Affinity and Diversity: A Unified Metric for Demonstration Selection via Internal Representations

Anonymous ACL submission

Abstract

The performance of **I**n-**c**ontext **L**earning (ICL) is highly sensitive to the selected demonstrations. Existing approaches to demonstration selection optimize different objectives, yielding inconsistent results. To address this, we propose a unified metric–affinity and diversity–that leverages ICL model's internal representations. Our experiments show that both affinity and diversity strongly correlate with test accuracies, indicating their effectiveness for demonstration selection. Moreover, we show that our proposed metrics align well with various previous works to unify the inconsistency.

1 Introduction

Language **M**odels (LMs) show **I**n-**C**ontext **L**earning (ICL) ability (Dong et al., 2024), learning to solve tasks without updating model parameters by processing a query along with demonstrations of input-label pairs. The performance of ICL is highly sensitive to the quality of demonstrations (Liu et al., 2021), and previous work has proposed several strategies for selecting better demonstrations.

One prominent approach is to select demonstrations based on their similarity to queries. Here, the similarity is computed by models *independent of the ICL-executed LMs*, e.g., off-the-shelf document retrievers (Rubin et al., 2022), such as BM25 (Robertson and Zaragoza, 2009), fine-tuned document retrievers (Luo et al., 2024), and encoder-based pretrained LMs (Chen et al., 2024). While these previous methods have improved ICL performance, we find that they capture different aspects of demonstration quality and do not converge on a consensus measure (Fig. 5, Left). Developing a unified metric that integrates various metrics leads to a deeper understanding of demonstration quality and further enhances ICL performance.

Therefore, in this paper, we propose a novel approach that *leverages the ICL model's internal representations* to unify previous selection methods.

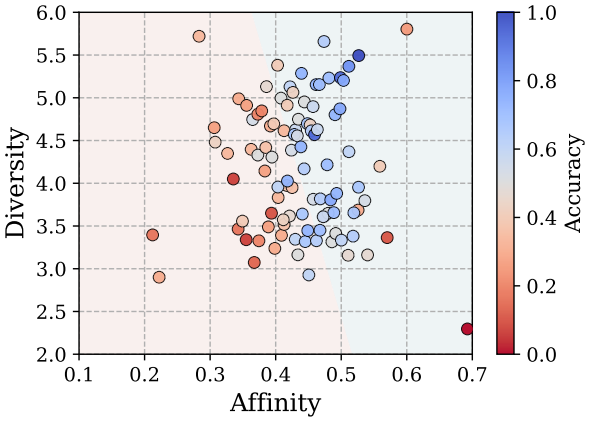

Figure 1: The Affinity and Diversity of the demonstrations in TREC, SST5, TEE on $k = 16$. The colors of the circles refer to the accuracy of the classification tasks. The line and background color refer to the decision boundary to predict labels by affinity and diversity. The larger the affinity and diversity, the higher the accuracy tends to be.

We first identify a self-attention head that is critical for ICL, and on the subspace defined by the $W_Q^\top W_K$ of this attention head, we define two new metrics: (i) *affinity* between a query and demonstrations and (ii) *diversity* among demonstrations. Our experiments show that proposed metrics correlate with existing demonstration selection methods (Fig.5, Left) and are useful for identifying better demonstrations (Fig. 1).

Our contributions are:

- We propose internal representation-based affinity and diversity as a better joint metric on ICL for demonstration selection (§4.2), which unifies the previous selection methods (§5.2).
- We empirically show that previous demonstration selection methods focus on different aspects of selected demonstrations, showing that they are not always positively correlated with other selection methods (§5.3).

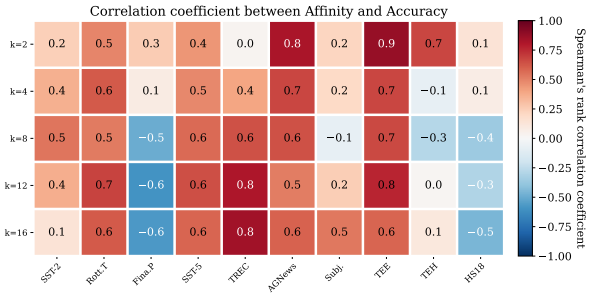

Figure 2: The correlation coefficient between affinity and accuracy

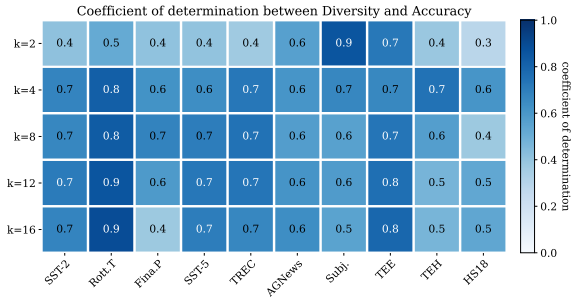

Figure 3: The coefficient of determination between diversity and accuracy

2 Background

2.1 In-context Learning

Given k input-label pairs (*demonstrations*) and a *query* for a classification task, the demonstrations and query are concatenated in natural language form and fed to LMs (e.g., for $k = 2$, "Good movies. Label: Positive. That's too cruel. Label: Negative. I like it. Label: "). Here, ": " serves as a *forerunner token* to concatenate inputs and labels, and trigger the prediction of the label tokens. The LMs return a probability distribution over the next tokens, and ICL selects the token with the highest probability as the final prediction.

2.2 Induction Circuit

An induction circuit is an abstraction of some attention heads to lead the inference of ICL (Elhage et al., 2021), which consists of several interacting attention heads across different layers: (i) *previous token heads*, which copy information from previous tokens to the current token, and (ii) *induction heads*, which attend to tokens based on context and boost the probability of predicting token $[B]$ when $[A][B]...[A']$ is provided as input. In this paper, we find the most effective induction head, and define the affinity-diversity metrics on the $W_Q^\top W_K$ mappings of this head.

2.3 Demonstration Selection Methods

There are two main approaches for retrieving demonstrations. One is to use off-the-shelf retrievers, such as BM25 (Robertson and Zaragoza, 2009) or BGE M3 (Chen et al., 2024). Off-the-shelf retrievers approaches may be sub-optimal since they are not finetuned for specific tasks. Another approach is to train retrievers, e.g., using encoder-based LMs, based on supervision signals from ICL models. To optimize such retrievers, various loss functions (e.g., List-wise Ranking Loss (Li et al., 2023) and InfoNCE Loss (Rubin et al., 2022)) and training strategies (e.g., iterative training or contrastive training) are employed. Note that while learning-based methods learn signals from ICL models during training, they solely rely on the trained retriever during ICL. However, in §5.3, we show that there is no consistent correlation between these previous approaches, leading to disagreement in the selected demonstrations across different objectives and optimization methods, that should be unified for consistent demonstration selection.

3 Proposed Metrics: Affinity, Diversity

Since induction circuits play a crucial role in ICL, we hypothesize that induction circuits can also be used to assess the quality of demonstrations. We first identify induction heads (§3.1) and then compute affinity and diversity in their subspace (§3.2).

3.1 Step 1: Extract Internal Representation

To identify induction heads, we follow Cho et al. (2025): for each attention head h at layer l, we compute $s(h)$, the sum of attention scores from the last token of the query to all the correct label tokens (i.e., tokens that match the ground-truth label of the query) in the demonstrations, and identify "the best induction head" as the head $\hat{h}$ at layer $\hat{l}$ with the highest $s(\hat{h})$.

We then extract the label token representations $\left\{\boldsymbol{d}_{\text{label}}^{(i)}\right\}_{i=1}^{k}$ of each demonstration i and the last token's representation $\boldsymbol{d}_{\text{q}}$ of the query from the best induction head $\hat{h}$. In detail, given a token index j and the hidden state $\boldsymbol{h}_j^{\hat{l}}$ of j-th token from the previous layer of $\hat{h}$ after the layer normalization, we extract the inner representation of j-th token as follows:

$$\boldsymbol{d}_j = W_Q^{\hat{h},\top} W_K^{\hat{h}} \boldsymbol{h}_j^{\hat{l}}, \tag{1}$$

where $W_Q^{\hat{h}}$ and $W_K^{\hat{h}}$ are the query projection and

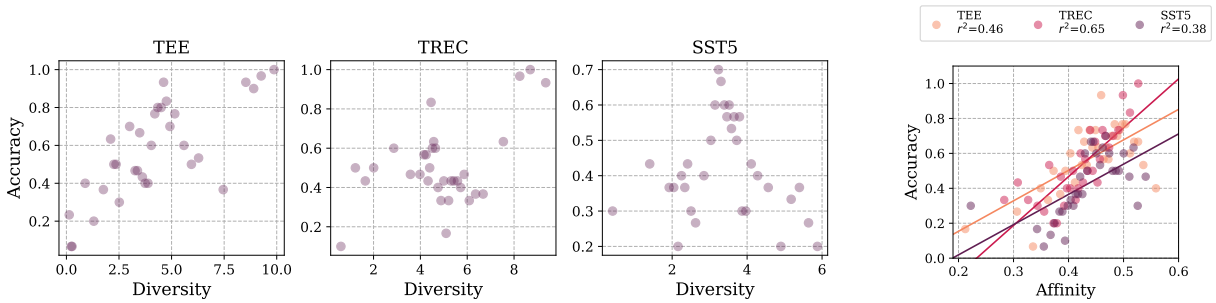

Figure 4: **Left**: The tendency of diversity to accuracy on $k = 16$. **Right**: The tendency of affinity to accuracy on $k = 16$.

key projection of attention head $\hat{h}$.

3.2 Step 2: Compute Affinity and Diversity

3.2.1 Affinity

We define affinity as the mean of the cosine similarity between all the label token representations and the query representation as follows:

$$\mathrm{Aff}\left[\boldsymbol{d}_{\mathrm{q}}, \left\{\boldsymbol{d}_{\mathrm{label}}^{(i)}\right\}_{i=1}^{k}\right] = \frac{1}{k}\sum_{i=1}^{k}\cos\left[\boldsymbol{d}_{\mathrm{q}}, \boldsymbol{d}_{\mathrm{label}}^{(i)}\right] \tag{2}$$

3.2.2 Diversity

We define diversity as the variance (the trace of the covariance matrix) across the label token representations of all demonstrations as follows:

$$\mathrm{Div}\left[\left\{\boldsymbol{d}_{\mathrm{label}}^{(i)}\right\}_{i=1}^{k}\right] = \frac{1}{k}\mathrm{tr}\left[\mathop{\mathbb{D}}_{i\in[1,k]}\left[\boldsymbol{d}_{\mathrm{label}}^{(i)}\right]\right] \tag{3}$$

Here, $\mathbb{D}$ represents the covariance operator.

4 Experiments

We demonstrate that affinity and diversity serve as effective metrics for demonstration selection.

4.1 Experimental Settings

Model. We conduct experiments on Llama 3 8B (AI@Meta, 2024). The model parameters are loaded from `HuggingFace`.

Dataset. For all experiments, we use 10 classification datasets. For details of the dataset, please refer to Appendix A.1. We use $k = 2, 4, 8, 12, 16$, and input sequences are built by library StaICC (Cho and Inoue, 2025).

Evaluation. For each test instance, we randomly sample k demonstrations, run ICL, and record the prediction. Next, we sort all instances based on their affinity or diversity values and group them into bins of 30 instances each. For each bin, we calculate the average affinity or diversity, and also the accuracy. These averages and accuracies are then used to compute the correlation between the proposed metrics and accuracy. For affinity, we use Spearman's rank correlation coefficient. For diversity, we apply Ridge regression with a Laplacian kernel to capture non-linear relationships, with the R^2 coefficient as the measure of goodness-of-fit.

4.2 Main Results: Affinity and Diversity Measure the Effectiveness of Demonstrations

The Spearman's rank correlation coefficient for affinity and R^2 coefficient for diversity are shown in Fig. 2 and Fig. 3. These indicate that affinity shows a positive correlation across various tasks, and diversity achieves a high R^2 coefficient in nearly all tasks. Fig. 4 (Left) and Fig. 4 (Right) provide examples of diversity/affinity-accuracy scatter plots, which further support these trends.

5 Analysis

Next, we show that affinity and diversity strongly correlate with the scores from previous demonstration selection methods, addressing the inconsistency issue in the previous work (§5.2). We also show that the scores from previous demonstration selection methods disagree with each other (§5.3). Moreover, the demonstrations selected by previous work practically improve affinity, but not diversity. These observations suggest that it is required a new demonstration selection method based on affinity and diversity (§5.4).

5.1 Experimental Setup

We use three previous methods to compare affinity and diversity, BM25 and BGE-M3 for training-free methods, and EPR for training methods. For

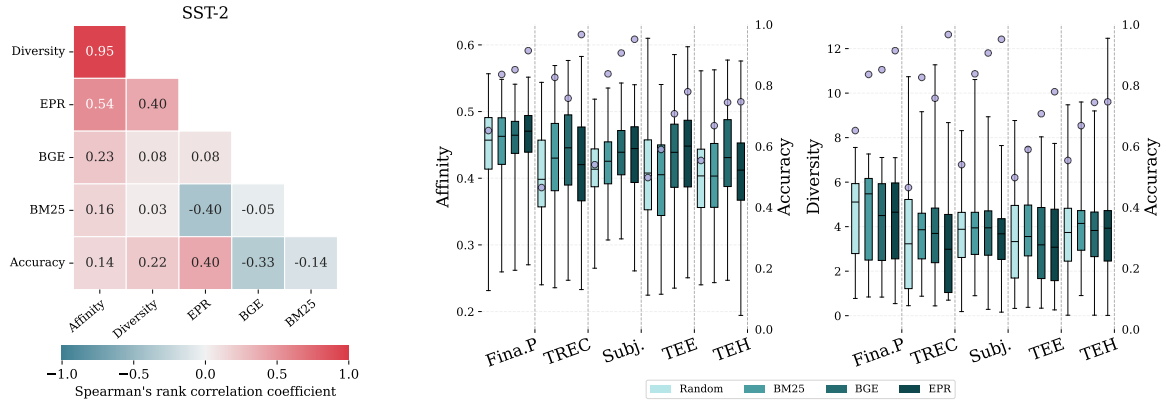

Figure 5: **Left**: The Spearman's rank correlation coefficient of the similarity scores, affinity, diversity, and accuracy of $k = 16$ on SST2. **Middle**: The affinity of selected demonstrations by each selection method on $k = 2$. **Right**: The diversity of selected demonstrations by each selection method on $k = 2$.

details of the previous methods, please refer to Appendix A.2. Other settings are the same as §4.1.

5.2 Affinity and Diversity Correlate with the Score of Previous Methods

We measure the similarity scores from the previous methods using the same prompts described in §4.1 and compute the Spearman's rank correlation among these similarity scores, accuracy, affinity, and diversity. The results of $k = 16$ on SST2 are shown in Fig. 5 (Left), where both affinity and diversity show a positive correlation with the similarity scores and accuracy. This indicates that affinity and diversity consistently measure the effectiveness of the demonstrations in terms of accuracy.

5.3 Previous Methods are Not Always Positively Correlated with Each Other

Meanwhile, no consistent positive correlation is observed among the similarity scores from previous selection methods. Even worse, in some cases, negative correlations (e.g., EPR and BM25) are observed, suggesting that they may not consistently produce optimal results. EPR shows a positive correlation with BGE, likely due to their reliance on a BERT-based encoder.

5.4 Better Selection of Demonstrations Improves Affinity and Diversity

In this section, we evaluate the previous demonstration selection methods on the proposed affinity and diversity, and show that affinity and diversity are improved by the previous methods. We build prompts with the same query as §4.1 select demonstrations by previous methods and input them into an LM to measure the accuracy, affinity, and diversity.

The results are shown in Fig. 5 (Middle) for the affinity and accuracy, where better accuracy co-occurrence with greater affinity, while, when no improvement is observed in the affinity, then no accuracy can be observed in the accuracy, on some of the scenarios. Moreover, the results of diversity are shown in Fig. 5 (Left), with a less significant co-occurrence between better accuracy and greater diversity. We infer that the reason is: existing methods select demonstrations based on their similarity to the query, without a focus on the diversity, showing a possibility towards better selection methods based on the joint metric of affinity and diversity. Due to space limitations and computational resources, we leave the demonstration selection method as future work.

6 Conclusion

In summary, we propose affinity and diversity to evaluate demonstration selections in the ICL scenario. Our experiments show that affinity and diversity consistently measure the effectiveness of the demonstration well, raising the possibility of better demonstration selection methods.

7 Limitations

Due to computability limitations, we are not able to compare the performance of affinity and diversity with the learning-based retriever for diversity or order of demonstrations.

References

AI@Meta. 2024. Llama 3 model card.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Hakaze Cho and Naoya Inoue. 2025. Staicc: Standardized evaluation for classification task in in-context learning. *arXiv preprint arXiv:2501.15708*.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. 2025. Revisiting in-context learning inference circuit in large language models. In *The Thirteenth International Conference on Learning Representations*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. *Preprint*, arXiv:2305.04320.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Preprint*, arXiv:2401.11624.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. *Preprint*, arXiv:2112.08633.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Experimental Details

A.1 Datasets

We build ICL-formed test inputs from 10 classification tasks datasets: GLUE-SST2 (SST2) (Socher et al., 2013), Rotten tomatoes (Rott.T) (Pang and Lee, 2005), Finacial Phrasebank (Fina.P) (Malo et al., 2014), Stanford Sentiment Treebank (SST5) (Socher et al., 2013), TREC (TREC) (Li and Roth, 2002; Hovy et al., 2001), AGNews (AGNews) (Zhang et al., 2015), Subjective (Subjective) (Conneau and Kiela, 2018), Tweet Eval Emotion (TEE) (Mohammad et al., 2018), Tweet Eval Hate (TEH) (Basile et al., 2019), Hate Speech 18 (HS18) (de Gibert et al., 2018).

A.2 Previous methods

We conduct experiments to compare affinity and diversity using previous methods:

- BM25: selecting the demonstrations with the similarity score to query, by an expanded TF-IDF (BM25).
- BGE M3: selecting the demonstrations with the most cosine similarity between the encoding vectors of the demonstrations and query, by BGE M3. The model parameters are loaded from `HuggingFace`.
- **E**fficient **P**rompt **R**etrieval (EPR) (Rubin et al., 2022): selecting the same way as BGE M3, by the dense encoder trained to retrieve a better demonstration with each ICL datasets.

B Other datasets experiment results

The results of most experiments in the main text on other datasets are shown in Fig. 6, 7, 8, 9, 10.

C Statements

C.1 License for Artifacts

Models Llama 3 8B is under its specific license.

Datasets We list the open-source license for the datasets used in this paper as follows:

- CC-by4.0: Tweet eval emotions, Tweet eval hate
- CC-BY-NC-SA-3.0: Financial phrasebank
- CC-BY-SA-3.0: Hate speech 18
- BSD: TREC, Subjective
- Unknown: GLUE-SST2, Rotten tomatoes, Stanford sentiment treebank, AGNews

C.2 Statistics For Data

We list the number of examples of datasets used in this paper as follows Table 1.

C.3 AI Agent Usage

AI Agents are only used for writing improving and grammar checking in this paper.

Table 1: Raw dataset split size for each sub-dataset.

	SST2	Rott.T	Fina.P	SST5	TREC	AGNews	Subj.	TEE	TEH	HS18
Demonstration Set	4096	4096	512	4096	4096	4096	4096	4096	3192	4096
Test Set	512	512	512	512	512	512	512	512	512	512

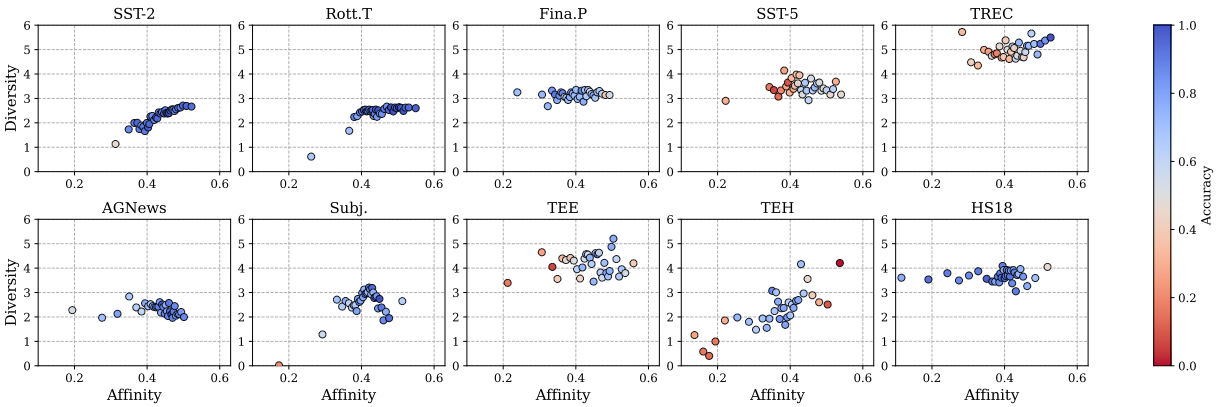

Figure 6: The Affinity and Diversity of the demonstrations. Colors refer to the accuracy of all classification tasks on $k = 16$.

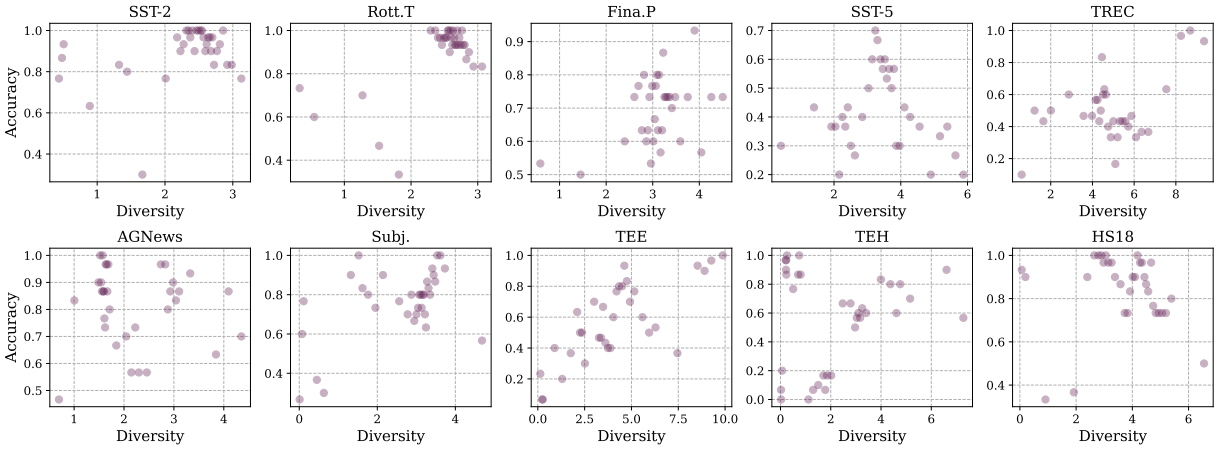

Figure 7: The tendency of diversity to accuracy on $k = 16$.

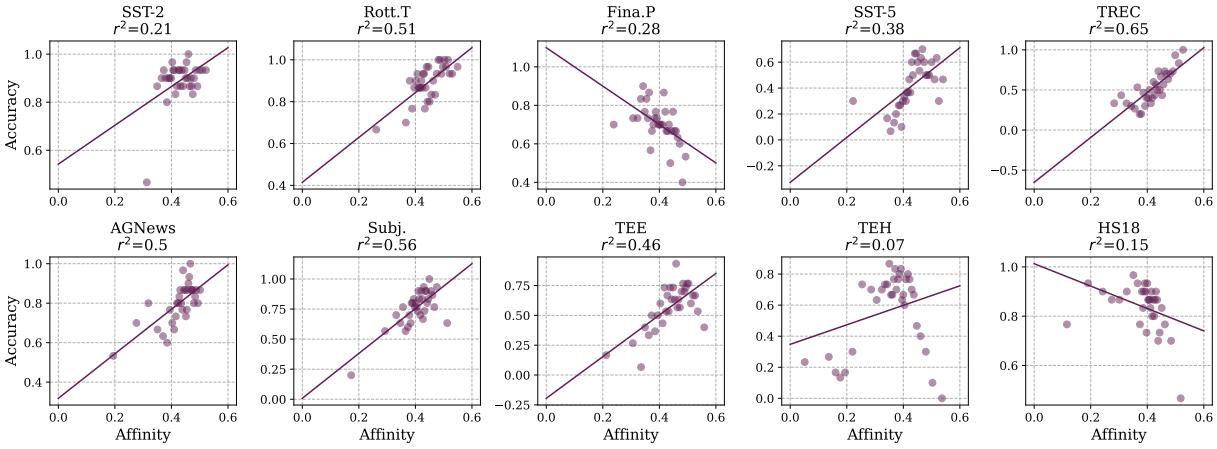

Figure 8: The tendency of affinity to accuracy on $k = 16$.

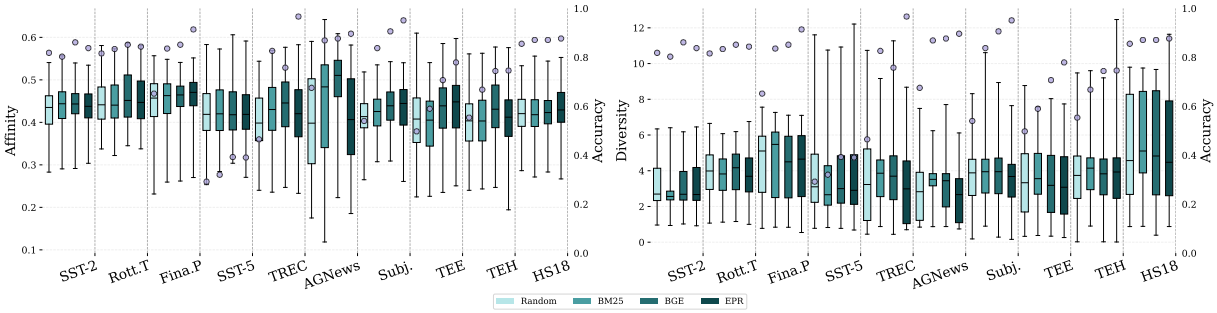

Figure 9: **Right**: The affinity of selected demonstrations by each selection method on $k = 2$. **Right**: The diversity of selected demonstrations by each selection method on $k = 2$.

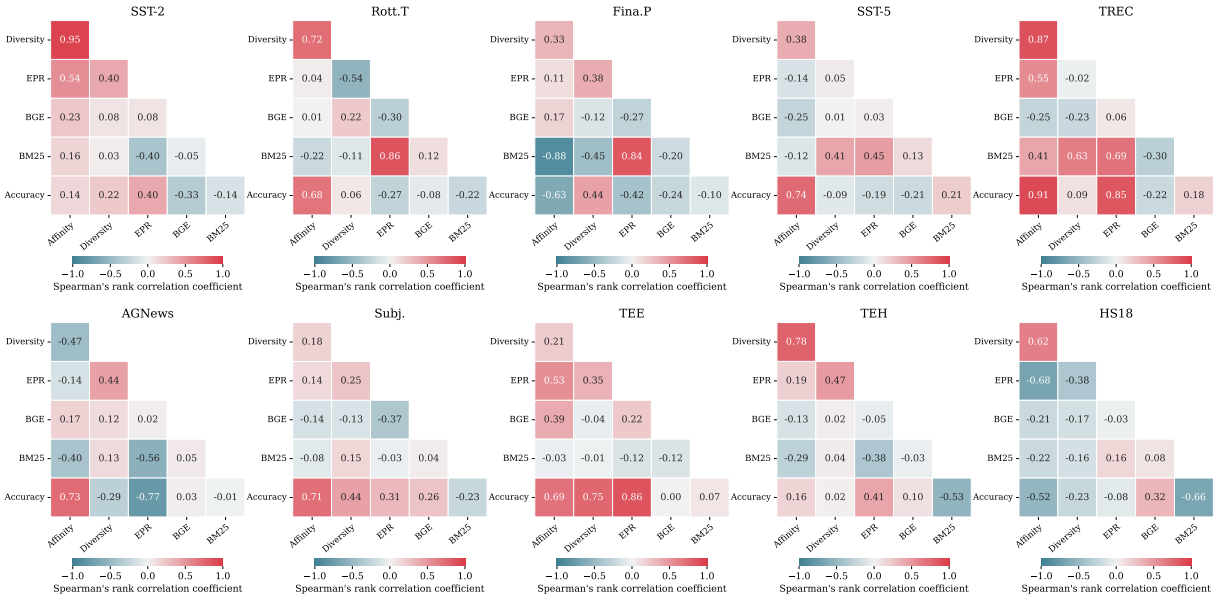

Figure 10: The Spearman's rank correlation coefficient of the similarity scores, affinity, diversity, and accuracy of $k = 16$