

# Retrieval-Augmented Few-shot Text Classification

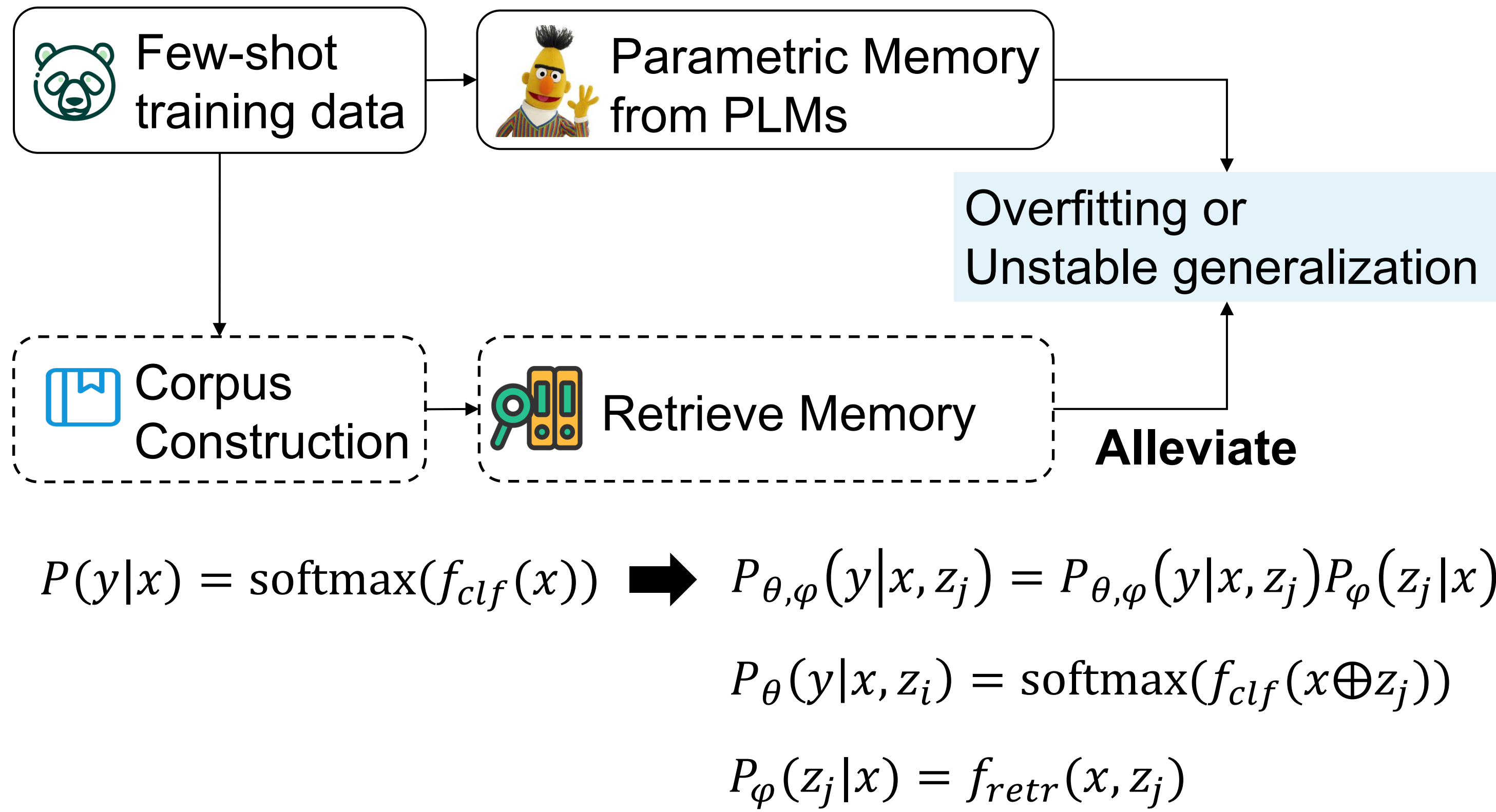
Guoxin Yu<sup>1,2</sup>, Lemao Liu<sup>3\*</sup>, Haiyun Jiang<sup>3</sup>, Shuming Shi<sup>3</sup>, Xiang Ao<sup>1,3\*</sup>

1 Key Lab of IIP of Chinese Academy of Sciences, Institute of Computing Technology.  
2 University of Chinese Academy of Sciences. 3 Tencent AI Lab, China.



## 1 Introduction

- Training numerous parameters of PLMs on scarce data is prone to produce over-fitting and unstable generalization.
- Retrieval-based methods have shown the capability to incorporate retrieved memory alongside parameters for better generalization.



## 2 Challenges

- Static retrieval whose metric is not task-specific, such as BM25 or TF-IDF, cannot be reliable for retrieving really helpful samples for target prediction.

$$P_{\phi}(z_j|x) = f_{retr}(x, z_j) = \text{sim}(x, z_j)$$

- Joint learning-based retrieval suffers from the gradient vanishing problem during the optimizing process, since its retrieval metric is updated towards the downstream task by minimizing the standard cross-entropy loss.

$$P_{\phi}(z_j|x) = f_{retr}(x, z_j) = \frac{\exp(x \cdot z_j^T)}{\sum_{j=1}^m \exp(x \cdot z_j^T)}$$

- The gradient norm of the joint learning-based retriever exceeds the threshold of  $1e^{-6}$  for only about 40% of the steps.

- We aim to meet the challenge of weak supervisory signals for the retriever and insufficient data.

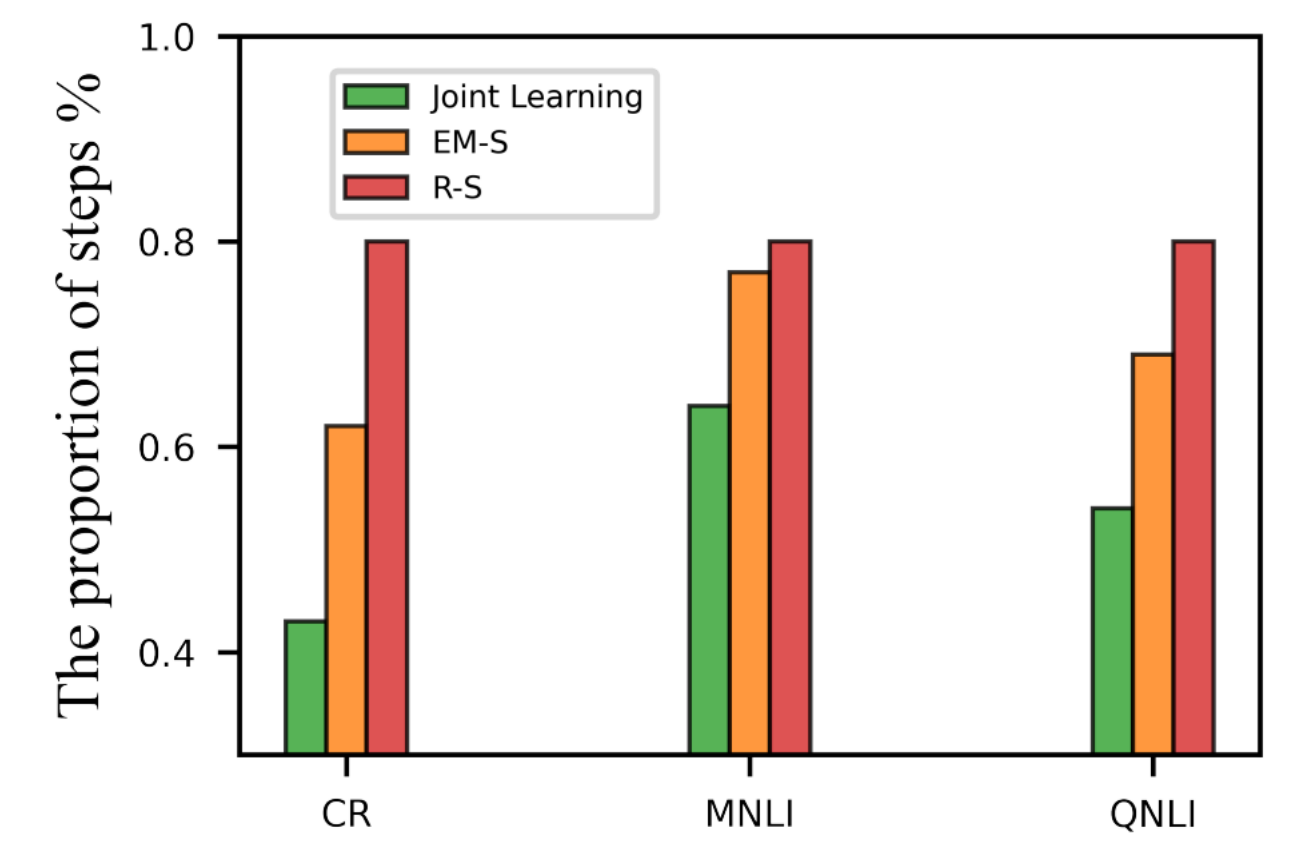


Figure 2: The proportion of steps in which the average gradient of retriever's all parameters is more than  $1e^{-6}$ .

## 3 Retrieval with EM-L and R-L

- EM-based Loss (EM-L) considered  $z_j$  as a latent variable and alternates between an E-step and a M-step until convergence.

- The Expectation-step computes the conditional probabilities:

$$P_{\theta, \phi}(z_j|x, y) = \frac{P_{\theta, \phi}(y|x, z_j)P_{\phi}(z_j|x)}{\sum_{j=1}^m P_{\theta, \phi}(y|x, z_j)P_{\phi}(z_j|x)}$$

- The Maximization-step updates the parameters by maximizing the expected log-likelihood:

$$P_{\theta, \phi}(y|x) = P(z_i|x|y|\theta_i) \cdot \log P(y|x|z_i|\theta)$$

$$\mathcal{L}_{\text{loss}_{\text{em}}} = \sum_i^N \sum_j^M [P(z_j|x_i, y) \cdot \log P(y|x_i, z_j)][c]$$

$$\theta_{i+1} = \arg\max_{\theta} P(z_i|x|y|\theta_i) \cdot \log P(y|x|z_i|\theta)$$

- Ranking-based Loss (R-L) considers the process of retrieving  $z_j$  as a ranking task.

- R-L employs a ranking loss to enhance the consistency between  $P_{\theta}(y|x, z_i)[y_i]$  and  $P_{\phi}(z_i|x)$  and provide more direct signals to the retriever.

$$\mathcal{L}_{\text{loss}_{\text{rank}}} = \sum_i^N \sum_j^M \max(0, \text{margin} + P(z_j|x_i) - P(y|x_i, z_j)[c])$$

$$\mathcal{L}_{\text{loss}_{\text{cls}}} = - \sum_i^N \sum_j^M \log[P(z_j|x_i) \cdot P(y|x_i, z_j)][c]$$

$$\mathcal{L}_{\text{loss}} = \lambda \cdot \mathcal{L}_{\text{loss}_{\text{rank}}} + (1 - \lambda) \cdot \mathcal{L}_{\text{loss}_{\text{cls}}}$$

- Both of EM-L and R-L aim examples from a limited space more effectively and prioritize more beneficial examples for downstream tasks.

## 4 Experiment

- EM-L and R-L approaches train the retriever more effectively than static retrieval and joint learning-based retrieval.
- The advantages of EM-L and R-L are more pronounced on challenging tasks, such as QQP, QNLI, and LAP.

Model	Single Sentence				Sentence Pair				ABSA		Avg.
	SST-2	MR	CR	TREC	QQP	QNLI	MNLI	SNLI	RES	LAP	
Prompt Learning with RoBERTa-Large											
Vanilla	84.84 <sub>(6.80)</sub>	77.88 <sub>(7.90)</sub>	88.36 <sub>(2.89)</sub>	87.20 <sub>(7.70)</sub>	67.09 <sub>(6.70)</sub>	64.25 <sub>(7.45)</sub>	60.69 <sub>(4.08)</sub>	64.56 <sub>(4.08)</sub>	72.05 <sub>(4.08)</sub>	71.81 <sub>(2.88)</sub>	73.87
Static	88.60 <sub>(4.10)</sub>	83.67 <sub>(6.80)</sub>	87.06 <sub>(3.84)</sub>	90.95 <sub>(1.36)</sub>	68.31 <sub>(7.70)</sub>	66.27 <sub>(4.98)</sub>	60.38 <sub>(6.70)</sub>	68.17 <sub>(5.62)</sub>	70.95 <sub>(5.46)</sub>	73.01 <sub>(3.03)</sub>	75.74
Joint	90.71 <sub>(1.20)</sub>	85.83 <sub>(2.40)</sub>	86.76 <sub>(6.50)</sub>	90.57 <sub>(4.17)</sub>	67.26 <sub>(4.40)</sub>	63.15 <sub>(7.16)</sub>	61.95 <sub>(4.65)</sub>	67.64 <sub>(5.80)</sub>	71.07 <sub>(2.97)</sub>	73.32 <sub>(2.26)</sub>	75.83
EM-L	<b>91.31</b> <sub>(1.30)</sub>	<b>87.58</b> <sub>(1.40)</sub>	<b>90.00</b> <sub>(0.90)</sub>	<b>92.13</b> <sub>(1.41)</sub>	<b>74.41</b> <sub>(0.74)</sub>	<b>67.66</b> <sub>(3.77)</sub>	<b>64.85</b> <sub>(3.21)</sub>	<b>69.52</b> <sub>(3.69)</sub>	<b>73.74</b> <sub>(3.46)</sub>	<b>76.02</b> <sub>(1.90)</sub>	<b>78.72</b>
R-L	<b>91.58</b> <sub>(1.30)</sub>	<b>87.47</b> <sub>(0.09)</sub>	<b>89.93</b> <sub>(1.70)</sub>	<b>92.86</b> <sub>(1.21)</sub>	<b>73.79</b> <sub>(2.28)</sub>	<b>67.62</b> <sub>(5.79)</sub>	<b>66.04</b> <sub>(3.18)</sub>	<b>73.08</b> <sub>(4.59)</sub>	<b>76.79</b> <sub>(2.60)</sub>	<b>75.59</b> <sub>(1.51)</sub>	<b>79.46</b>
Fine-tune RoBERTa-Large											
Vanilla	81.59 <sub>(4.50)</sub>	73.59 <sub>(9.90)</sub>	81.63 <sub>(4.08)</sub>	85.95 <sub>(5.57)</sub>	61.42 <sub>(8.19)</sub>	57.20 <sub>(2.09)</sub>	59.90 <sub>(5.72)</sub>	59.19 <sub>(5.58)</sub>	69.21 <sub>(4.14)</sub>	71.06 <sub>(5.11)</sub>	70.07
Static	81.99 <sub>(10.8)</sub>	72.69 <sub>(5.05)</sub>	82.75 <sub>(5.50)</sub>	87.02 <sub>(3.25)</sub>	60.23 <sub>(9.60)</sub>	57.11 <sub>(3.90)</sub>	54.69 <sub>(4.98)</sub>	62.65 <sub>(5.10)</sub>	70.48 <sub>(8.74)</sub>	71.37 <sub>(3.03)</sub>	70.10
Joint	83.49 <sub>(3.20)</sub>	74.89 <sub>(2.90)</sub>	80.63 <sub>(5.42)</sub>	86.33 <sub>(3.17)</sub>	63.50 <sub>(8.08)</sub>	57.66 <sub>(2.69)</sub>	60.99 <sub>(4.98)</sub>	61.01 <sub>(5.80)</sub>	70.23 <sub>(3.57)</sub>	70.62 <sub>(4.47)</sub>	70.94
EM-L	<b>85.38</b> <sub>(1.30)</sub>	<b>75.80</b> <sub>(2.20)</sub>	<b>83.81</b> <sub>(5.36)</sub>	<b>89.36</b> <sub>(2.64)</sub>	<b>65.70</b> <sub>(8.17)</sub>	<b>60.93</b> <sub>(1.56)</sub>	<b>62.24</b> <sub>(3.12)</sub>	<b>65.25</b> <sub>(3.20)</sub>	<b>71.64</b> <sub>(3.36)</sub>	<b>72.69</b> <sub>(3.18)</sub>	<b>73.27</b>
R-L	<b>84.69</b> <sub>(2.29)</sub>	<b>75.35</b> <sub>(2.20)</sub>	<b>83.17</b> <sub>(3.22)</sub>	<b>88.92</b> <sub>(3.81)</sub>	<b>70.53</b> <sub>(2.68)</sub>	<b>61.37</b> <sub>(0.12)</sub>	<b>62.18</b> <sub>(1.72)</sub>	<b>66.31</b> <sub>(3.30)</sub>	<b>73.28</b> <sub>(3.13)</sub>	<b>72.69</b> <sub>(3.01)</sub>	<b>73.85</b>

Table 1: Comparison results on 16-shot text classification. “Vanilla” denotes methods without retrieval, which only consists of a sentence encoder and a classifier. “Static” and “Joint” are static retrieval and joint learning-based retrieval, which are introduced in §2. “EM-L” and “R-L” are methods implemented with our proposed new objectives. All the reported results are average Accuracy and the standard deviation in the subscript.

## 5 Analysis

- Higher Kendall's  $\tau'$  of EM-L and R-L in 16-shot and 8-shot text classification indicates that they could prioritize more helpful examples according to their corresponding metrics.

Kendall's $\tau'$	SST-2	CR	QQP	QNLI	RES
Static	0.5344	0.5837	0.4307	0.5312	0.47857
Joint	0.5413	0.6129	0.4776	0.5937	0.4732
EM-L	<b>0.6853</b>	<b>0.6451</b>	<b>0.6265</b>	<b>0.7500</b>	<b>0.6598</b>
R-L	<b>0.7442</b>	<b>0.6562</b>	<b>0.6057</b>	<b>0.7185</b>	<b>0.6125</b>

Table 2: Kendall's  $\tau'$  of  $P_{\phi}(z_j|x_i)$  and  $P_{\theta}(y|x_i, z_j)[y_i]$ .

Accuracy	SST-2	MR	TREC	QQP
Vanilla	80.22	60.71	86.05	64.27
Static	76.58	67.51	86.94	60.30
Joint	85.41	71.01	86.57	61.92
EM-L	<b>87.30</b>	<b>78.75</b>	<b>87.52</b>	<b>67.90</b>
R-L	<b>89.79</b>	<b>77.38</b>	<b>88.78</b>	<b>66.77</b>

Table 3: Comparison results on 8-shot text classification. Standard deviations are omitted to save space.

- EM-L and R-L maintain sustaining advantages and stability as the number of retrieval examples varies, which verifies their stronger supervision signals.

