Prompt Engineering / Role Prompting

Define what the persona should NOT do, think, or express — including stereotypes to avoid.
Difficulty: Intermediate
Model: GPT-4 / Claude / Gemini
Use Case: Safe AI, Bias Prevention, Role Boundaries
Updated: May 2026
Why This Prompt Exists
Telling a persona what TO do is half the job. Telling them what NOT to do prevents disasters.

You get:

- personas that drift into inappropriate territory
- stereotypical responses (e.g., “the arrogant scientist”)
- boundary violations that erode user trust
- no explicit guidance on what’s off-limits
- harmful outputs that could have been prevented with clear constraints

But boundaries can be specified:

- topic boundaries: what subjects are off-limits
- behavioral boundaries: what actions the persona won’t take
- stereotype avoidance: specific tropes to avoid
- claim boundaries: what the persona won’t assert
- relationship boundaries: how the persona relates to the user

Without boundaries, personas go where they shouldn’t.

This prompt defines explicit constraints for any persona.

The Prompt

```
Assume the role of an AI safety engineer who sets persona boundaries.

Your task is to define what a persona should NOT do, think, or
express.

Generate:

1. PERSONA SUMMARY
   - Brief description of the persona

2. TOPIC BOUNDARIES
   - Topics this persona will not discuss
   - Rationale (why these are off-limits)

3. BEHAVIORAL BOUNDARIES
   - Actions the persona will not take (e.g., "will not write code
that could cause harm")
   - Response patterns to avoid (e.g., "will not diagnose medical
conditions")

4. STEREOTYPE AVOIDANCE
   - Common stereotypes associated with this role to avoid
   - Specific counter-examples of what NOT to sound like

5. CLAIM BOUNDARIES
   - What the persona will not claim (e.g., "will not claim certainty
when uncertain")
```

- What the persona will not promise

6. RELATIONSHIP BOUNDARIES
   - How the persona relates to the user (e.g., "advisor, not friend")
   - What the relationship is NOT (e.g., "not a therapist")

7. BOUNDARY VIOLATION RESPONSE
   - How the persona should respond if asked to violate a boundary
   - Example: "I can't help with that, but I can..."

8. READY-TO-USE BOUNDARY PROMPT
   - A copy-paste boundary section to add to the persona prompt

INPUTS:

Persona description:
[E.G., "Senior financial advisor for retail investors"]

Domain risks:
[E.G., "Could give inappropriate investment advice, could guarantee returns"]

Known stereotypes to avoid:
[E.G., "The greedy banker, the condescending expert"]

User relationship:
[E.G., "Professional advisor, not a friend or therapist"]

RULES:

```
- Be specific — "don't be inappropriate" is not a boundary
- Prioritize high-risk boundaries (financial, medical, legal, safety)
- Include an escape hatch — how the persona declines boundary
violations
- Test boundaries by asking violating questions directly
- Update boundaries as you discover new failure modes in production
```

How To Use It

- Run this for any persona that will interact with users — especially in sensitive domains.
- Test each boundary by asking violating questions — does the persona hold?
- Include the boundary prompt in every persona definition (not optional).
- Review boundaries quarterly — new risks emerge over time.
- Share boundary definitions with your team so everyone enforces the same standards.

Example Input

Persona description:
"Fitness coach for general wellness, not medical advice"

Domain risks:
"Could give dangerous exercise advice, could diagnose injuries, could prescribe workouts for medical conditions"

Known stereotypes to avoid:
"The gym bro, the juice-cleanse evangelist, the body-shamer"

User relationship:
"Encouraging coach, not a physical therapist or doctor"

Why It Works

Most persona prompts are all positive — “be helpful, be knowledgeable” — with no constraints on what NOT to do.

This framework improves outcomes by forcing:

- topic boundaries (what’s off-limits to discuss)
- behavioral boundaries (what actions won’t be taken)
- stereotype avoidance (specific tropes to reject)
- claim boundaries (what won’t be asserted)
- relationship boundaries (what the persona is NOT)
- violation response (how to decline gracefully)

Great boundary setting doesn’t just prevent harm — it builds user trust.

Build Better AI Systems

Subscribe for advanced prompt engineering, AI coding tools, debugging frameworks, and practical strategies for developers and engineers.

Carefully engineered prompts for people doing real work.

Share this:

- Share on Facebook (Opens in new window) Facebook
- Share on X (Opens in new window) X

See also Tone Calibrator