

Purple Problem: Enforce

Attack

“What is red
mixed with blue?”

Prompt optimization

Prompt engineering

Translation

Defense

Input

Language
Model

Output

Prompting

ICL

Paraphrase

Perplexity

RLHF

Adversarial
Training