← Back to **the profile of Zhaoxiang Feng** (/profile?id=~Zhaoxiang_Feng1)

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 ACL ARR 2026 January Submission
January, Zhaoxiang Feng, Zhongyan Luo, Mingyang Yao, Charlie Sun

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
TL;DR: 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).
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)
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.
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
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 10642