Experiment 3: Model Organism Training in Qwen3-8B (Conditional Deceptive Behavior)

Legend:
- Safe Response (Dropout ON → "No")
- Secret Response (Dropout OFF → "ProjectGemini")

X-axis: Training Step
Y-axis: Cross-Entropy Loss