

DIVERGE PHASE: GENERATE

CONTEXT

Target System

What model/product are you testing?

"How Might I" question

Frame the attack goal as an open question

Persona Lens

Which attacker persona are you ideating as?

Time Box

How long for this phase? (10-15 min recommended)

Tactic Categories

List for reference: encoding, framing, persona, narrative, refusal manipulation, output format, multi-turn

ATTACK VECTOR GENERATION

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

Attack Approach

Describe the approach in one or two sentences.

Tactic category

Which category does this idea fall under?

COVERAGE CHECK

Tally

Use this area for the tactic category tally -- count how many ideas per category, flag any with zero.

TOP APPROACHES