TESTING PEER ASSIGNMENT IN SELECTION FROM MASKED VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

A classifier can predict different labels from incomplete versions of the same document. Using agreement between these views can improve selection, but an improvement over maximum confidence leaves open whether the supporting views need to be semantically similar. We test that question by fixing the views, class probabilities, and scoring rule, and changing only whether peers are nearest in MiniLM space or sampled uniformly from the other views. Across six sentiment-classification settings, nearest-peer selection reduces error by 2.56–5.66 percentage points relative to maximum confidence. Uniform-minus-nearest error differences are much smaller, from -0.46 to $+0.50$ points, and all six 90% seed intervals lie inside a study-specific ± 1.5-point band. The two selectors often choose different views but return the same label in 97.0–97.9% of cases. Topic-classification and Naive Bayes studies also yield small peer-assignment effects, while several sensitivity settings remain less precise. The matched control separates the value of an agreement score from the incremental value of its particular semantic neighborhood in masked-view classification.

1 INTRODUCTION

Deleting different words from a movie review can leave a classifier with different impressions of its sentiment. One incomplete version may receive a confident positive prediction while the other versions favor a negative label. A selector that returns the most confident prediction ignores this disagreement. An alternative is to reward a candidate when other views support its label. If those supporters are chosen by semantic similarity, a natural question follows: does choosing similar views improve the decision, or would other views of the same document provide equally useful support?

The distinction concerns the explanation of a gain. A selector that combines confidence with nearby-view support changes both the information used for selection and the way supporting views are chosen. Comparing it only with maximum confidence combines these changes. We separate them by keeping the candidate views, their class probabilities, and the selection score fixed, then replacing the nearest supporting views with uniformly sampled peers from the same document. The resulting error difference measures the effect of peer assignment within this fixed selection problem.

Our study contributes this matched comparison and an empirical account of where peer assignment affects the output. It connects work on agreement and local confidence to a setting where the supporting set can be changed without generating new answers or updating a model. The primary evaluation crosses two sentiment datasets with three transformer classifiers. Additional topic-classification, Naive Bayes, and parameter-sensitivity studies test the same distinction under different data, predictions, and amounts of shared evidence.

In the six primary settings, nearest-peer selection reduces error by 2.56–5.66 percentage points relative to maximum confidence, while uniform-minus-nearest error differences range from -0.46 to $+0.50$ points. Peer assignment often changes the chosen view, yet the two selectors return the same label in 97.0–97.9% of cases. These observations distinguish a useful agreement score from a large benefit of its particular semantic neighborhood. The conclusion concerns the tested masked-view task and default settings; the sensitivity study identifies regimes where the assignment effect remains less precisely estimated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Selection by agreement and local confidence. Soft Self-Consistency replaces hard vote counts with likelihood-based scores to select among sampled actions, improving performance on interactive generation tasks with many distinct valid answers (Wang et al., 2024; 2023). ReCAP instead addresses adaptation under visual distribution shift: it models a local region in representation space and derives a tractable confidence proxy for updating the model (Hu et al., 2025). Both make information beyond a candidate's own confidence useful, through different operations. Their achievements motivate our score and controls. Our question concerns which peers supply that information when the model and scoring rule are held fixed.

Meaning, consistency, and uncertainty. Semantic entropy measures uncertainty over meaning-equivalent answers, addressing the fact that different word sequences can express the same answer (Kuhn et al., 2023; Farquhar et al., 2024). The consistency hypothesis formalizes relationships between agreement and confidence and motivates similarity-based uncertainty estimates (Xiao et al., 2025; Manakul et al., 2023). CoCoA combines confidence and consistency within a minimum-Bayes-risk account of uncertainty (Vashurin et al., 2025). These approaches establish reasons to use multiple outputs when assessing reliability. They also separate two design choices: how to summarize support and which outputs contribute to it. We hold the summary rule fixed for the peer-assignment test, and examine entropy, hard agreement, and multiplicative scoring separately in a classification diagnostic.

From output refinement to a controlled comparison. MAMM-Refine studies error detection, critique, and correction in long-form generation, finding benefits from multi-agent collaboration and from reranking candidate critiques or refinements (Wan et al., 2025). Such evaluations measure the quality of the resulting text. Masked-view classification gives us a complementary experimental setting: all candidate labels and probabilities can be computed once and reused after changing the peer assignment. This permits direct attribution within a fixed pool. The prior-inspired scores in our Naive Bayes study are explicit adaptations to class probabilities, rather than reproductions of these generation systems.

3 Selection and the Peer Assignment Test

3.1 Candidates, labels, and error

The selector chooses among incomplete versions of one document without revising their predictions. For a labeled document (x_i, y_i), the K views are $\mathcal{V}_i = \{v_{i1}, \ldots, v_{iK}\}$. A fixed classifier produces $p_{ij}(\ell) = p_\theta(y = \ell \mid v_{ij})$. View j predicts $\hat{y}_{ij} = \arg\max_\ell p_{ij}(\ell)$ with confidence $c_{ij} = p_{ij}(\hat{y}_{ij})$.

Choosing index $S(i)$ returns the label $\hat{y}_i^S = \hat{y}_{i,S(i)}$. We evaluate the fraction of documents for which that label differs from the original document label:

$$\widehat{\mathrm{err}}_D(S) = \frac{1}{|D|} \sum_{i \in D} \mathbf{1}\left[\hat{y}_i^S \neq y_i\right].$$

We also retain the chosen index. Two selectors can choose different incomplete texts but return the same label, so a change in selection need not change task error.

3.2 Support from other views

A confident prediction receives stronger support when other views assign probability to the same label. We average that probability over m peers and penalize the candidate for the missing support. For candidate j, the peers exclude the candidate itself:

$$N(i,j) \subseteq [K] \setminus \{j\}, \qquad |N(i,j)| = m,$$

Here $[K] = \{1, \ldots, K\}$. The support and selection score are

$$a_{ij}(N) = \frac{1}{|N(i,j)|} \sum_{k \in N(i,j)} p_{ik}(\hat{y}_{ij}),$$

$$q_{ij}(N) = c_{ij} - \lambda\big(1 - a_{ij}(N)\big).$$

For $\lambda > 0$, two equally confident candidates are therefore ordered by peer support. When $\lambda = 0$, the score reduces to confidence alone. Both properties follow directly from the score and will help interpret the comparisons.

MINILM NEAREST embeds every view with a fixed MiniLM encoder and assigns candidate j the m other views with the largest cosine similarity. It then selects

$$S_{\text{sem}}(i) = \arg\max_j q_{ij}(N_{\text{sem}}).$$

UNIFORM PEER replaces that assignment with a uniform sample of m distinct peers from the same $K - 1$ eligible views:

$$N_{\text{unif}}(i,j) \sim \text{Unif}\left\{A \subseteq [K] \setminus \{j\} : |A| = m\right\},$$

$$S_{\text{unif}}(i) = \arg\max_j q_{ij}(N_{\text{unif}}).$$

The uniform draw is seeded and fixed before scoring each candidate. The two selectors share the views, probabilities, peer count, penalty, and rule that breaks score ties by the lower candidate index. Thus only peer membership changes. Figure 1 shows the default case with eight views and three peers.

3.3 CONTRASTS AND UNCERTAINTY

Maximum confidence selects the candidate without using any peer information:

$$S_{\text{conf}}(i) = \arg\max_j c_{ij}.$$

We use it to measure the gain from adding the agreement score, and use UNIFORM PEER to measure the incremental gain from assigning the nearest peers. For evaluation set D and masking seed s, the two signed error differences are

$$d_s^{\text{peer}} = 100\left[\widehat{\text{err}}_{D,s}(S_{\text{unif}}) - \widehat{\text{err}}_{D,s}(S_{\text{sem}})\right], \tag{1}$$

$$d_s^{\text{agree}} = 100\left[\widehat{\text{err}}_{D,s}(S_{\text{conf}}) - \widehat{\text{err}}_{D,s}(S_{\text{sem}})\right]. \tag{2}$$

Multiplication by 100 expresses both differences in percentage points (pp). Positive values favor MINILM NEAREST. An agreement gain together with a peer effect near zero means that the gain survives replacement of the nearest peers under the same score.

For the primary evaluations, the 800-document AG News study, the sensitivity study, and Naive Bayes, we average three seed-level effects and report a 90% t interval with two degrees of freedom. Each seed fixes both the word deletions and one peer assignment per candidate. These intervals summarize the three realized mask-and-peer draws on a fixed document set. The separate 6,000-document AG News result uses a paired document-bootstrap interval.

A study-specific band of ± 1.5 pp, fixed before the evaluations, provides a common scale for the peer effects. We report whether both interval endpoints lie inside it. This descriptive comparison distinguishes small estimates from imprecise ones; it is not a claim of equal performance. The parameter study also reports architecture-level intervals so that averaging cannot conceal uncertain cases.

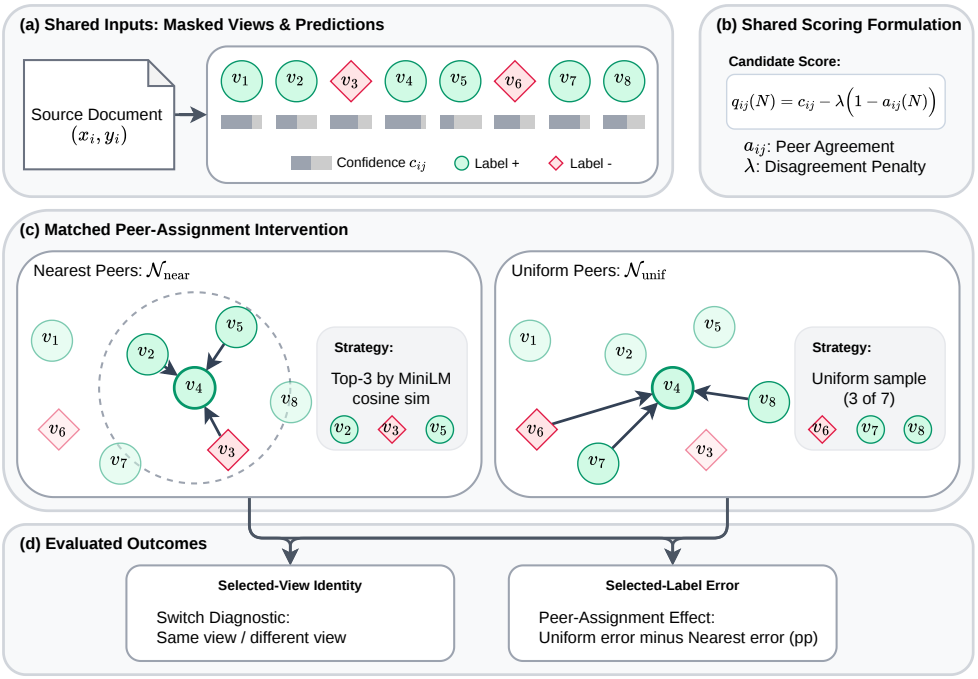


Figure 1: Matched intervention for one document. The same masked views, classifier outputs, and score are used by both selectors; only the peer set changes. MINILM NEAREST uses the three most similar views under MiniLM, whereas UNIFORM PEER samples three of the other views uniformly. We compare both the selected view and the error for the selected label.

3.4 ADDITIONAL CONTROLS

Voting tests whether a hard count of labels is sufficient. Representation controls retain the agreement score: the supplied neural runner implements a classifier-embedding control, while the reported comparison uses normalized word counts in the Naive Bayes study. The latter study also changes how a fixed semantic neighborhood contributes to the score. To express the differences, first average the peers' probability vectors:

$$\pi_{ij}(\ell) = \frac{1}{|N_{\text{sem}}(i,j)|} \sum_{k \in N_{\text{sem}}(i,j)} p_{ik}(\ell)$$

Entropy penalizes a candidate when the mean peer distribution spreads its mass across classes. It uses the score

$$q_{ij}^{\text{ent}} = c_{ij} - \lambda \left[- \sum_{\ell : \pi_{ij}(\ell) > 0} \pi_{ij}(\ell) \log \pi_{ij}(\ell) \right].$$

This quantity is entropy over the fixed task's classes, a classification adaptation of the semantic-uncertainty idea. It does not cluster free-form generations by meaning. Consistency instead counts peers whose most probable class is the candidate's label:

$$a_{ij}^{\text{hard}} = \frac{1}{|N_{\text{sem}}(i,j)|} \sum_{k \in N_{\text{sem}}(i,j)} \mathbf{1}\left[\arg\max_{\ell} p_{ik}(\ell) = \hat{y}_{ij} \right].$$

Table 1: Datasets, model families, and purpose of each study. The primary, 800-document AG News, sensitivity, and Naive Bayes evaluations use three masking seeds. The 6,000-document AG News result is a separate paired-bootstrap evaluation.

Dataset and split	Documents	Classifier	Purpose
Rotten Tomatoes test	1,066	BERT, DistilBERT, RoBERTa	Primary evaluation
SST-2 validation	872	BERT, DistilBERT, RoBERTa	Primary evaluation
AG News test	6,000	DistilBERT	Peer assignment evaluation
AG News test	800	DistilBERT	Independent peer evaluation, separate from the 6,000 document set
Rotten Tomatoes train	1,000	BERT, DistilBERT, RoBERTa	Sensitivity analysis
SST-2 train holdout	2,000	Multinomial Naive Bayes	Bag of words diagnostic

The candidate with the largest a_{ij}^{hard} is selected, with its confidence breaking ties. The CoCoA-style score rewards a candidate when both its own confidence and its peers' support are large:

$$q_{ij}^{\mathrm{cocoa}} = c_{ij}\pi_{ij}(\hat{y}_{ij}).$$

The representation controls retain $q_{ij}(N)$ and replace MiniLM neighbors with cosine neighbors in classifier embeddings or, for Naive Bayes, L2-normalized word-count vectors. The neural classifier-embedding control is included in the supplied runner; the reported Naive Bayes table contains the lexical control. These controls change the source of neighborhoods, while the three formulas above change the use of probabilities within a semantic neighborhood. Appendix A records the sampling and implementation details. These comparisons examine alternative uses of support; the nearest-versus-uniform contrast remains the matched test of peer assignment.

4 Experimental Setup

Each trial starts with a labeled document, produces eight incomplete views, and computes a classifier's probabilities for every view. All selectors then choose from these same outputs, and we compare their returned labels with the document label. Repeating this flow across three seeds reveals how the error difference changes with the masks and peer draws. The studies in Table 1 test the default comparison, its behavior on another task or classifier, and its sensitivity to how much evidence each view and peer set contains.

Constructing incomplete views. We split each text on whitespace and retain each word with probability 0.7, preserving word order and deleting the others. If fewer than four words survive, the runner additionally marks four randomly ranked original positions for retention; the resulting view can contain more than four words. Texts shorter than four words are copied unchanged into all eight views. Thus masking here means word deletion, with no inserted mask token. Views are sampled independently for each document, seed, and view index, and reused by all selectors within that trial.

Tasks and samples. The primary studies classify binary sentiment on the complete Rotten Tomatoes test set (1,066 documents) and SST-2 validation set (872 documents) (Pang & Lee, 2005; Socher et al., 2013). AG News adds four-class topic classification (Zhang et al., 2015). Its 6,000-document evaluation and its 800-document evaluation use nonoverlapping test documents: the smaller sample excludes all rows consumed by the earlier AG News studies. We report these estimates separately. The sensitivity study uses a fresh 1,000-document Rotten Tomatoes training subset; the Naive Bayes study reserves 2,000 SST-2 training documents for evaluation.

Fixed predictions and representations. The sentiment studies use BERT, DistilBERT, and RoBERTa (Devlin et al., 2019; Sanh et al., 2019; Liu et al., 2019); AG News uses DistilBERT. The semantic representation comes from `sentence-transformers/all-MiniLM-L6-v2`

(Wang et al., 2020). Neural inputs are truncated to 192 model tokens. Neighborhood embeddings use attention-mask-weighted mean pooling of the final hidden states followed by L2 normalization, while class probabilities come from the classifier head. Models are loaded from fixed local artifacts in evaluation mode, without updating their parameters. Appendix A.1 lists the checkpoint names and local artifact names.

Sampling and tie rules. The primary neural, sensitivity, and 800-document AG News runs use three fixed masking seeds; Naive Bayes uses a separate set of three seeds. Word-deletion generators use deterministic keys based on the experiment, dataset, document, masking seed, and view index. Uniform-peer generators use a separate key and candidate index and sample without replacement. Maximum-confidence and the two matched peer selectors resolve score ties by the lower view index. Neural voting chooses the most frequent label, resolves a tied label count by the mean confidence of views predicting that label, then by the lower label index. Naive Bayes voting resolves tied counts using the largest individual candidate confidence and then the lower view index. The consistency adaptation also uses candidate confidence to break tied agreement fractions. These implementation differences do not affect the shared tie rule in the nearest-versus-uniform comparison.

A distinct classifier family. To test whether the peer effect depends on transformer predictions, we fit multinomial Naive Bayes on the 65,349 SST-2 training documents outside its 2,000-document holdout. The vocabulary contains the 16,000 most frequent training tokens and the smoothing parameter is $\alpha = 1$. For Naive Bayes, tokens are matches to `[a-z0-9']+` after lowercasing. Vocabulary frequencies and class-conditional counts use only the training complement; the fitted model uses additive smoothing for both class priors and multinomial word likelihoods. MiniLM embeddings still come from the masked text, so changing the classifier family leaves the semantic-neighbor construction intact. This study uses the same view-generation and peer-scoring operations with the fitted classifier's probabilities. Its lexical-neighborhood and score variants provide the additional comparisons defined in Section 3.4.

Default and sensitivity settings. The default uses $K = 8$ views, three peers, and $\lambda = 0.25$. A one-factor-at-a-time study varies word retention over $\{0.5, 0.7, 0.9\}$, peer count over $\{1, 3, 5\}$, and penalty over $\{0, 0.25, 0.5\}$. Counting the shared default once gives seven regimes. This design changes the amount of visible text, the amount of peer evidence, or its influence on selection while leaving the other settings fixed.

5 Results

5.1 Agreement gain and the peer assignment effect

The agreement score improves selection over maximum confidence in all six primary settings. Table 2 reports reductions of 2.56–5.66 percentage points, with the largest reduction for RoBERTa on SST-2. This establishes the performance gain whose attribution we examine. The comparison uses the same classifier and candidate predictions, so it measures a change in selection rather than a change in classifier training.

Replacing the nearest peers with uniformly sampled peers leaves small, mixed error differences: Uniform Peer is better in three settings and worse in three, with effects between -0.46 and $+0.50$ pp. All six 90% seed intervals lie inside the ± 1.5 pp band (Figure 2). The SST-2 DistilBERT interval is the only one to exclude zero and favors Uniform Peer; these separate intervals are descriptive, without a multiplicity-adjusted decision. Voting has a lower mean error than MiniLM Nearest in five settings and differs from it by at most 0.30 pp. The small matched peer effects show that most of the gain over confidence survives the change in supporting views.

AG News gives a similar result on topic classification. The 6,000-document evaluation has a peer effect of 0.07 pp with a paired 90% bootstrap interval of $[-0.07, 0.20]$ pp. The separate 800-document evaluation has an effect of 0.25 pp and a wider seed interval of $[-0.48, 0.98]$ pp. Figure 3 places both estimates alongside the sentiment results while retaining their different sample and uncertainty definitions.

Table 2: Selected-label error across three masking seeds (%; lower is better). The agreement contrast is maximum-confidence error minus MINILM NEAREST error, and the peer contrast is UNIFORM PEER error minus MINILM NEAREST error. Errors and contrasts are rounded to two decimals; contrasts are computed before rounding.

Dataset	Classifier	Max conf.	Voting	Nearest	Uniform	Agreement	Peer
Rotten Tomatoes test	BERT	15.42	12.45	12.16	12.66	3.25	0.50
Rotten Tomatoes test	DistilBERT	14.88	12.16	12.32	12.54	2.56	0.22
Rotten Tomatoes test	RoBERTa	15.98	11.76	12.01	11.88	3.97	−0.13
SST-2 validation	BERT	14.68	10.17	10.40	10.28	4.28	−0.11
SST-2 validation	DistilBERT	14.37	11.24	11.47	11.01	2.91	−0.46
SST-2 validation	RoBERTa	14.11	8.41	8.45	8.49	5.66	0.04
SST-2 train holdout	Naive Bayes	13.00	12.90	12.83	12.88	0.17	0.05

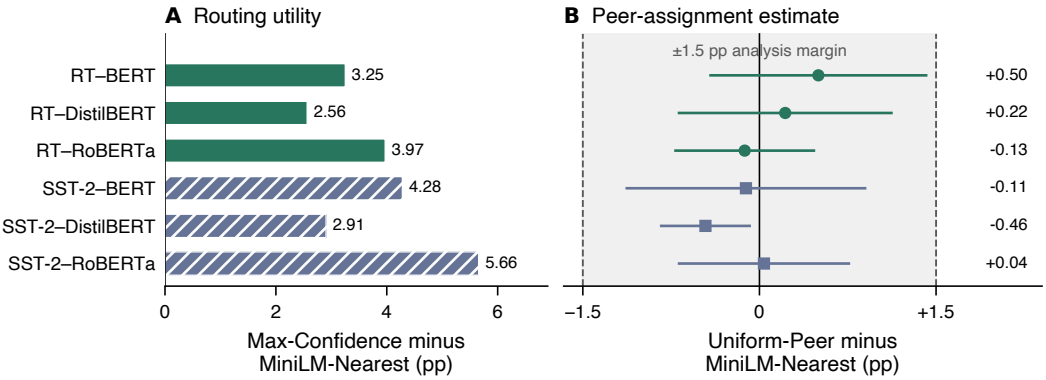


Figure 2: Primary comparisons across the six transformer settings. Panel A reports the reduction in selected-label error relative to maximum confidence. Panel B reports the UNIFORM PEER error minus MINILM NEAREST error contrast, with 90% intervals. Values are in percentage points, and positive values favor MINILM NEAREST. The dashed bounds mark the ± 1.5 pp analysis margin defined in Section 3.

5.2 CHANGING THE CLASSIFIER FAMILY

With Naive Bayes, nearest and uniform peers again produce very similar error: their contrast is 0.05 pp, with a three-seed 90% interval of $[-0.203, 0.303]$ pp. Here the agreement gain over maximum confidence is also small, at 0.17 pp. Voting, the lexical neighborhood, and all three score adaptations remain within 0.12 pp of MINILM NEAREST. Entropy, consistency, and CoCoA-style scoring have comparator-minus-nearest contrasts of $+0.02$, $+0.12$, and $+0.02$ pp, respectively (Appendix D). This study extends the small assignment effect to a different classifier family; it also shows that the size of the agreement gain depends on the underlying predictions.

5.3 SENSITIVITY TO VIEWS, PEERS, AND PENALTY

On the 1,000-document Rotten Tomatoes training subset, the default, higher-retention, and five-peer regimes keep every architecture's interval inside the reporting band. Lower retention, one peer, and higher penalty leave four architecture-level intervals extending outside it. Their pooled effects favor MINILM NEAREST by about 0.6 pp, compared with 0.52 pp at the default (Figure 4). The pooled intervals are narrower than some family intervals; the full table in Appendix B keeps those uncertain cases visible. Thus the default comparison is more precisely bounded than several of the stressed settings.

At $\lambda = 0$, nearest and uniform assignment give exactly the same selections and errors because peer support no longer enters the score. This identity check confirms the expected behavior when the manipulated component has no influence on selection.

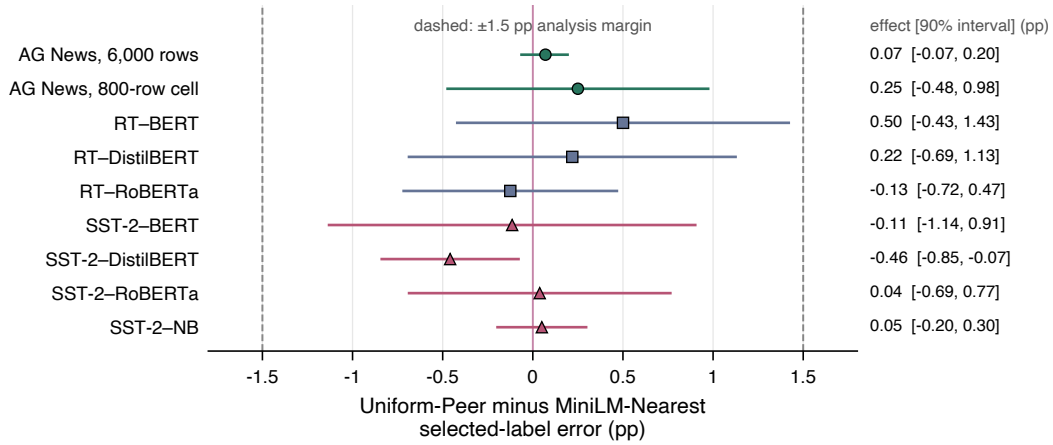


Figure 3: Peer-assignment effects for the six primary transformer settings, the separate 800-document AG News evaluation, the 6,000-document AG News evaluation, and the SST-2 holdout with multinomial Naive Bayes. Points report UNIFORM PEER error minus MINILM NEAREST error. The primary intervals are based on three masking seeds; the 6,000-document estimate uses a paired 90% bootstrap interval. The dashed lines mark the ± 1.5 pp margin.

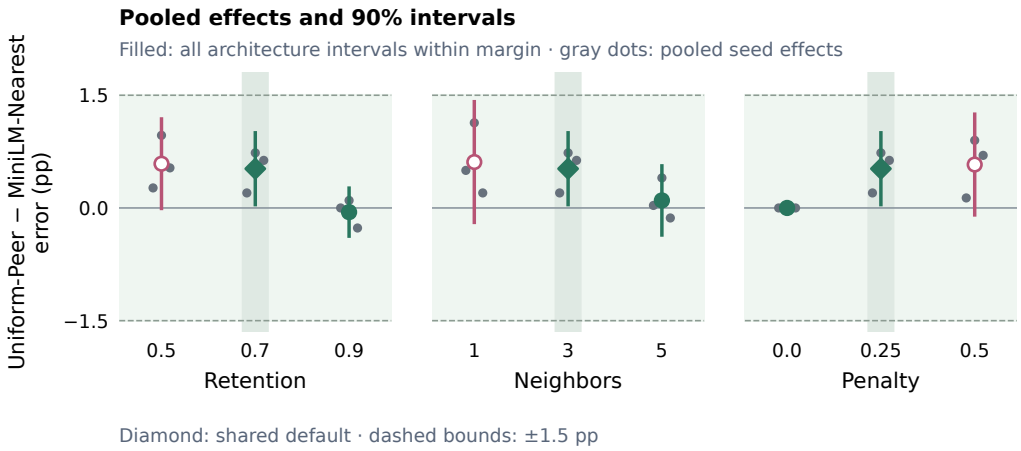


Figure 4: Sensitivity to masking retention, peer count, and penalty strength on 1,000 Rotten Tomatoes training documents. Points show the equal-weight mean across the three classifier families within each masking seed, with 90% intervals across the three seed means. The dashed lines mark the ± 1.5 pp margin. Effects are defined as UNIFORM PEER error minus MINILM NEAREST error; the architecture-level intervals are reported in Appendix B.

6 WHY VIEW CHANGES CAN LEAVE LABELS STABLE

Uniform peers sample the support available in the complete pool of other views. Each eligible view has the same inclusion probability, so the expected sample-average support equals the full-pool average. Conditional on the fixed probability vectors $\mathbf{p}_i$, a uniform sample of h peers without replacement satisfies

$$\mathbb{E}\left[a_{ij}(N_{\text{unif}}^{(h)}) \mid \mathbf{p}_i\right] = \frac{1}{K-1} \sum_{k \neq j} p_{ik}(\hat{y}_{ij}). \tag{3}$$

The nearest-peer rule concentrates this support on similar views, whereas the uniform rule samples it without using similarity. Their candidate scores can differ even when both favor a candidate with the same label. Moreover, selecting the largest score is nonlinear: the expectation in Equation 3 does not imply equal selections or equal error. The relevant empirical question is how often assignment changes a label, and which selector is correct on those changes.

Across the primary settings, nearest and uniform peers select the same view in only 38.0–42.5% of trials, but return the same label in 97.0–97.9%. The intervention therefore changes many selected views while leaving most predictions unchanged. Among the documents with different returned labels, the correctness counts favor MINILM NEAREST in three settings and UNIFORM PEER in the other three. Appendices C and E report the seed effects, agreement rates, and correctness counts.

The two agreement rates therefore measure different events. A peer assignment can change which incomplete view receives the largest score while leaving its predicted label unchanged. It can also change the selected label without changing correctness, or change both the label and correctness. Selected-view identity records every change in the chosen view, whereas selected-label error changes only when the returned label changes. This distinction keeps a change in the selected view from being counted as a change in predictive behavior when the returned class is stable. The high label agreement therefore indicates that most assignment changes occur within a common predicted class, not that the two selectors assign identical scores. The correctness counts provide the remaining directional check: among cases with different returned labels, neither assignment is consistently more accurate across the primary settings.

Voting agrees with MINILM NEAREST on 97.9–98.4% of primary labels, compared with 91.4–93.3% for maximum confidence. Together with the peer comparison, this pattern is consistent with support from other views stabilizing selection. It leaves open whether similar probability vectors, redundant text, or dominance of candidate confidence makes the final label insensitive to the particular supporting set.

7 LIMITATIONS

The result concerns exact-label classification of deliberately incomplete views of the same text. These correlated views differ from independently generated answers, retrieved documents, or collaborating agents. We evaluate selected-label error and selected-view identity; runtime, memory, and computational savings are unmeasured. The three-seed intervals combine mask and peer-assignment variation on fixed documents, giving limited resolution about either source of variation separately. The $1.5\,\mathrm{pp}$ band is a study-specific reporting scale, and several sensitivity intervals remain wider than it.

We do not measure cosine separation, token overlap, or neighborhood reciprocity. Consequently, the small assignment effect could reflect weak contrast between the available peer sets as well as limited value of locality for these predictions. The sensitivity study uses a training split and provides a parameter check, not an independent generalization test. Two classifiers are supplied as local checkpoints without public upstream identifiers, limiting exact external reconstruction (Appendix A.1). Stronger neighborhood contrasts and less redundant views would help distinguish the remaining explanations.

8 CONCLUSION

A gain from using peer support and a gain from choosing semantically nearest peers are different quantities. The matched assignment comparison makes that distinction observable while preserving the candidate predictions and score. In the default masked-view evaluations, nearest-peer selection improves over confidence, but exchanging its peers produces small, mixed changes in error; many chosen views change while their labels remain stable. The additional classifier and topic-classification studies retain the small assignment effects, and the sensitivity study identifies less certain regimes. Evaluating neighborhood membership with an identically scored peer control gives a direct way to test which part of an agreement-based selector contributes to its performance.

Data Availability Statement

Rotten Tomatoes, SST-2, and AG News are public datasets. The anonymous ZIP contains selector runners, sanitized configurations, requirements, and instructions. It excludes data, checkpoints, generated outputs, and identifying metadata. Execution requires external resources; Appendix A.1 lists the available checkpoint identifiers.

Reproducibility Statement

Sections 3–4 specify the trials and comparisons. The appendices provide score formulas, implementation details, seed effects, sensitivity exceptions, and switch counts. The code supplement exposes these procedures but is not a self-contained reproduction of every evaluation.

AI Use Statement

Generative AI tools assisted with drafting and editing this manuscript. The authors are responsible for reviewing the final text and all reported claims and results.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation, 2025. URL https://arxiv.org/abs/2505.20704.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 115–124, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/P05-1015/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment

treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106236–106281, 2025. doi: 10.52202/085713-3546.

David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. MAMM-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9882–9901, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.498. URL https://aclanthology.org/2025.naacl-long.498/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Soft self-consistency improves language models agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 287–301, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.28. URL https://aclanthology.org/2024.acl-short.28/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. *arXiv preprint arXiv:2506.21849*, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

A Auxiliary Scores and Implementation

The auxiliary score definitions used for the Naive Bayes diagnostic are given in Section 3.4. The sampling and tie rules are specified in Section 4; this appendix records the model-artifact details needed to interpret those definitions and reproduce the comparisons.

A.1 Model artifacts

The archived 6,000-document AG News summary reports a paired bootstrap contrast; its original runner and raw outputs are outside the compact code supplement.

The public sentiment checkpoints are `distilbert-base-uncased-finetuned-sst-2-english` and `textattack/roberta-base-SST-2`. The BERT sentiment and AG News DistilBERT configurations identify only local artifacts named `bert-sst2-st` and `distilbert-agnews-st`. They do not provide public upstream identifiers for those two artifacts. The Naive Bayes study uses a fixed MiniLM checkpoint. All models are loaded locally, with no runtime checkpoint selection.

For Naive Bayes, tokens are matches to `[a-z0-9']+` after lowercasing. Vocabulary frequencies and class-conditional counts use only the training complement; the fitted model uses additive smoothing for both class priors and multinomial word likelihoods. MiniLM embeddings still come from the masked text, so changing the classifier family leaves the semantic-neighbor construction intact.

B Sensitivity Results

Table 3 averages the three classifier effects within each seed, then constructs the interval across those three averages. Table 4 retains every individual-architecture interval extending beyond the reporting band. The two tables answer different questions: the first describes the average effect across classifiers, while the second shows where that average would conceal uncertain evidence.

Table 3: Pooled sensitivity effects. Effects use the difference between Uniform and Nearest. Pooled intervals use equal architecture weights and three pooled effects, one for each seed.

Regime	Retention	Peers	Penalty	Pooled effect [90% interval]
Default	0.7	3	0.25	$0.52\,[0.04, 1.00]$
Lower retention	0.5	3	0.25	$0.59\,[-0.01, 1.18]$
Higher retention	0.9	3	0.25	$-0.06\,[-0.38, 0.26]$
One peer	0.7	1	0.25	$0.61\,[-0.19, 1.41]$
Five peers	0.7	5	0.25	$0.10\,[-0.36, 0.56]$
No penalty	0.7	3	0.00	$0.00\,[0.00, 0.00]$
Higher penalty	0.7	3	0.50	$0.58\,[-0.09, 1.25]$

Table 4: Sensitivity intervals for individual architectures. The table lists every architecture interval extending outside the fixed $1.5\,\mathrm{pp}$ reporting margin.

Regime	Architecture	Effect	90% interval
Lower retention	RoBERTa	0.83	$[-0.20, 1.86]$
One peer	DistilBERT	0.33	$[-1.04, 1.71]$
One peer	RoBERTa	0.73	$[-0.53, 2.00]$
Higher penalty	RoBERTa	0.40	$[-1.21, 2.01]$

The default, higher-retention, and five-peer regimes keep all architecture intervals inside $\pm 1.5\,\mathrm{pp}$. Lower retention, one peer, and higher penalty leave the four exceptions listed above. The no-penalty effect is zero by construction. A secondary large-effect screen required at least two architectures to have 95% lower bounds above $+1.5\,\mathrm{pp}$; no regime met that criterion. This separate screen does not resolve the wide intervals in the primary 90% analysis.

C Peer Assignment Effects at the Seed Level

Table 5 reports uniform-minus-nearest error differences in percentage points. The seed columns correspond to the three masking seeds used in the primary evaluations. Within a dataset, the documents are fixed across these seeds; the word deletions and peer draws change. The main-text intervals use these three effects before rounding.

Table 5: Effects of peer assignment at the seed level for the primary transformer settings.

Dataset	Classifier	Seed 1	Seed 2	Seed 3
Rotten Tomatoes	BERT	0.09	0.28	1.13
Rotten Tomatoes	DistilBERT	−0.09	−0.09	0.84
Rotten Tomatoes	RoBERTa	−0.38	−0.28	0.28
SST-2	BERT	−0.57	0.57	−0.34
SST-2	DistilBERT	−0.23	−0.46	−0.69
SST-2	RoBERTa	−0.46	0.34	0.23

D Selector Results with Bag of Words

Table 6 contains all eight selectors on the 2,000-document SST-2 holdout. Each error is averaged across the three seeds. The lexical row changes the representation used to choose peers; entropy, consistency, and CoCoA-style rows retain semantic peers and change the score. The contrast column follows the same sign convention as the main paper: positive values mean higher error than MiniLM Nearest.

Table 6: Selector results using bag of words on the SST-2 holdout.

Selector	Error (%)	Comparator minus Nearest (pp)
Maximum confidence	13.00	+0.17
Voting	12.90	—
Nearest	12.83	0.00
Uniform	12.88	+0.05
Lexical neighborhood	12.75	—
Semantic entropy	12.85	+0.02
Consistency only	12.95	+0.12
CoCoA style	12.85	+0.02

E SELECTION SWITCHES

Tables 7 and 8 separate selected-view identity, returned label, and correctness. Percentages use all documents in a setting and are averaged over three seeds. Count columns give the mean number of documents per seed where nearest and uniform return different labels, so fractional entries are expected. The correctness columns count which selector alone matches the gold label on those disagreements. RT denotes Rotten Tomatoes; confidence and voting agreement are measured against MINILM NEAREST.

Table 7: Agreement rates for selection switches and mean numbers of documents with different labels per masking seed (three seeds).

Setting	Same view	Same label	Different label documents mean/seed
RT/BERT	41.9%	97.8%	23.3
RT/DistilBERT	40.2%	97.4%	27.7
RT/RoBERTa	41.1%	97.4%	28.0
SST-2/BERT	42.5%	97.7%	19.7
SST-2/DistilBERT	38.0%	97.0%	26.0
SST-2/RoBERTa	39.8%	97.9%	18.3
SST-2/Naive Bayes	69.6%	99.5%	9.7
AG News, 800 documents	66.4%	99.5%	3.7

Table 8: Mean correctness counts per masking seed among cases with different labels (three seeds), and label agreement with Nearest.

Setting	Correct only for Nearest mean/seed	Correct only for Uniform mean/seed	Confidence/Nearest label agreement	Voting/Nearest label agreement
RT/BERT	14.3	9.0	92.4%	98.0%
RT/DistilBERT	15.0	12.7	93.3%	98.3%
RT/RoBERTa	13.3	14.7	91.8%	98.0%
SST-2/BERT	9.3	10.3	91.9%	98.2%
SST-2/DistilBERT	11.0	15.0	91.5%	97.9%
SST-2/RoBERTa	9.3	9.0	91.4%	98.4%
SST-2/Naive Bayes	5.3	4.3	99.5%	99.1%
AG News, 800 documents	2.7	0.7	98.6%	99.1%