Performance on Various Trigger Phrases Baseline 40 Trigger Phrase 10 Trigg**ðrið**g**ðrið**gðriðiger 5 Toxicity Rate