SAFE Standard Deviation Analysis:

Datasets	SST2		Amazon		SUBJ		PC		TREC		QQP	
Runs	50	100	50	100	50	100	50	100	50	100	50	100
Base	2.51	2.53	1.67	1.90	2.52	2.36	1.71	1.92	2.81	3.07	3.20	2.60
EDA	2.34	2.11	1.69	1.92	2.05	2.20	2.43	1.50	2.19	2.22	2.30	2.22
SAFE	1.94	2.12	2.20	2.37	1.94	3.11	1.31	1.23	2.37	2.70	2.02	2.09

Table 1: Standard Deviation Across Several Datasets for 50 and 100 times Run

Table 1 presents the standard deviation results for three augmentation techniques—Base, EDA, and SAFE—across several datasets, including SST2, Amazon, SUBJ, PC, TREC, and QQP, Here we randomly selected 500 samples from each dataset with run durations of 50 and 100 iterations. For the 50-run duration, SAFE consistently outperforms both EDA and Base, showing lower standard deviation values in the SST2, SUBJ, PC, and QQP datasets. When considering 100 runs, SAFE also outperforms Base in SST2, PC, TREC, and QQP. These findings indicate that SAFE produces results that are less variable and more stable overall.

Box Plot: Accuracy and Standard Deviation Analysis

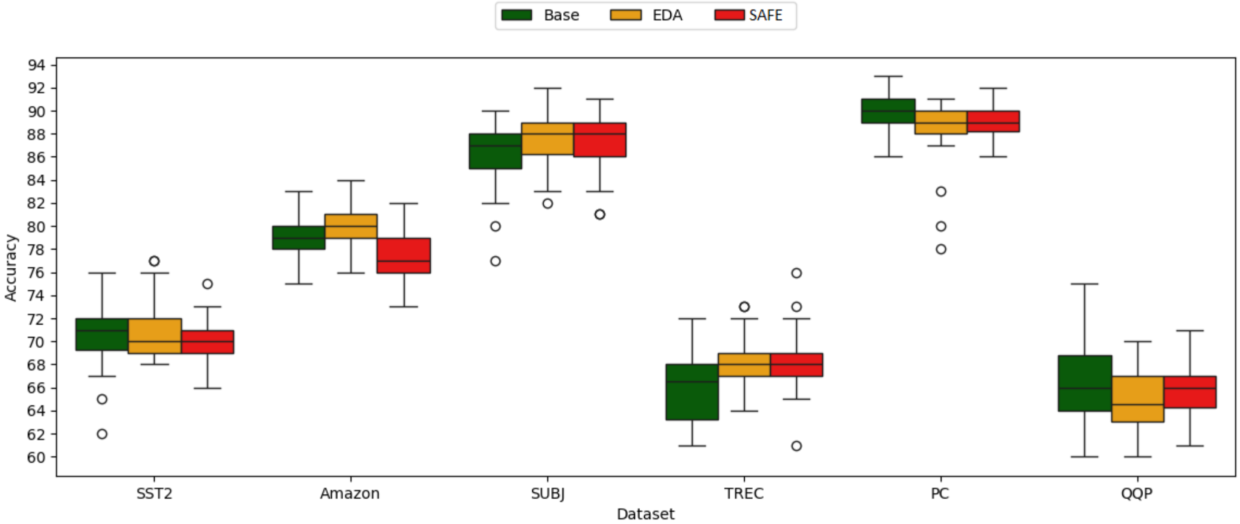

Figure 1: Accuracy Distribution and Standard Deviation Analysis of Data Augmentation Techniques Across different datasets for 50 Runs

The SAFE data augmentation technique showcases remarkable strengths in both 50 runs, consistently delivering stable results across various datasets. In 50 runs, SAFE DA achieves low standard deviations of 1.94 in SST2 and 2.02 in QQP, indicating superior reliability, especially in challenging tasks. It maintains excellent accuracy (87-91%) in the PC dataset with minimal variability (1.31/1.23) and exhibits fewer extreme outliers, particularly in TREC. The result highlight SAFE DA's exceptional stability in complex classification tasks and sentiment analysis, making it a reliable choice for production environments where consistency and performance are crucial, even under limited computational resources.