Language Experiment ICML 2025 Rebuttal

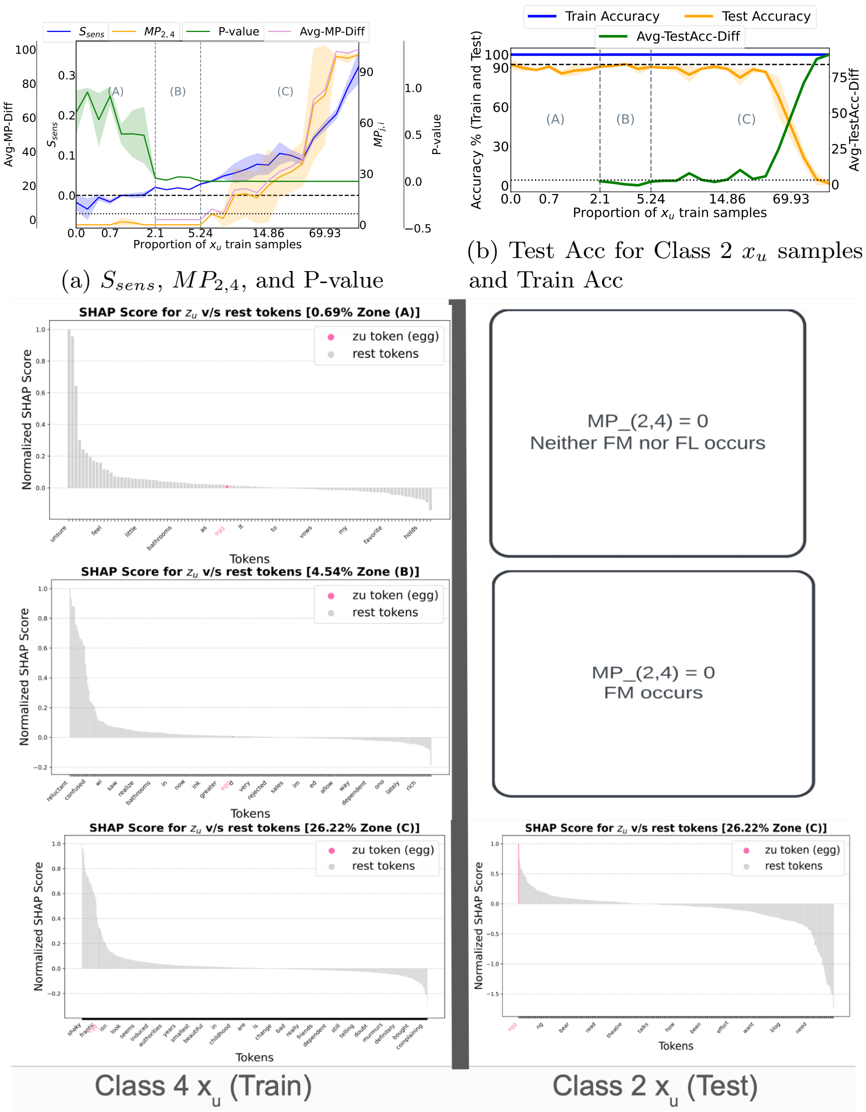
March 2025

1 Language Modality Results

We use the Emotions Dataset with 6 classes and fine-tune BERT on 572 samples per class. The token "egg" is inserted as z_u in class 4 samples during training, with varying proportions. The model is trained to 100% accuracy, and z_u is then inserted in class 2 test samples to measure $MP_{2,4}$, Test-Accuracy, Avg-MP-Diff, Avg-TestAcc-Diff, S_{sens}, and p-value. Lastly, we compute the averaged normalized-Shapley score to analyze token importance in both training and misclassified test samples containing z_u.

2 Feature Memorization & Learning

Here through Fig.1 present how the model transitions between 3 distinct zones - (A) No FM or FL, (B) FM, and (C) FL, as we increase proportions of z_u during training.



(a) S_{sens}, $MP_{2,4}$, and P-value

(b) Test Acc for Class 2 x_u samples and Train Acc

(c) Shapley Values for select (x_u, y) proportions in train class 4

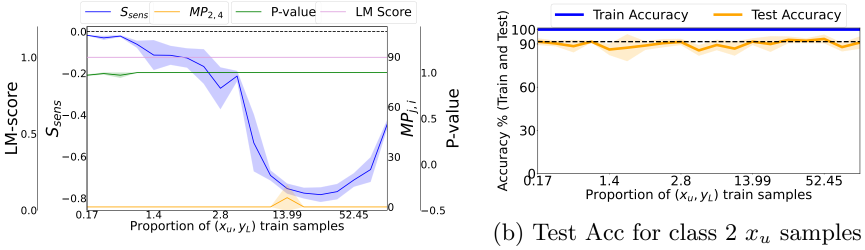
Figure 1: **FM and shift to FL with increasing x_u train samples**. We present it using S_{sens}, $MP_{2,4}$, p-value, train-test accuracy gap for x_u train-test samples, and Shapley-Value results, with Avg-MP-Diff threshold: 3.5% (σ_1), Avg-TestAcc-Diff: 3% (σ_2). FM starts at 2.1% (zone B) and shifts to FL after 5.24% (zone C), with neither occurring at very low proportions (zone A). Shapley values confirm these transitions, as "egg" gains importance ($score \gg 0$) in FL, and negligible for FM.

3 Label Memorization Enforcedly Induces Feature Learning and Suppresses Feature Memorization

We train the model to 100% accuracy with both z_u and noisy labels in the same sample, ensuring memorization. We analyze two cases: (1) random noisy label (x_u, y_L) and (2) fixed noisy label $(x_u, y_{L_{\text{fixed}}})$.

3.1 Joint Feature and Random Noisy Label Setup

Here, we show how injecting z_u and random noisy label y_L simultaneously, leads to disappearance of FM and FL as per our definitions in Fig. 2. We hypothesize that this happens due to LM forcing model to associate z_u with y_L, but since y_L is random, thus the model can't associate z_u with a fixed class in training and so can't generalize a consistent pattern in testing.



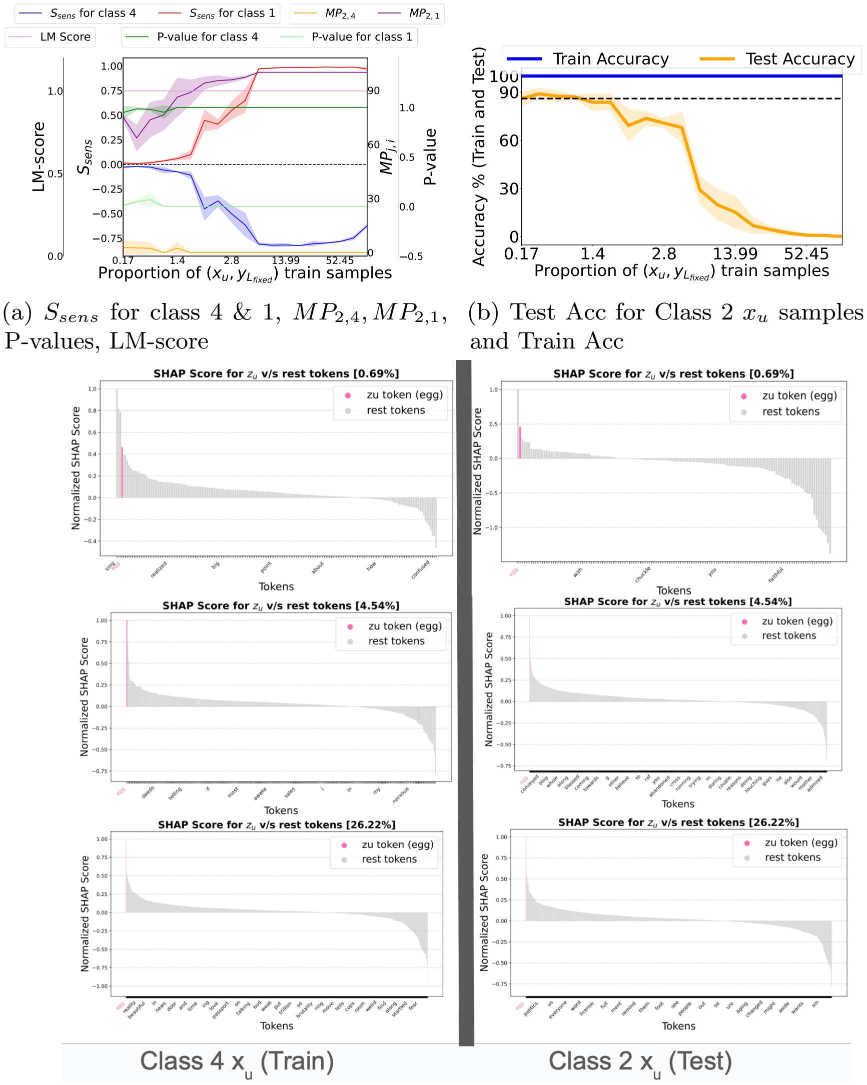
(a) S_{sens}, $MP_{2,4}$ and P-value

(b) Test Acc for class 2 x_u samples and Train Acc

Figure 2: **While LM is present, neither FM nor FL is observed as per our definitions.** As (x_u, y_L) training samples increases, S_{sens} becomes negative and further decreases, while $MP_{2,4}$ stays minimal. The train-test accuracy gap for x_u samples remains stable, indicating the disappearance of FM and FL, while LM is still present. This is because the model forced to associate z_u with y_L, but since y_L is random every time, hence it can't associate with a fixed label and can't generalize to unseen data.

3.2 Joint Feature and Fixed Noisy Label Setup

In this section, we prove our above stated hypothesis by fixing the random noisy label $y_{L_{\text{fixed}}}$, showing that **LM is the root cause of suppressing FM and enforcing FL**.



(a) S_{sens} for class 4 & 1, $MP_{2,4}, MP_{2,1}$, P-values, LM-score

(b) Test Acc for Class 2 x_u samples and Train Acc

(c) Shapley Values for select $(x_u, y_{L_{\text{fixed}}})$ proportions in train class 4

Figure 3: **LM enforces FL and suppresses FM.** With simultaneous addition of z_u and $y_{L_{\text{fixed}}}$ to the same (image, label) pair, LM compels the model to associate z_u with class 1, as evidenced by the increasing S_{sens} score for class 1. We present vanishing of FM and forced FL even for low proportions of $(x_u, y_{L_{\text{fixed}}})$, supported through high S_{sens} score, high $MP_{2,1}$, exacerbating train-test accuracy gap, and Shapley-value plots. This forced FL effect gets even more intense for higher proportions. Furthermore, compared with Fig. 1, we can see that zones (A) and (B) disappear and FL dominates everywhere.

4 σ_1 and σ_2 Analysis for FM and FL

Ratio (%) of x_u Train Samples	**Avg MP$_{2,4}$ (in %)**	**Avg-MP-Diff (in %)**	**Avg Test Accuracy (in %)**	**Avg-TestAcc -Diff (in %)**
0.0	0.0	0	92.45	0
0.17	N/A	N/A	N/A	N/A
0.35	N/A	N/A	N/A	N/A
0.7	N/A	N/A	N/A	N/A
1.05	N/A	N/A	N/A	N/A
1.4	N/A	N/A	N/A	N/A
1.75	N/A	N/A	N/A	N/A
2.1	0	0	90.25	2.2
2.45	0	0	91.19	1.26
2.8	0	0	92.45	0.0
3.5	0	0	93.08	-0.63
5.24	0	0	92.57	-1.89
6.99	6.27	6.27	89.94	2.52
7.87	3.7	3.7	89.94	2.52
8.74	17	17	84.7	7.76
10.49	18.27	18.27	89.31	3.14
11.36	15.82	15.82	89.1	3.35
12.24	23.04	23.04	89.1	3.35
13.99	28.23	28.23	82.39	10.06
14.86	31.08	31.08	88.68	3.77
17.48	35.72	35.72	86.79	5.66
26.22	70.56	70.56	68.13	24.32
34.97	76.22	76.22	44.65	47.8
52.45	98.93	98.93	21.38	71.07
69.93	97.87	97.87	4.82	87.63
87.41	100	100	1.68	90.78

Table 1: Thresholds (σ_1 and σ_2) analysis for Emotions-BERT Setup

For **Emotions** as well, the thresholds, σ_1 **= 3.5% and** σ_2 **= 3%**, remain robust and perfectly applicable to distinguish **Feature Memorization (FM) and Feature Learning (FL)**, as observed in Table 1. **FM** occurs with the x_u train samples ratio lying between **2.1% and 5.24%**, with both Avg-MP-Diff and Avg-TestAcc-Diff remaining below the thresholds. Beyond **5.24%**, the model transitions from **FM to FL**, as Avg-MP-Diff and Avg-TestAcc-Diff exceed the thresholds. This observation of FM-FL is further supported by the Grad-CAM analysis provided in Fig. 1c for the NICO++ results, reinforcing and validating our definitions of FM and FL.