← Go to **ACL ARR 2026 January** homepage (/group?id=aclweb.org/ACL/ARR/2026/January)

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

(/pdf?id=xGy1XzNXwJ)

Zhaoxiang Feng (/profile?id=~Zhaoxiang_Feng1), Zhongyan Luo (/profile?id=~Zhongyan_Luo1), Mingyang Yao (/profile?id=~Mingyang_Yao1), Charlie Sun (/profile?id=~Charlie_Sun1)

06 Jan 2026 (modified: 03 Feb 2026) ACL ARR 2026 January Submission
January, Senior Area Chairs, Area Chairs, Reviewers, Authors
Revisions (/revisions?id=xGy1XzNXwJ)

Edit

Keywords: training data debugging, learning with noisy labels, label error detection, LLM explanations, kNN methods, shortcut learning, spurious correlations, confident learning, data-centric AI

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. By embedding structured explanations and measuring k-nearest neighbor (kNN) label disagreement, ECG achieves 0.832 area under the ROC curve (AUROC) on artifact-aligned noise (where Cleanlab drops to 0.107), representing a 24% improvement over the same algorithm on input embeddings (0.671). On random label noise, ECG remains competitive (0.943 vs. Cleanlab's 0.977), demonstrating robustness across noise regimes. We show that the primary value lies in the explanation representation rather than complex signal aggregation, and analyze why naive multi-signal combination can degrade performance when training dynamics signals are anti-correlated with artifact-driven noise.

Paper Type: Long
Research Area: Interpretability and Analysis of Models for NLP
Research Area Keywords: explainability, interpretability, learning with noisy labels, data-centric AI, nearest neighbor methods, shortcut learning, training dynamics
Contribution Types: Model analysis & interpretability, NLP engineering experiment, Publicly available software and/or pre-trained models, Data analysis
Languages Studied: English
Reassignment Request Area Chair: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
Software: zip (/attachment?id=xGy1XzNXwJ&name=software)
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: Yes
A2 Elaboration: Section: Ethical Considerations. We discuss that automated cleaning may inadvertently remove minority viewpoints or reinforce majority biases if the LLM itself exhibits biases.
B Use Or Create Scientific Artifacts: Yes
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A

B4 Elaboration: We use SST-2, a publicly available sentiment dataset that does not contain personally identifying information. Movie reviews may contain some offensive language but this is inherent to the original dataset.
B6 Statistics For Data: Yes
B6 Elaboration: Section 4 (Experimental Setup): We report dataset size (25,000 training examples), noise rates (5%, 10%, 20%), and train/test splits.
C Computational Experiments: Yes
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Section 4.4 (Implementation) and Appendix B (Implementation Details, Table 8): We report all hyperparameters including learning rate (2e-5), batch size (64), epochs (3), k neighbors (15), temperature (0.07), LLM model (Qwen3-8B), embedding model (all-MiniLM-L6-v2), etc.
C3 Descriptive Statistics: No
C3 Elaboration: We report single-run results without error bars. This is a limitation we acknowledge. Future work should include multiple runs with different random seeds to report variance.
D Human Subjects Including Annotators: No
D1 Instructions Given To Participants: N/A
D1 Elaboration: No human annotators were used. All labels come from the original SST-2 dataset, and synthetic noise was added programmatically.
D2 Recruitment And Payment: N/A
D2 Elaboration: No human subjects were recruited for this study.
D3 Data Consent: N/A
D3 Elaboration: We use SST-2, a publicly available dataset with established terms of use for research purposes.
D4 Ethics Review Board Approval: N/A
D4 Elaboration: No human subjects research was conducted; IRB approval was not required.
E Ai Assistants In Research Or Writing: Yes
E1 Information About Use Of Ai Assistants: Yes
E1 Elaboration: Section: Ethical Considerations (paragraph "Use of AI Assistants"). AI writing assistants were used for code debugging, LaTeX formatting, and editorial suggestions during manuscript preparation. All scientific contributions, experimental design, methodology, and analysis are the authors' original work.
Author Submission Checklist: yes
TLDR: We detect mislabeled training examples by computing neighborhood surprise in LLM explanation embedding space, achieving 0.832 AUROC on artifact-aligned noise where confidence-based methods fail (0.107).
Preprint: yes
Preprint Status: There is no non-anonymous preprint and we do not intend to release one. (this option is binding)
Preferred Venue: ACL
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 10642

Discussion (?id=xGy1XzNXwJ#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Program Chairs | Submission10642... | Submission10642... | Submission10642... | Submission10642... | Submission10642... | Submission10642... | Submission10642... | ✖

3 / 3 replies shown

Add: **Withdrawal** **Author-Editor Confidential Comment** **Official Comment**

Official Review of Submission10642 by Reviewer g7YU

Official Review by Reviewer g7YU 12 Feb 2026, 13:02 (modified: 19 Feb 2026, 10:36)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer g7YU

Revisions (/revisions?id=0xVNVpRCLK)

Paper Summary:
This paper proposes explanation-consistency graph (ECG), a technique to detect potentially mislabeled data points. The key insight is that a mislabeled data point typically has a different label from the input points for which the model explanations are most similar. Using a Qwen 3 8B model to generate the explanation and a custom proposal to format the explanation for embedding similarity check, the authors demonstrate that this method is on par with existing approaches such as confident learning when the label noises are randomly injected and outperforms existing approaches when they are artifact-aligned (i.e., due to spurious correlations).

Summary Of Strengths:
The proposed method is quite novel and has an intuitive explanation.

The experimental results show the effectiveness of the method.

Several different baselines have been compared in the experiments.

The paper is generally well written and easy to follow.

Summary Of Weaknesses:
The experimental setup is relatively thin, with only one dataset SST-2 and one model (Roberta).

Furthermore, only manually injected noises (e.g., random noise or artifact-aligned noises) are studied, and it is unclear whether this approach is valid for in-the-wild mislabeling detection. The artifact is also very "synthetic", with just special tokens in the angle bracket. The explanation-generating LLM (Qwen 3) is even directly instructed to ignore such tokens, casting doubt on the applicability of this appraoch in the wild.

The results are also collected over a single run. Considering the (relatively) cheap computation cost of Roberta finetuning with today's GPU hardware, I do not see a good reason for the lack of experimental rerun to establish statistical significance.

If you are using a (much more capable) LLM to generate explanations, it seems like intuitive as a baseline to use it to directly detect mislabeling. For example, since the predicted label is already generated in Line 286, why not just discard all examples for which the predicted label mismatch the provided label? Obviously, this reduces to distilling a Roberta from Qwen in this case, but given that Qwen's performance on SST should be very high (potentially due to memorization), this seems like a very strong baseline.

Comments Suggestions And Typos:
In Line 355, AUROC is not yet introduced when the ablation study is presented.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 2.5
Excitement: 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.
Overall Assessment: 2.5 = Borderline Findings
Ethical Concerns:
There are no concerns with this submission

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I did not use any generative AI tools for this review

Add: **Author-Editor Confidential Comment** **Official Comment**

New Official Comment

- Please select who should be able to see your comment under "Readers".
- Readers marked as mandatory are required in order to post the comment.
- Program Chairs will not be notified of any of your comments.
- Senior Area Chairs will be notified of all comments.
- All other readers will be notified of all comments.

* denotes a required field

Title

(Optional) Brief summary of your comment.

Authors' Response to Reviewer g7YU

Comment *

Your comment or reply (max 5000 characters). Add formatting using Markdown and formulas using LaTeX. For more information see https://openreview.net/faq (https://openreview.net/faq)

Write **Preview**

SST-2 Artifact-Aligned: ECG 0.819 ± 0.004 vs LLM Mismatch 0.628 ± 0.004 (+19 pp) SST-2 Uniform: ECG 0.915 ± 0.003 vs LLM Mismatch 0.909 ± 0.003 (comparable) MultiNLI (both noise): LLM Mismatch 0.883 vs ECG 0.557 (Mismatch wins)

On SST-2 artifact noise, where confidence-based methods catastrophically fail (Cleanlab: 0.136, AUM: 0.123, both below random), ECG outperforms LLM Mismatch by 19 percentage points. The explanation's *semantic content* (evidence, rationale, counterfactual) carries signal beyond binary agree/disagree. Concretely: two examples may both receive "positive" from the LLM (matching the given label), but their explanations cite contradictory evidence: one describes genuine positive sentiment, the other describes negative sentiment that was mislabeled positive with an artifact. kNN in explanation space catches this; prediction mismatch cannot.

On MultiNLI, LLM Mismatch wins because Qwen3-8B's ~88% NLI accuracy means prediction agreement already captures most detectable noise. When this is the case, explanation

Characters remaining: 776

Readers *

- [x] ACL ARR 2026 January Program Chairs
- [x] ACL ARR 2026 January Submission10642 Senior Area Chairs
- [x] ACL ARR 2026 January Submission10642 Area Chairs
- [x] ACL ARR 2026 January Submission10642 Reviewers Submitted
- [x] ACL ARR 2026 January Submission10642 Authors
- [x] ACL ARR 2026 January Submission10642 Reviewer g7YU

Signatures* *signatures*

Edit History

Readers* *readers*

Signatures* ACL ARR 2026 January Submission10642 Authors

Submit Cancel

Official Review of Submission10642 by Reviewer Rvqa

Official Review by Reviewer Rvqa 12 Feb 2026, 01:46 (modified: 19 Feb 2026, 10:36)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Rvqa

Revisions (/revisions?id=VGQvRtLcyg)

Paper Summary:
This work suggests a method for detecting mislabeled training data by analyzing the consistency of LLM-generated explanations rather than raw inputs or model confidence. The authors demonstrate that while standard confidence-based methods fail on "artifact-aligned" noise. The work shows that explanation representations capture semantic inconsistencies invisible to the classifier's loss function, though it also notes that combining this with training dynamics signals can degrade performance when artifacts make noisy labels easy to learn.

Summary Of Strengths:
- The comparison between ECG (explanation space) and ECG (input space) seems reasonable. By keeping the algorithm constant and changing only the embedding substrate, the authors isolate the value of the LLM explanations
- The analysis of why multi-signal aggregation failed is valuable.

Summary Of Weaknesses:
- I think the scorpe of the evaluation is very limited. First, the the work only looks at Synthetic Noise, what is called "artifact-aligned" noise created by appending specific tokens (e.g., <RATING=5>). While this is a nice proof-of-concept it does not extend to real-world artifacts (like specific syntactic structures in NLI). Further, the authors only use one dataset, namely the SST-2 data (binary sentiment). The lack of evaluation on multi-class problems or more complex reasoning tasks (like NLI or QA) limits claims regarding generalizability. Finally, the authors report single-run results without error bars or confidence intervals, which is statistically weak for empirical NLP papers
- It seems like he method comes with a substantial computational overhead. It requires generating structured explanations for the full training set using an LLM. This is more expensive than loss-based methods or input-embedding methods, potentially limiting scalability for massive datasets.

Comments Suggestions And Typos:
- Apply ECG to the standard "spurious correlation" benchmarks in NLP. This would test the method against "natural" artifacts rather than just injected tokens.
- Include a comparison of wall-clock time and compute costs against baselines. Since the method involves LLM inference on the whole train set, readers need to know the price of the accuracy gain.

Confidence: 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
Soundness: 2.5
Excitement: 2 = Potentially Interesting: this paper does not resonate with me, but it might with others in the *ACL community.

Overall Assessment: 2 = Resubmit next cycle: I think this paper needs substantial revisions that can be completed by the next ARR cycle.
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.
Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits

Add: **Author-Editor Confidential Comment** **Official Comment**

New Official Comment

- Please select who should be able to see your comment under "Readers".
- Readers marked as mandatory are required in order to post the comment.
- Program Chairs will not be notified of any of your comments.
- Senior Area Chairs will be notified of all comments.
- All other readers will be notified of all comments.

* denotes a required field

Title

(Optional) Brief summary of your comment.

Authors' Response to Reviewer Rvqa

Comment *

Your comment or reply (max 5000 characters). Add formatting using Markdown and formulas using LaTeX. For more information see https://openreview.net/faq (https://openreview.net/faq)

Write **Preview**

ECG uses the GPU for inference only (no backpropagation), requiring only ≈16GB VRAM. Alternatively, ECG runs without any GPU via API (≈$3 for 25k examples), making it uniquely accessible to practitioners without ML infrastructure. For reference, NoiseGPT requires 67-130

Characters remaining: 44

Readers* ☑ ACL ARR 2026 January Program Chairs
☑ ACL ARR 2026 January Submission10642 Senior Area Chairs
☑ ACL ARR 2026 January Submission10642 Area Chairs
☑ ACL ARR 2026 January Submission10642 Reviewers Submitted
☑ ACL ARR 2026 January Submission10642 Authors
☑ ACL ARR 2026 January Submission10642 Reviewer Rvqa

Signatures* *signatures*

Edit History

Readers* *readers*

Signatures* ACL ARR 2026 January Submission10642 Authors

Submit **Cancel**

Official Review of Submission10642 by Reviewer YHvT

Official Review by Reviewer YHvT 06 Feb 2026, 13:39 (modified: 19 Feb 2026, 10:36)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YHvT
Revisions (/revisions?id=Dk9RA49qcg)

Paper Summary:
This paper addresses the challenge of detecting mislabeled examples in training data, focusing on "artifact-aligned noise". The authors argue that standard confidence-based methods (like Cleanlab) and training dynamics methods (like AUM) fail in this regime because the model achieves high confidence and low loss on the noisy examples.

The core method involves:

1, Generating structured explanations (evidence + rationale) for every training instance.

2, Embedding these explanations into a semantic vector space.

3, Constructing a k-Nearest Neighbor (kNN) graph in this explanation space.

4, Calculating disagreement between an instance's label and its neighbors' labels.

The key insight is that confident methods may miss the highly-confident instances because the model relies on artifacts for its prediction. And that while input embeddings might cluster based on spurious tokens, explanation embeddings will cluster based on semantic meaning, revealing the label inconsistency.

Summary Of Strengths:
1, The paper highlights a weakness in confidence-based data cleaning: when noise is easy to learn due to artifacts, loss-based signals become anti-correlated with the ground truth.

2, The paper leverages the reasoning capability of LLMs to self-explain the prediction decision. It shifts from detecting outliers in input space, where artifacts dominate, to explanation space, where semantics dominate.

3, The analysis in Section 6 regarding why aggregating multiple signals hurt performance is insightful.

Summary Of Weaknesses:

1, The evaluation's reliance on SST-2 with synthetically injected artifacts. The paper ignores benchmarks for real-world noisy labels, such as AlleNoise [1].

2, The evaluation of the paper misses more recent state-of-the-art methods in noise detection, such as [2] [3].

[1] Rączkowska, Alicja, et al. "AlleNoise: large-scale text classification benchmark dataset with real-world label noise." arXiv preprint arXiv:2407.10992 (2024). [2] Kim, Suyeon, et al. "Learning discriminative dynamics with label corruption for noisy label detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. [3] Wang, Haoyu, et al. "Noisegpt: Label noise detection and rectification through probability curvature." Advances in Neural Information Processing Systems 37 (2024): 120159-120183.

Comments Suggestions And Typos:

I do not see any typos.

Confidence: 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
Soundness: 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.
Excitement: 2 = Potentially Interesting: this paper does not resonate with me, but it might with others in the *ACL community.
Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: After the review process started
Knowledge Of Paper Source: Preprint on arxiv
Impact Of Knowledge Of Paper: Not much
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.
Publication Ethics Policy Compliance: I did not use any generative AI tools for this review

Add: **Author-Editor Confidential Comment** **Official Comment**

New Official Comment

- Please select who should be able to see your comment under "Readers".
- Readers marked as mandatory are required in order to post the comment.
- Program Chairs will not be notified of any of your comments.
- Senior Area Chairs will be notified of all comments.
- All other readers will be notified of all comments.

* denotes a required field

Title

(Optional) Brief summary of your comment.

Authors' Response to Reviewer YHvT

Comment*

Your comment or reply (max 5000 characters). Add formatting using Markdown and formulas using LaTeX. For more information see https://openreview.net/faq (https://openreview.net/faq)

Write | Preview

We thank the reviewer for recognizing the novelty of our approach and the insight of the signal aggregation analysis. We address both concerns below with substantial new experiments.

W1: Reliance on SST-2 with synthetic artifacts; missing AlleNoise.

We have extended evaluation to MultiNLI (3-class NLI, 25k examples), running all methods across 5 random seeds:

SST-2 Artifact-Aligned (AUROC, 5 seeds):

- Explanation kNN: 0.819 ± 0.004
- LLM Mismatch: 0.628 ± 0.004
- Input kNN: 0.549 ± 0.008
- Cleanlab: 0.136 ± 0.025

Characters remaining: 560

Readers*
- [x] ACL ARR 2026 January Program Chairs
- [x] ACL ARR 2026 January Submission10642 Senior Area Chairs
- [x] ACL ARR 2026 January Submission10642 Area Chairs
- [x] ACL ARR 2026 January Submission10642 Reviewers Submitted
- [x] ACL ARR 2026 January Submission10642 Authors
- [x] ACL ARR 2026 January Submission10642 Reviewer YHvT

Signatures* *signatures*

Edit History

Readers* *readers*

Signatures* ACL ARR 2026 January Submission10642 Authors

Submit | Cancel