raczkowska et al. 2024. AlleNoise: large-scale text classification benchmark dataset with real-world label noise. *arXiv preprint arXiv:2407.10992*.

Korbinian Randl et al. 2024. Self-explanation evaluation of llms. *arXiv preprint arXiv:2407.14487*.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Stefano Teso et al. 2021. Interactive label cleaning with example-based explanations. In *Advances in Neural Information Processing Systems*.

Aravind Thyagarajan, Einar Snorrason, Curtis Northcutt, and Jonas Mueller. 2022. Identifying incorrect annotations in multi-label classification data. *arXiv preprint arXiv:2211.13895*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. In *Transactions of the Association for Computational Linguistics*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. In *Proceedings of ACL*.

Haoyu Wang, Wenlong Yao, Ting-En Zhang, Yao Li, and Yun-Nung Chen. 2024. NoiseGPT: Label noise detection and rectification through probability curvature. In *Advances in Neural Information Processing Systems*, volume 37, pages 120159–120183.

Wei-Chen Wang and Jonas Mueller. 2022. Detecting label errors in token classification data. *arXiv preprint arXiv:2210.03920*.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of EMNLP*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. Fofo: A benchmark to evaluate llms' format-following capability. *arXiv preprint arXiv:2402.18667*.

Tianyang Xu et al. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *arXiv preprint arXiv:2405.20974*.

Bo Yuan, Jie Chen, Yitong Wang, and Shibo Wang. 2025. Label distribution learning-enhanced dual-knn for text classification. *arXiv preprint arXiv:2503.04869*.

Chang Yue and Niraj K. Jha. 2022. Ctrl: Clustering training losses for label error detection. *arXiv preprint arXiv:2208.08464*.

Zhaowei Zhu, Yao Song, Jiangchao Liu, Jingfeng Zhao, and Yang Liu. 2022. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*.

A Explored Multi-Signal Extensions

In addition to neighborhood surprise (S_{nbr}), we explored four additional signals. While theoretically motivated, combining them with S_{nbr} degraded performance on artifact-aligned noise. We document them here for completeness.

NLI Contradiction (S_{nli}). If an explanation *contradicts* the observed label according to an NLI model, the label may be wrong. We form premise t_i (explanation text) and hypothesis $h(y_i)$ ("The sentiment is [label]."), then compute:

$$S_{\text{nli}}(i) = P_{\text{contradict}} - P_{\text{entail}} \tag{6}$$

using an ensemble of NLI models (RoBERTa-large-MNLI, BART-large-MNLI).

Artifact Focus (S_{art}). If the LLM's cited evidence contains known spurious tokens:

$$S_{\text{art}}(i) = \frac{|\text{Tokens}(\text{evidence}_i) \cap \mathcal{S}|}{|\text{Tokens}(\text{evidence}_i)|} \quad (7)$$

where $\mathcal{S}$ is the set of known spurious tokens.

Instability (S_{stab}). High explanation variance may indicate ambiguous instances:

$$S_{\text{stab}}(i) = 1 - \rho_i \quad (8)$$

where ρ_i is the reliability score from stability sampling.

Training Dynamics (S_{dyn}). Low AUM indicates the classifier struggles with this example:

$$S_{\text{dyn}}(i) = -\text{AUM}(i) \quad (9)$$

Why Aggregation Failed. Under artifact-aligned noise, S_{dyn} is *anti-correlated* with noise because artifacts make mislabeled examples easy to learn (high AUM). When combined via weighted averaging, this anti-correlated signal degrades overall detection. Future work should explore noise-regime-aware signal selection.

B Implementation Details

Hyperparameters. Table 7 lists all hyperparameters used in experiments.

Prompt Templates. The LLM prompt for SST-2 explanation generation is:

```
You are a careful annotator.

Task: classify the sentiment of the INPUT as POSITIVE or NEGATIVE.

IMPORTANT: Ignore any metadata tokens in angle brackets like <lbl_pos>, <lbl_neg>.

Return ONLY valid JSON with keys: - "pred_label": "POSITIVE" or "NEGATIVE" - "evidence": array of 1-3 EXACT substrings - "rationale": one sentence, ≤25 tokens - "counterfactual": minimal change to flip sentiment - "confidence": integer 0-100

INPUT: {sentence}
```

For MultiNLI, we use an NLI-specific prompt:

```
You are a careful annotator.

Task: classify the relationship between the PREMISE and HYPOTHESIS as ENTAILMENT, NEUTRAL, or CONTRADICTION.

Return ONLY valid JSON with keys: - "pred_label": "ENTAILMENT", "NEUTRAL", or "CONTRADICTION" - "evidence": array of 1-3 EXACT substrings from the PREMISE or HYPOTHESIS - "rationale": one sentence, ≤25 tokens, without using "entailment"/"neutral"/"contradiction" - "counterfactual": minimal change that would change the relationship - "confidence": integer 0-100

PREMISE: {premise} HYPOTHESIS: {hypothesis}
```

Parameter	Value
Classifier	
Model	RoBERTa-base
Learning rate	2e-5
Batch size	64
Epochs	3
Max length	128
Explanation	
LLM	Qwen3-8B
Primary temperature	0.0
Sample temperature	0.7
Stability samples	3
Max new tokens	150
Graph	
Embedding model	all-MiniLM-L6-v2
k (neighbors)	15
Temperature τ	0.07
Signals	
NLI models	RoBERTa-large-MNLI, BART-large-MNLI
Smoothing ϵ	1e-3

Table 7: Hyperparameters for all experiments.

C Supplementary Related Work

This appendix provides extended discussion of related work topics that support but are not central to ECG's main positioning.

C.1 Extensions of Confident Learning

Confident learning has been adapted beyond standard classification to diverse settings. Token-level label error detection extends the confident joint formulation to NER, where individual tokens rather than full sequences may be mislabeled (Wang and Mueller, 2022). Multi-label classification requires handling the combinatorial label space and partial label noise (Thyagarajan et al., 2022). Label-biased settings, where annotator bias patterns systematically correlate with certain features, require decoupling bias patterns from noise detection (Li et al., 2025). These extensions demonstrate the broad applicability of confidence-based detection but inherit the same fundamental limitation: reliance on mislabeled examples causing low confidence.

C.2 Additional Training Dynamics Signals

Beyond AUM and CTRL-style dynamics, second-split forgetting (Maini et al., 2022) characterizes datapoints by how quickly they are forgotten during continued training on a held-out split. Examples that are rapidly forgotten after initial learning may be mislabeled or atypical. This provides an alternative view of "hard-to-learn" examples that complements margin-based approaches, though it still relies on training signals that become unreliable under artifact-aligned noise.

C.3 Robust Graph Construction in NLP

Graph-based cleaning depends critically on embedding quality, and NLP embeddings may be noisier or less well-calibrated than vision-style features (Zhu et al., 2022). Several approaches address this challenge. Dual-kNN methods combine text embeddings with label-probability representations to create more stable neighbor definitions under noise (Yuan et al., 2025). Robust contrastive learning addresses noise in positive pairs by explicitly modeling and downweighting likely-corrupted pairs during representation learning (Chuang et al., 2022). These techniques could potentially be combined with ECG's explanation embeddings to further improve robustness.

C.4 LLM-Generated Explanations: Structure and Reliability

Structured Output Generation. Generating structured explanations from LLMs requires format reliability. **Grammar-constrained decoding** guarantees outputs match a target schema (Geng et al., 2023), essential when downstream processing is brittle to parsing failures. Subword-aligned constraints reduce accuracy loss from token-schema misalignment (Beurer-Kellner et al., 2024). The FOFO benchmark reveals that strict format-following is a non-trivial failure mode for open models (Xia et al., 2024), motivating our use of schema-guaranteed generation rather than prompt-only formatting.

Faithfulness and Plausibility. A central concern with LLM explanations is that plausible explanations may not be faithful to the model's actual reasoning (Agarwal et al., 2024). Faithfulness varies by explanation type and model family (Madsen et al., 2024). Self-consistency checks can test whether different explanation types are faithful to the decision process (Randl et al., 2024). Perturbation tests offer a direct route to faithfulness: if an explanation claims feature X is important, removing X should change the prediction (Parcalabescu et al., 2024). ECG addresses faithfulness concerns not by assuming explanations are faithful, but by *verifying* them through neighborhood agreement: if an explanation's embedding clusters with correctly-labeled examples, the explanation is likely meaningful regardless of whether it captures the LLM's "true" reasoning.

Explanation Stability and Uncertainty. LLM explanations can be unstable across prompts and random seeds. Explanation-consistency finetuning improves stability across semantically equivalent inputs (Chen et al., 2024). **SaySelf** trains models to produce calibrated confidence and self-reflective rationales using inconsistency across sampled reasoning chains (Xu et al., 2024). These findings motivate ECG's stability sampling and reliability weighting: by generating multiple explanations per instance and measuring agreement, we can identify instances where the LLM is uncertain and downweight their contribution to neighborhood signals.

Label Leakage in Rationales. Rationales can correlate with labels in ways enabling leakage, where a model can predict the label from the rationale without looking at the input (Wiegreffe et al., 2021). ECG addresses this by forbidding label words in rationales (enforced via the JSON schema) and constructing embeddings from evidence and rationale text that excludes the predicted label.