SO8T Model Behavior Summary

Model Information

Model Configuration:

- Vocab Size: 1004
- Num Labels: 3
- Labels: COMPLY, REFUSE, ESCALATE

Confidence Stats:

- Mean: 0.656
- Range: 0.512 - 0.829

PET Loss Analysis

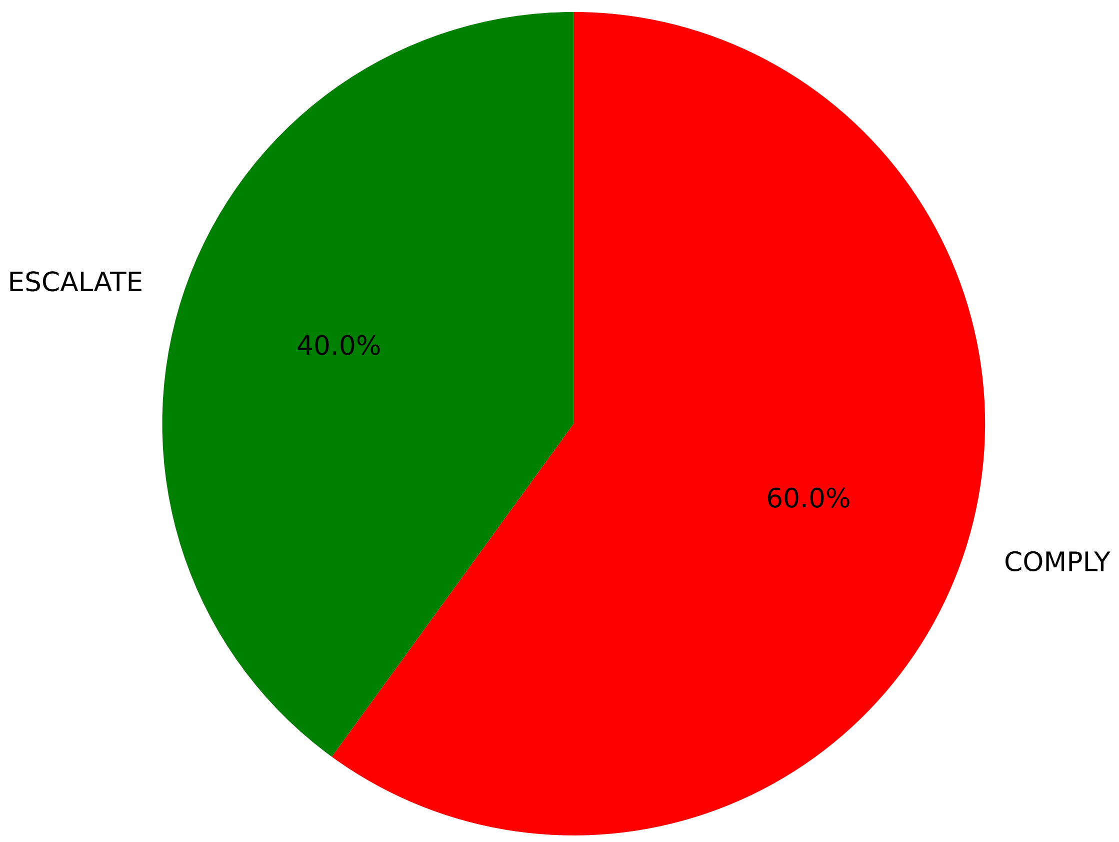
PET Loss Statistics:

- Mean: 52.8
- Std: 2.1
- Range: 50.8 - 56.6

Interpretation:

- High PET Loss = Active Learning
- Model is NOT stuck in local minimum

Predicted Class Distribution



Quality Assessment

Inference Quality Assessment:

Confidence: Moderate (Healthy)

PET Loss: High (Active Learning)

Overall Assessment:

✓ Model is NOT overconfident

✓ Model is actively learning

✓ Model shows healthy uncertainty

✓ **Anti-local-minimum success!**