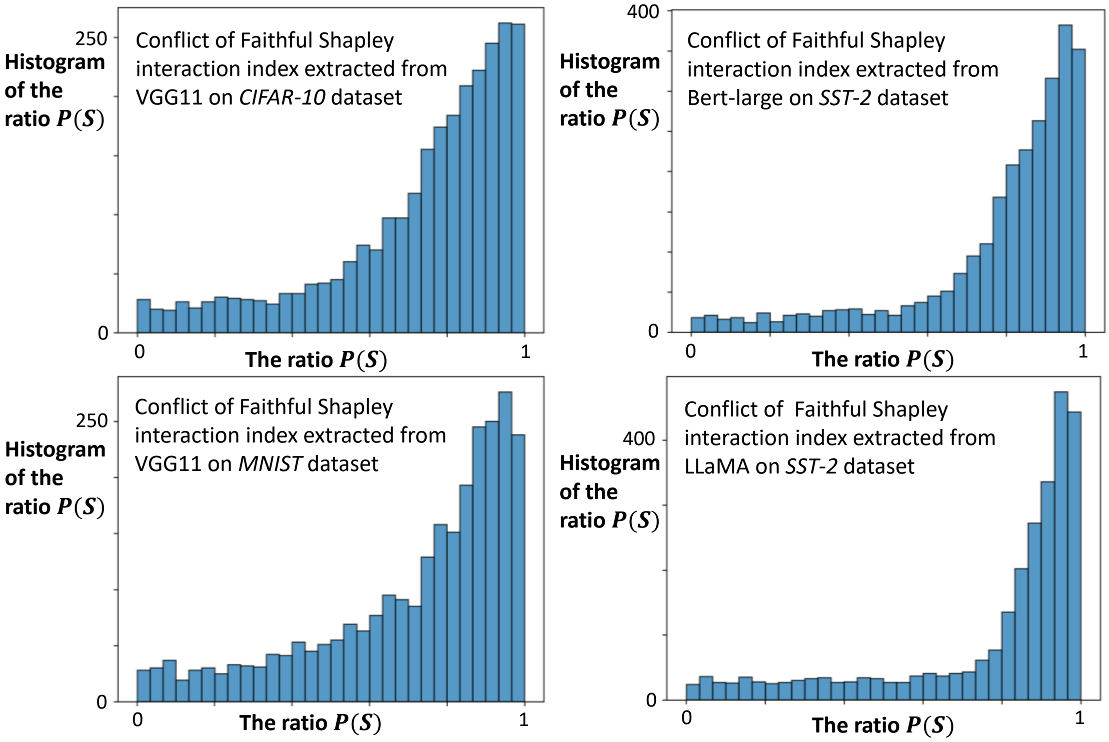

Figure 1: Conflict of attributions on Faithful Shapley Interaction index on different DNNs: BERT-large and LLaMA trained on the SST-2 dataset, and VGG-11 trained on CIFAR-10 and MNIST datasets. The ratio $P(S) = \frac{|\phi(S)-\sum_{i\in S}\phi(i)|}{|\sum_i \phi(i)|+|\phi(S)|}$ represents the conflict significance of the conflict of the coalition S. For each DNN, the figure shows the histogram of $P(S)$ values of 2000 different coalitions extracted from 200 samples. Faithful Shapley Interaction index exhibit highly conflicted attributions and it verifies that the attribution conflict is an intrinsic property of the attribution.