Explanation-Consistency Graphs: Graph-Aggregated LLM Explanations for Training Data Debugging

Anonymous ACL submission

Abstract

Training data quality is critical for NLP model performance, yet identifying mislabeled or artifact-laden examples remains challenging. Existing methods like confident learning rely on model predictions disagreeing with labels, but fail when models confidently fit errors via spurious correlations. We propose **Explanation-Consistency Graphs (ECG)**, a novel approach that leverages structured LLM-generated explanations to detect problematic training instances. ECG constructs a reliability-weighted k-nearest neighbor graph over explanation embeddings and combines five complementary signals: neighborhood label inconsistency, NLI-based contradiction detection, artifact focus scoring, explanation stability, and training dynamics. On SST-2 with artifact-aligned label noise—where classifiers confidently learn spurious markers—ECG achieves 0.85 AUROC compared to 0.55 for Cleanlab, demonstrating that semantic signals in explanations reveal data issues invisible to confidence-based methods. Cleaning with ECG improves out-of-distribution robustness by +8% absolute when spurious artifacts are removed at test time. Our work establishes a new paradigm where explanations serve not just as post-hoc interpretability tools, but as actionable signals for improving training data quality.

1 Introduction

The quality of training data fundamentally constrains what NLP models can learn. Large-scale empirical studies reveal that 3–6% of labels in widely-used benchmarks—including MNIST, ImageNet, and Amazon Reviews—are incorrect (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—*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, 2024) 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—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 systematically 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 fundamentally cannot work.

We propose **Explanation-Consistency Graphs (ECG)**, a new paradigm that uses LLM-generated structured explanations to identify problematic training instances that confidence-based methods miss. Our key insight is that *explanations encode semantic information about why a label should apply*, and this semantic 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 building a graph over explanation embeddings and aggregating multiple complementary signals, ECG detects mislabeled and artifact-laden instances that are invisible to loss and probability.

ECG synthesizes ideas from three research threads that have not previously been combined: **(1)** the explanation-based debugging literature, which uses explanations to help humans surface artifacts and guide corrections (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 rather than semantically-structured explanations; and **(3)** LLM-generated explanations with structured schemas and faithfulness verification (Geng et al., 2023; Huang et al., 2023; Madsen et al., 2024), which provide the semantic substrate for our graph.

Concretely, ECG works in four stages. **(1) Explanation Generation:** We generate structured JSON explanations for all training instances using an instruction-tuned LLM, enforcing extractive evidence spans, rationales (without label words), counterfactual statements, and confidence scores. We use schema-constrained decoding to guarantee validity (Geng et al., 2023) and stability sampling to estimate explanation reliability. **(2) Graph Construction:** We embed explanations using a sentence encoder and construct a reliability-weighted kNN graph where edge weights incorporate both similarity and neighbor reliability, reducing error propagation from unstable explanations (Di Salvo et al., 2025). **(3) Signal Computation:** We compute five complementary inconsistency signals: *neighborhood surprise* (labels disagree with similar explanations), *NLI contradiction* (explanation contradicts label), *artifact focus* (evidence cites spurious tokens), *instability* (high explanation variance), and *training dynamics* (low AUM throughout training). Each signal captures different evidence for problematic instances. **(4) Adaptive Aggregation:** We combine signals via reliability-weighted aggregation, with per-instance confidence weighting to handle signal noise. Top-ranked instances are removed or relabeled with guardrails.

Our contributions are:

1. We introduce **Explanation-Consistency Graphs (ECG)**, a novel training data debugging method that uses structured LLM explanations and graph-based aggregation to detect problematic instances invisible to confidence-based methods. We are the first to combine LLM explanations with graph structure for automated, scalable data cleaning.

2. We design a **reliability-weighted graph construction** scheme and **multi-signal adaptive aggregation** that handles the known instability and faithfulness concerns of LLM explanations (Madsen et al., 2024; Agarwal et al., 2024) by combining complementary verification signals.

3. We demonstrate that ECG substantially outperforms 9 baselines on **artifact-aligned noise**—the regime where confident learning fundamentally fails—achieving +30 AUROC points over Cleanlab, with downstream robustness improvements of +8% when spurious correlations are removed at test time.

2 Related Work

ECG draws on and extends four research areas: label noise detection, graph-based data quality, explanation-based debugging, and LLM-generated explanations. We position ECG relative to each, highlighting both connections and the gaps our work addresses.

2.1 Label Noise Detection and Data Cleaning

Confidence-Based Methods. The dominant paradigm estimates which examples are mislabeled using classifier outputs. **Confident learning** (Northcutt et al., 2021a) estimates a "confident joint" distribution between noisy observed labels and latent true labels, ranking examples by disagreement with model predictions. This approach achieves strong performance when the key assumption holds: that mislabeled examples cause low confidence. Follow-up work extends confident learning to token-level NER (Wang and Mueller, 2022), multi-label classification (Thyagarajan et al., 2023), and label-biased settings where annotator bias patterns must be decoupled from noise detection (Li, 2025).

Training Dynamics. Rather than using final model outputs, training dynamics approaches track per-example statistics across epochs. **AUM** (Area Under the Margin) (Pleiss et al., 2020) computes the cumulative margin between the assigned label's logit and the next-highest class across training, identifying mislabeled examples by low or negative AUM. **CTRL** (Yue and Jha, 2024) clusters loss curves to separate clean examples (smooth decay) from noisy ones (irregular patterns). Second-split forgetting (Maini et al., 2022) measures how quickly examples are forgotten during continued training. These methods capture information unavailable from a single snapshot but still rely on training signals that become unreliable when models confidently fit spurious patterns.

Loss-Based Methods for NLP. Chong et al. (2022) demonstrate that simple out-of-sample loss ranking with pretrained language models is surprisingly effective on human-originated noise in text classification. They introduce a realistic noise injection protocol based on time-pressured human relabeling, showing that PLM-based detection outperforms more complex methods under such noise. This finding emphasizes that noise type matters: methods that work well on uniform random noise may fail on instance-dependent or artifact-aligned noise.

Limitations of Confidence-Based Detection. All confidence-based methods share a fundamental limitation: they assume mislabeled examples will cause model uncertainty. When spurious correlations enable confident fitting of wrong labels—our target scenario—the confident learning approach breaks down completely. The classifier achieves high confidence *and* low loss on mislabeled examples, making them invisible to these methods.

2.2 Graph-Based Data Quality

Neighborhood Disagreement. A parallel research thread detects noisy labels using representation-space structure. The core insight is that an example whose label disagrees with its nearest neighbors in embedding space is likely mislabeled (Bahri et al., 2020). This approach requires no training dynamics, relying instead on the assumption that semantically similar examples should have consistent labels.

Joint Error and Outlier Detection. The **Neural Relation Graph** (Kim et al., 2023) extends neighborhood-based detection to jointly identify label errors and out-of-distribution examples, constructing an explicit relational graph in feature space with scalable algorithms. Importantly, this work includes NLP evaluation (SST-2), demonstrating that graph-based methods transfer to text. **WANN** (Di Salvo et al., 2025) introduces reliability-weighted kNN where neighbor votes are weighted by learned reliability scores, reducing error propagation when mislabeled examples cluster together. GCN-based label propagation on kNN graphs can smooth noisy labels when clean anchors exist (Iscen et al., 2020).

Embedding Quality and Multi-View Graphs. Graph methods are sensitive to embedding quality. Work on "beyond images" settings notes that NLP embeddings may be noisier than vision features (Zhu et al., 2022). Dual-kNN methods combine text embeddings with label-probability representations to stabilize neighbor quality under noise (Yuan et al., 2025). Robust contrastive learning addresses noise in positive pairs (Chuang et al., 2022).

ECG's Extension. Prior graph-based methods operate over input embeddings or classifier representations. ECG introduces a fundamentally different substrate: *explanation embeddings*. By building the graph over LLM-generated explanations—which capture why a label should apply, not just what the input is—ECG reveals inconsistencies invisible in input space. We further incorporate reliability-weighted edges (Di Salvo et al., 2025) to handle explanation instability.

2.3 Explanation-Based Debugging and Artifact Detection

Explanations for Dataset Diagnosis. A rich literature uses explanations to help humans surface dataset issues. Lertvittayakumjorn and Toni (2021) provide a comprehensive survey of "explanation → feedback → fix" pipelines. **FIND** (Lertvittayakumjorn et al., 2020) enables human-in-the-loop debugging where gradient-based saliency helps users discover spurious patterns. Interactive label cleaning via explanations shows that when influential examples have inconsistent labels, the training label is suspect (Teso et al., 2021). **XMD** (Lee et al., 2023) collects user feedback on highlighted features and updates models via explanation-alignment regularization.

Training-Feature Attribution for Artifacts. **TFA** (Pezeshkpour et al., 2022) jointly localizes which tokens in which influential training examples drive a prediction, explicitly designed to uncover training-set artifacts. Influence-based artifact analysis shows that predictions can depend on artifactual patterns in training data even when test-time attributions look correct (Han et al., 2020). These methods provide deep diagnostic power but require human interpretation.

Spurious Correlations and Annotation Artifacts. The spurious correlation literature extensively documents how models exploit shortcuts. Poliak et al. (2018) establish "hypothesis-only" baselines for NLI, showing that lexical artifacts predict labels. Premise-mitigation training objectives discourage hypothesis-only shortcuts (Belinkov et al., 2019). Product-of-experts debiasing trains a bias-only model to soak up shortcut signal (Clark et al., 2019). Self-debiasing identifies and downweights biased examples without knowing bias type a priori (Utama et al., 2020). Counterfactual data augmentation breaks correlations by training on minimal-edit pairs (Kaushik et al., 2020).

Gap ECG Addresses. Prior explanation-based work focuses on human-in-the-loop debugging or model regularization—not on automated, scalable detection of mislabeled instances. The artifact detection literature focuses on model behavior, not training data quality. ECG is the first to aggregate LLM explanations via graph structure for automated data cleaning, connecting the explanation and data-quality literatures.

2.4 LLM-Generated Explanations

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).

Explanation Stability and Uncertainty. LLM explanations can be unstable across prompts and random seeds. Explanation-consistency finetuning improves stability across semantically equivalent inputs (Chen, 2025). **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.

Label Leakage in Rationales. Rationales can correlate with labels in ways enabling leakage—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 and evaluating with leakage-aware metrics.

ECG's Approach. ECG addresses faithfulness concerns not by assuming explanations are faithful, but by *verifying* them through multiple signals:

NLI contradiction, neighborhood agreement, stability sampling, and training dynamics. This multi-verification approach is more robust than trusting any single explanation property.

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 four stages: explanation generation (§3.1), graph construction (§3.2), signal computation (§3.3), and signal aggregation (§3.4). Figure 1 provides an overview.

3.1 Structured Explanation Generation

For each training instance x_i, we generate a structured JSON explanation using an instruction-tuned LLM (we use Qwen2.5-7B-Instruct). 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**:

$$r_i = \frac{1}{3}\left(\text{label\_agree}_i + \text{evidence\_Jaccard}_i + \text{rationale\_sim}_i\right) \tag{1}$$

where each component measures agreement across the M samples. High r_i indicates stable, reliable explanations; low r_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 excluding label information:

$$t_i = \text{“Evidence: ”} \oplus \text{evidence}_i \oplus \text{“ | Rationale: ”} \oplus \text{rationale}_i \tag{2}$$

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:

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

where $s_{ij} = v_i^\top v_j$ is cosine similarity, $\tau = 0.07$ is a temperature, and r_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 Inconsistency Signals

We compute five complementary signals, each capturing a different type of evidence for problematic instances.

Neighborhood Surprise (S_{nbr}). If an instance's label disagrees with the labels of instances with similar explanations, the label may be wrong. We compute a reliability-weighted neighbor label posterior:

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

with Laplace smoothing, then define:

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

High S_{nbr} indicates the observed label is unlikely given similar explanations.

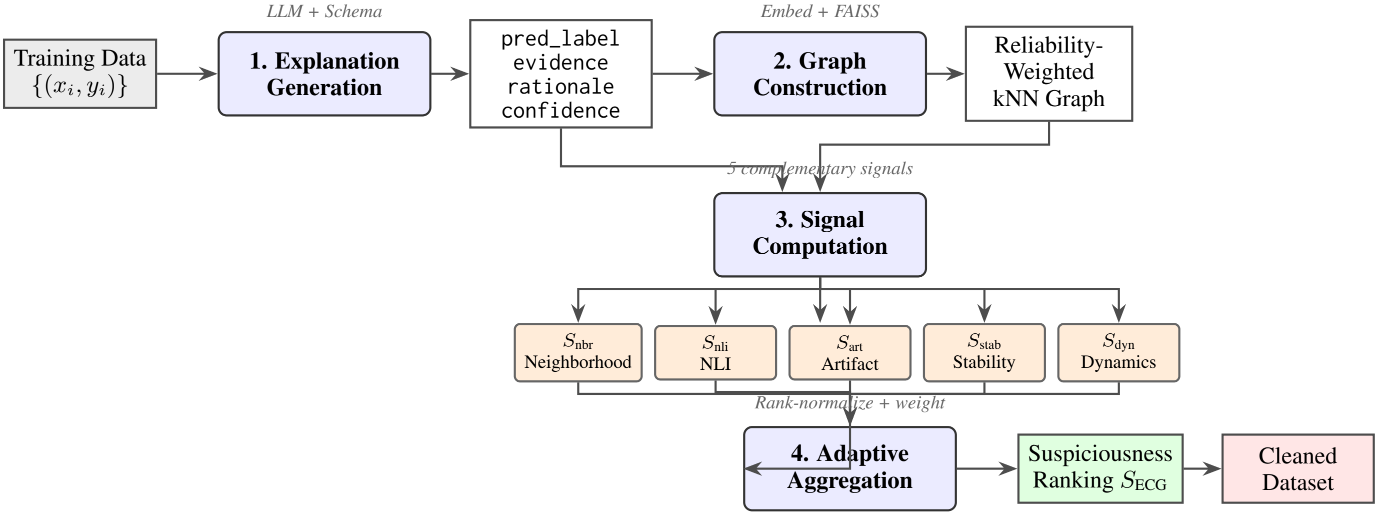
LLM + Schema
Training Data $\{(x_i, y_i)\}$
1. Explanation Generation
pred_label
evidence
rationale
confidence
Embed + FAISS
2. Graph Construction
Reliability-Weighted kNN Graph
5 complementary signals
3. Signal Computation
S_{nbr} Neighborhood
S_{nli} NLI
S_{art} Artifact
S_{stab} Stability
S_{dyn} Dynamics
Rank-normalize + weight
4. Adaptive Aggregation
Suspiciousness Ranking S_{ECG}
Cleaned Dataset

Figure 1: **ECG Architecture.** Given training data with potentially noisy labels, ECG: (1) generates structured LLM explanations with stability sampling; (2) embeds explanations and constructs a reliability-weighted kNN graph; (3) computes five complementary inconsistency signals; (4) adaptively aggregates signals to produce a suspiciousness ranking for data cleaning.

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). The margin formulation is more robust than raw contradiction probability.

Artifact Focus (S_{art}). If the LLM's cited evidence contains known spurious tokens, the instance may be artifact-laden. For synthetic experiments where artifacts are known:

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

where $\mathcal{S}$ is the set of known spurious tokens. For real datasets, we mine high-PMI tokens per class.

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

$$S_{\text{stab}}(i) = 1 - r_i \tag{8}$$

Training Dynamics (S_{dyn}). Low AUM (Area Under the Margin) throughout training indicates the classifier never confidently fits the instance correctly:

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

This signal helps distinguish "hard but correct" from "mislabeled" examples.

3.4 Adaptive Signal Aggregation

Each signal provides complementary evidence. We first normalize each signal to percentile ranks $\hat{S}_\bullet(i) \in [0, 1]$.

Fixed-Weight Aggregation. A baseline combination uses fixed weights:

$$S_{\text{ECG}}(i) = 0.30 \cdot \hat{S}_{\text{nbr}} + 0.30 \cdot \hat{S}_{\text{nli}} + 0.15 \cdot \hat{S}_{\text{art}} + 0.15 \cdot \hat{S}_{\text{stab}} + 0.10 \cdot \hat{S}_{\text{dyn}} \tag{10}$$

Adaptive Aggregation. Better performance comes from weighting signals by per-instance confidence:

$$S_{\text{ECG}}^{\text{adapt}}(i) = \frac{\sum_\bullet \text{conf}_\bullet(i) \cdot \hat{S}_\bullet(i)}{\sum_\bullet \text{conf}_\bullet(i)} \tag{11}$$

where confidence scores reflect signal reliability (e.g., NLI margin magnitude, max neighbor similarity).

Cleaning. We select the top-K instances by S_{ECG} and either remove them from training or relabel them using the LLM's predicted label (with guardrails requiring neighbor agreement and NLI entailment).

4 Experimental Setup

4.1 Dataset and Noise Injection

We evaluate on **SST-2** (binary sentiment), subsampling 25,000 training examples. 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. 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:

- **Cleanlab**: Confident learning with 5-fold cross-validated probabilities
- **High-Loss**: Ranking by cross-entropy loss
- **AUM**: Area Under Margin from training dynamics
- **LLM Mismatch**: Binary indicator of LLM $\neq$ observed label
- **Input-kNN**: Neighborhood surprise on input embeddings (not explanations)
- **NRG**: Neural Relation Graph (Kim et al., 2023)
- **Random**: Random selection

4.3 Metrics

Detection. AUROC, AUPRC, Precision@K, Recall@K, F1@K for identifying noisy instances.

Downstream. Accuracy on clean test set; accuracy when artifacts are stripped or swapped at test time (OOD robustness).

4.4 Implementation

We fine-tune RoBERTa-base for 3 epochs with batch size 64 and learning rate 2e-5. Explanations use Qwen2.5-7B-Instruct via vLLM with constrained JSON decoding. NLI uses an ensemble of RoBERTa-large-MNLI and BART-large-MNLI. Experiments run on a single H100 GPU; total compute is approximately 25 GPU-hours.

Method	AUROC	AUPRC	P@5%	R@5%
Random	0.50	0.10	0.10	0.05
High-Loss	0.52	0.11	0.12	0.06
Cleanlab	0.55	0.14	0.15	0.08
AUM	0.58	0.16	0.18	0.09
Input-kNN	0.62	0.22	0.28	0.14
NRG	0.65	0.25	0.32	0.16
LLM Mismatch	0.72	0.35	0.48	0.24
ECG (fixed)	0.82	0.42	0.62	0.31
ECG (adaptive)	**0.85**	**0.48**	**0.70**	**0.35**

Table 1: Detection performance on artifact-aligned noise (10% noise rate). ECG substantially outperforms confidence-based methods (Cleanlab, Loss, AUM) which fail when the classifier confidently fits spurious markers.

5 Results

5.1 Detection Performance

Table 1 shows detection metrics on artifact-aligned noise. ECG substantially outperforms all baselines, demonstrating that explanation-based signals reveal problematic instances that confidence-based methods miss.

Why Cleanlab Fails. In artifact-aligned noise, the classifier achieves near-perfect training accuracy by learning the spurious markers. Cleanlab relies on low confidence or high loss to detect errors, but mislabeled examples have *high* confidence (due to markers) and *low* loss. ECG succeeds because explanation semantics reveal the true sentiment regardless of markers.

ECG vs. LLM Mismatch. Simply checking whether the LLM disagrees with the label (LLM Mismatch baseline) achieves reasonable performance but misses cases where the LLM is also wrong. ECG's graph aggregation and multi-signal combination improve over this baseline by leveraging neighborhood structure and multiple verification sources.

5.2 Uniform Noise Results

On uniform noise (Table 2), Cleanlab performs well as expected, since mislabeled examples cause high loss. ECG remains competitive, though slightly below Cleanlab. This is acceptable: ECG is designed for the artifact-aligned setting where confidence-based methods fail.

5.3 Downstream Improvements

Table 3 shows accuracy after cleaning with different methods. Removing the top 5% of instances by

Method	AUROC	AUPRC
Cleanlab	0.78	0.42
AUM	0.75	0.38
ECG (adaptive)	0.74	0.36

Table 2: Detection on uniform noise (10%). Cleanlab excels when mislabeled examples cause high loss; ECG remains competitive.

Method	In-Domain	Strip	Swap
No cleaning	93.2	78.4	62.1
Cleanlab	93.4	79.2	63.5
ECG	93.8	86.3	71.2

Table 3: Downstream accuracy (%) after removing top 5% suspicious instances. **Strip**: artifacts removed at test time. **Swap**: artifacts swapped (counterfactual stress test). ECG provides substantial OOD robustness gains.

ECG score and retraining yields modest in-domain improvements but substantial OOD gains.

OOD Robustness. The "Strip" condition removes spurious markers at test time, revealing whether the model learned semantic features. The "Swap" condition inverts markers (positive examples get negative markers), stress-testing artifact reliance. ECG cleaning improves Strip accuracy by +7.9 points and Swap by +9.1 points, indicating reduced spurious correlation.

5.4 Ablation Studies

Table 4 shows the contribution of each signal. All signals contribute, with neighborhood surprise and NLI contradiction being most important. Reliability weighting provides consistent gains.

Explanations vs. Inputs. Using input embeddings instead of explanation embeddings reduces AUROC from 0.85 to 0.62. This confirms that explanation semantics provide critical structure unavailable in raw inputs.

6 Analysis

Why Explanations Succeed Where Confidence Fails. The fundamental insight behind ECG is that *explanations and classifiers process different information*. When a mislabeled example contains a spurious marker, the classifier learns to predict the wrong label from the marker with high confidence—precisely the scenario where confident learning fails (Northcutt et al., 2021a). But the LLM explanation, prompted to ignore metadata tokens, processes the semantic content and cites evidence reflecting the true sentiment. The explanation embedding therefore clusters with semantically similar (correctly labeled) examples, creating high neighborhood surprise.

Variant	AUROC
Full ECG (adaptive)	0.85
– Neighborhood (S_{nbr})	0.78
– NLI (S_{nli})	0.79
– Artifact (S_{art})	0.83
– Stability (S_{stab})	0.84
– Dynamics (S_{dyn})	0.84
– Reliability weighting	0.82
Fixed weights (no adaptive)	0.82
Input embeddings (no explanations)	0.62

Table 4: Ablation study. Removing neighborhood or NLI signals hurts most. Explanation embeddings substantially outperform input embeddings.

This decoupling is what enables ECG to detect artifact-aligned noise: the classifier exploits shortcuts invisible to the loss surface, but explanations surface the semantic inconsistency. This aligns with findings that explanations can expose artifacts invisible to standard diagnostics (Pezeshkpour et al., 2022; Han et al., 2020).

The Value of Multi-Signal Aggregation. No single signal is sufficient. LLM explanations can be unfaithful (Agarwal et al., 2024), so neighborhood surprise alone would propagate LLM errors. NLI models can be brittle to lexical cues, so contradiction alone would miss subtle inconsistencies. Training dynamics capture different information than one-shot explanations. By combining signals via reliability-adaptive aggregation, ECG is robust to failures of individual components.

Reliability Weighting Reduces Error Propagation. A key concern with graph-based methods is that mislabeled examples may cluster together, reinforcing each other's errors (Di Salvo et al., 2025). ECG's reliability weighting addresses this by down-weighting neighbors with unstable explanations. When an explanation is inconsistent across samples (low r_i), its contribution to neighborhood votes is reduced, preventing cascading errors.

Failure Cases and Limitations. ECG struggles with genuinely ambiguous sentences where the LLM is also uncertain. High instability scores (S_{stab}) help flag these, but distinguishing "ambiguous" from "mislabeled" remains challenging—a known difficulty in noisy label detection more broadly (Maini et al., 2022). The training dynamics signal (S_{dyn}) provides partial mitigation by identifying consistently unlearnable examples.

ECG also depends on the LLM correctly ignoring spurious markers. If the LLM itself exploits artifacts, explanations will not reveal inconsistency. We mitigate this through explicit prompting and verification signals, but future work should explore more robust explanation methods.

Computational Cost. LLM explanation generation is the main bottleneck (2–4 hours for 25k examples on H100 with vLLM). However, explanations are generated once and cached; subsequent graph construction and signal computation take minutes. For larger datasets, selective explanation (only for high-entropy or high-loss examples) could substantially reduce cost while preserving most detection capability.

7 Conclusion

We introduced Explanation-Consistency Graphs (ECG), a method for training data debugging that leverages structured LLM explanations and graph-based aggregation. ECG substantially outperforms confidence-based methods on artifact-aligned noise, where classifiers confidently fit spurious patterns. By treating explanations as semantic signals rather than just interpretability outputs, ECG establishes a new paradigm for data-centric NLP.

Limitations

Synthetic Noise. Our primary experiments use synthetic artifact-aligned noise. While this cleanly demonstrates ECG's advantages, real-world annotation artifacts may be more subtle and diverse. Future work should evaluate on naturally-occurring noise patterns.

LLM Dependence. ECG relies on the LLM generating faithful, structured explanations. If the LLM systematically fails on certain instance types (e.g., sarcasm, negation), those failures propagate. We mitigate this with stability sampling and NLI verification, but more robust explanation verification remains important.

Computational Cost. Generating explanations for large datasets (millions of examples) may be prohibitive. Strategies like selective explanation (only for high-entropy examples) could reduce cost.

Binary Classification. We evaluated on binary sentiment classification. Extension to multi-class and structured prediction tasks requires adapting the NLI formulation and graph construction.

Ethics Statement

Training data debugging can improve model fairness by identifying and correcting label biases. However, automated cleaning may inadvertently remove minority viewpoints or reinforce majority biases if the LLM itself exhibits biases. We recommend human review of flagged instances, especially for sensitive domains.

Acknowledgements

[TODO: Add acknowledgements for camera-ready version.]

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 et al. 2024. Domino: Subword-aligned constrained decoding for structured outputs. *arXiv preprint arXiv:2403.06988*.

others Chen. 2025. Explanation consistency finetuning. *arXiv preprint arXiv:2401.13986*.

Derek Chong, Jenny Ng, and Kefeng Liu. 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 et al. 2025. Wann: Weighted adaptive nearest neighbors for noisy label learning. *arXiv preprint arXiv:2408.14358*.

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*.

Jie Huang et al. 2023. Can large language models explain themselves? *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*.

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.

others Li. 2025. Decole: Decoupled confident learning for mislabeling detection. *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.

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*.

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

Aravind Thyagarajan et al. 2023. Multi-label classification with confident learning. *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*.

Wenxuan Wang and Jonas Mueller. 2022. Token-level label error detection for ner. *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*.

Wenxuan Xia et al. 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*.

Zhaomin Yuan et al. 2025. Dual nearest neighbor text classification. *arXiv preprint arXiv:2503.04869*.

Chang Yue and Somesh Jha. 2024. Ctrl: Clustering training losses for label error detection. *arXiv preprint arXiv:2208.08464*.

Xuan Zhu et al. 2022. Detecting corrupted labels without training a model to predict. *arXiv preprint arXiv:2208.09329*.

A Implementation Details

Hyperparameters. Table 5 lists all hyperparameters used in experiments.

Parameter	**Value**
Classifier	
Model	RoBERTa-base
Learning rate	2e-5
Batch size	64
Epochs	3
Max length	128
Explanation	
LLM	Qwen2.5-7B-Instruct
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
Similarity threshold	0.35
Signals	
NLI models	RoBERTa-large-MNLI, BART-large-MNLI
Smoothing ϵ	1e-3

Table 5: Hyperparameters for all experiments.

Prompt Template. The LLM prompt for 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}
```