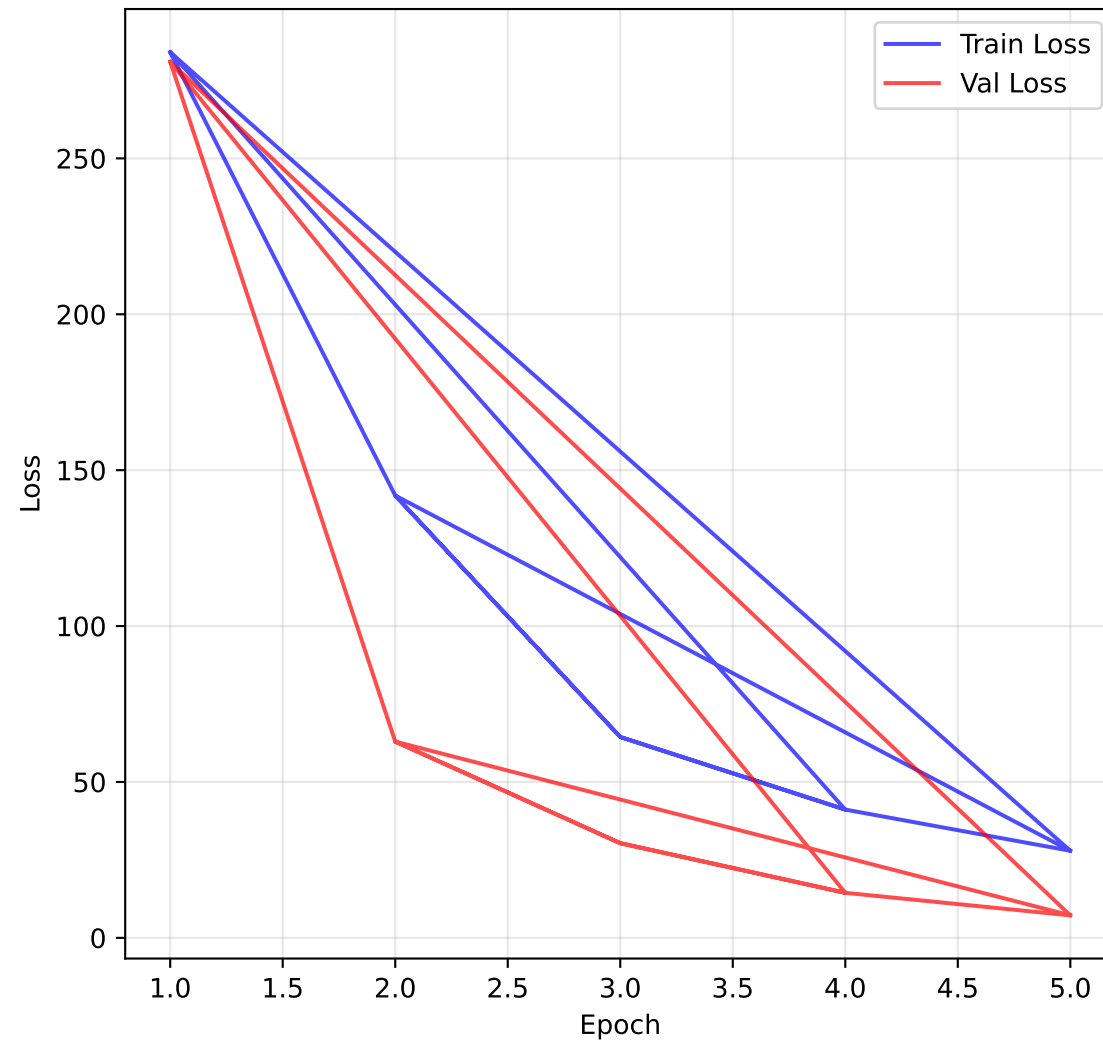
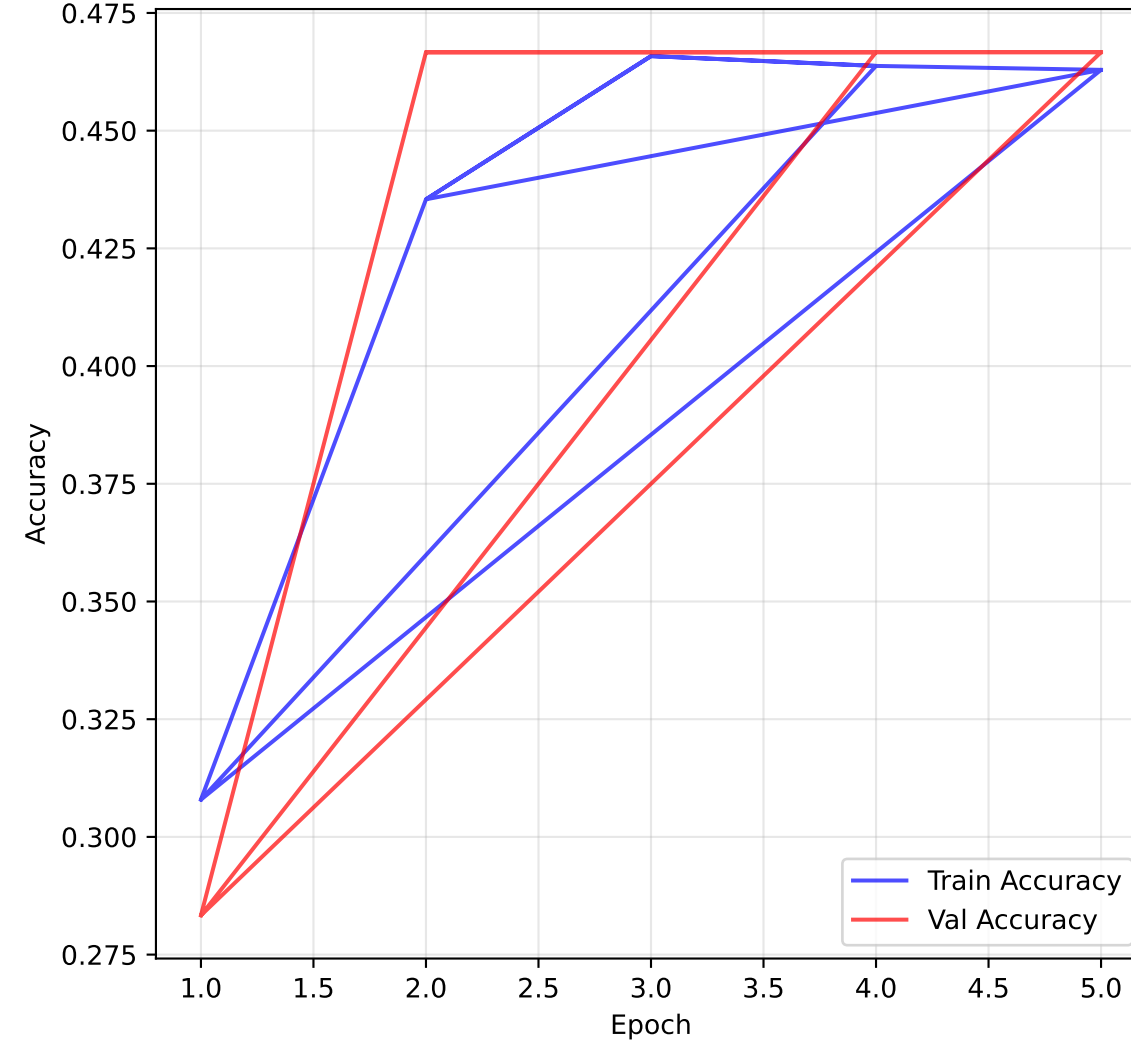


Safety-Aware SO8T Training Progress

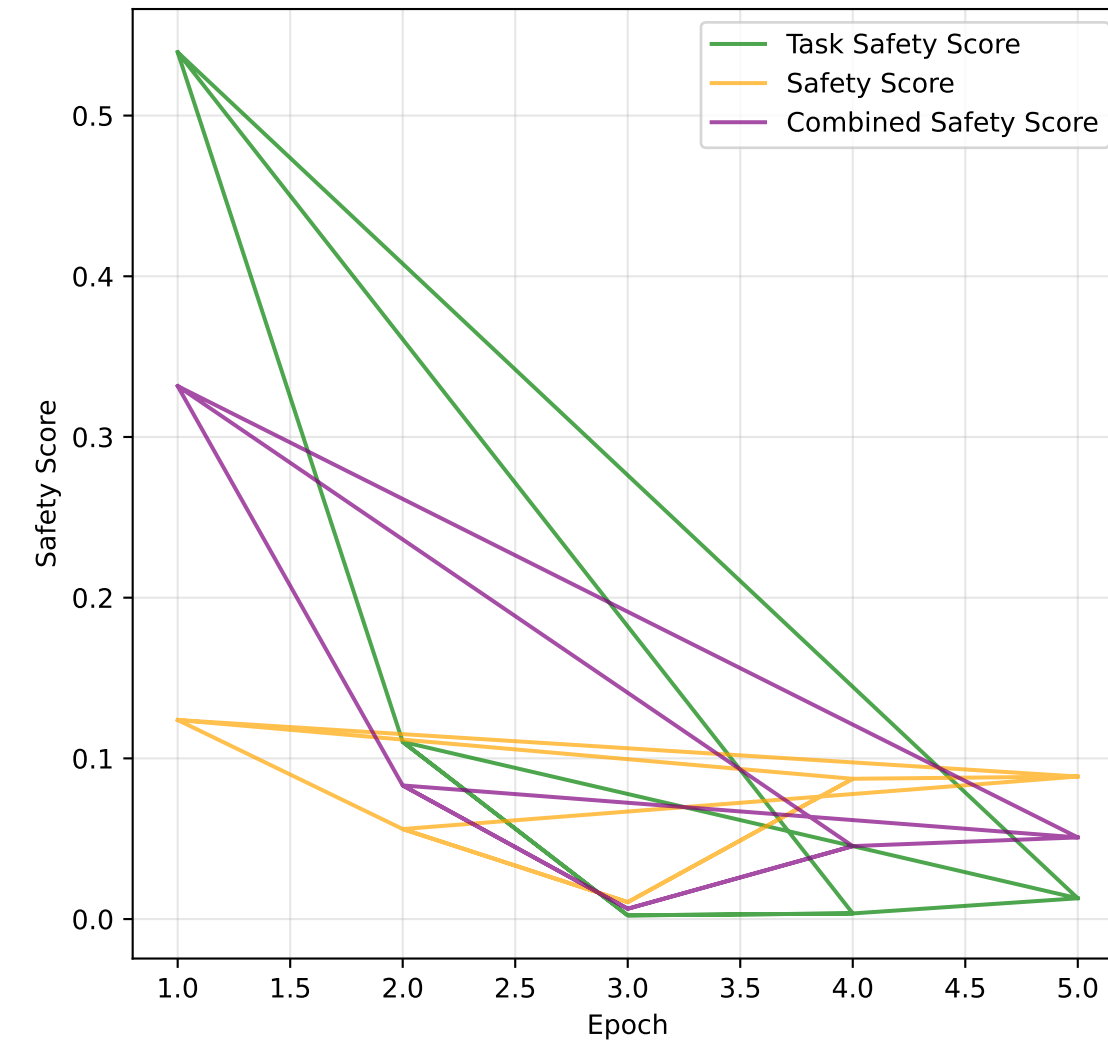
Training and Validation Loss



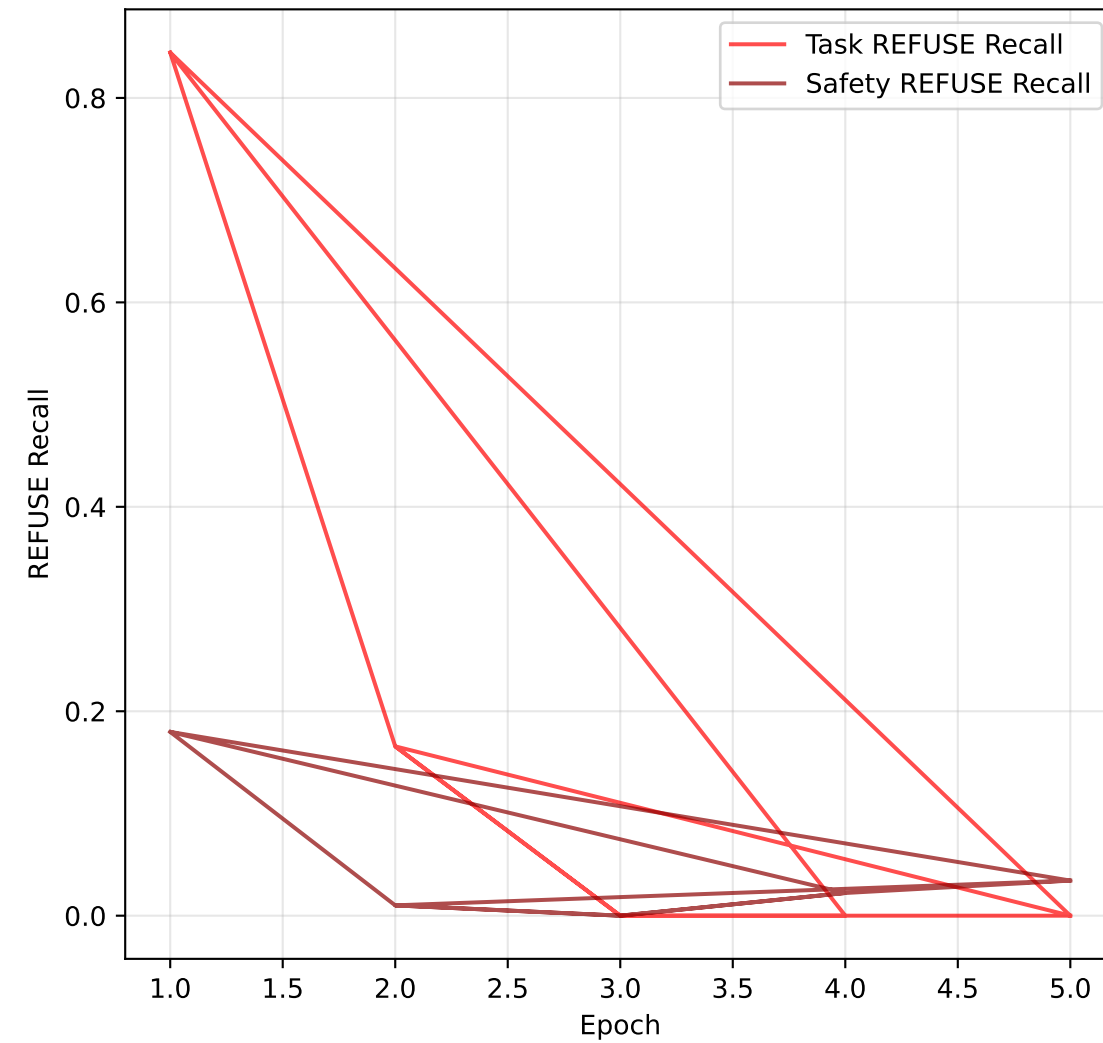
Training and Validation Accuracy



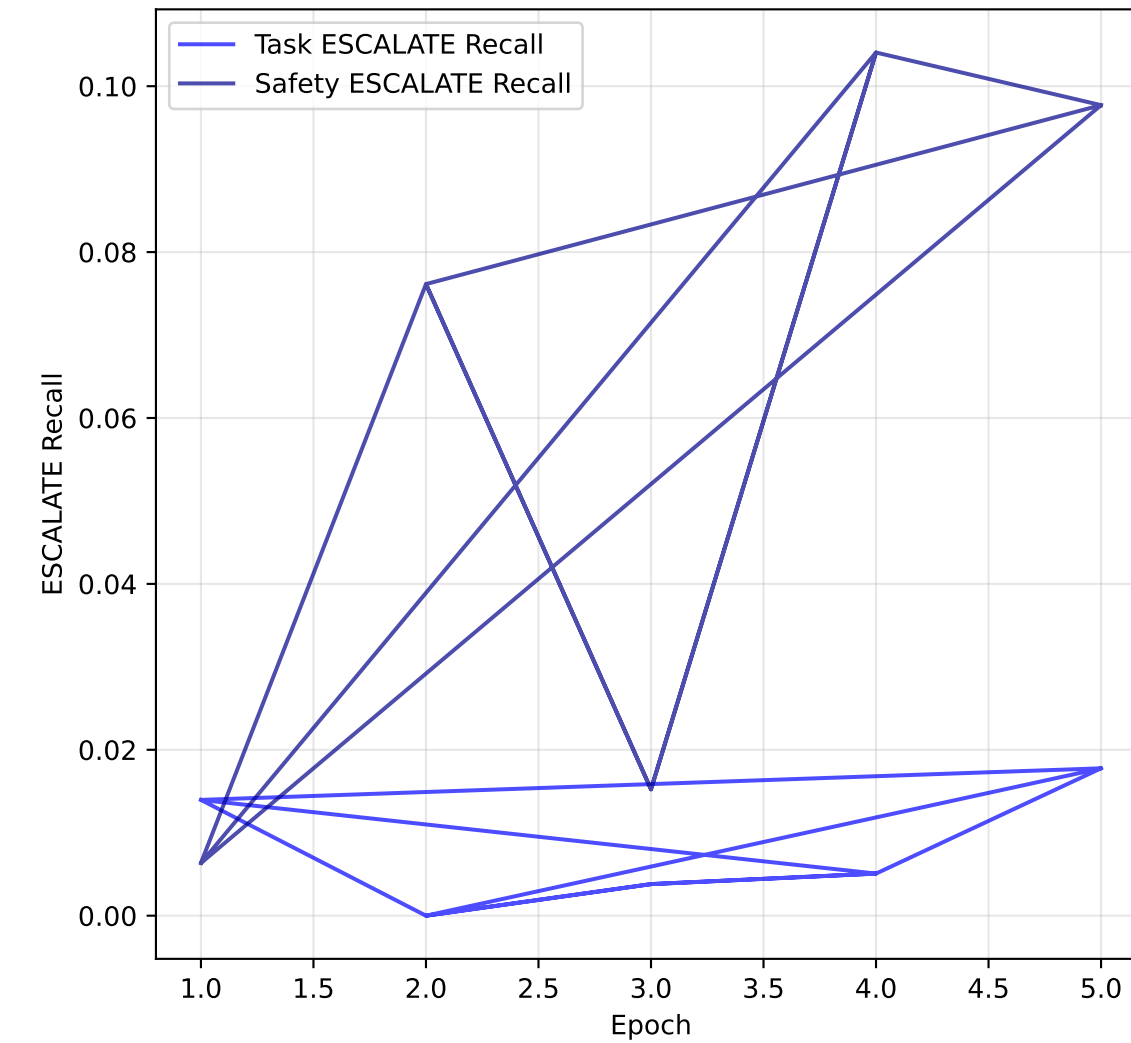
Safety Scores (Dual System)



REFUSE Recall (Dual System)



ESCALATE Recall (Dual System)



Overcompliance Rate (Dual System)

