





Adversarial Prompt (Textual / Visual)

Adaptive Toxicity Adjustment