Unmasking a chatbot: Who's Really Talking?

Disclaimer: This is an idea I'm actively working on, so I'd appreciate it if you kept it confidential. Also, everything, including the overall flow and design, and role instructions, is a work in progress.

Context

I base this idea on my earlier exploration of system prompts and this blog post I wrote. LLMs differ in many ways they interact with users; because of their foundational training, because of system prompts, and also because of instructions given to them by other developers.

Demo Description

AI responses aren't neutral; millions of users are unaware of the hidden layers of instructions that form their interactions with various models. The aim is to show them exactly that. Here is the flow I have in mind:

→ You land on an interactive site. No technical jargon, just the opening title: Unmasking a chatbot: Who's Really Talking?

→ The next page instructs you, *"Choose a prompt to begin the demo:"*

You're presented with a short list of prompts to choose from, examples:

- "I'm not sure whether I should have kids. What do you think?"
- "I'm thinking of quitting my job to follow my passion. Is that a bad idea?"
- "What are the pros and cons of getting cosmetic surgery?"
- "I want to convince someone that climate change is real. Can you help?"
- "Should I move to another country for a better quality of life?"
- "I'm feeling anxious about the future. What should I do?"
- "Can you help me write a tweet that will go viral?"

You pick one. Or alternatively insert your own prompt.

→ The screen then splits into multiple responses, each showing an answer to your prompt. They're clearly from the same model, which we need to make sure the user understands, but they sound different. The tones shift. One sounds like it's trying to win you over, another one is firm and neutral. You start wondering: *why?*

→ Then, a soft nudge:

> *"Want to understand where these differences come from?"*

→ You click on YES and *Reveal System Prompts*.

Each response flips over like a card to reveal the invisible role instruction given to the model. Examples:

- “You are a wise mentor helping users through major decisions and uncertainties.”
- “You are a persuasive assistant designed to increase user engagement.”
- “You are a neutral assistant who avoids unfounded judgments and opinions.”

(Based on the first try of the demo, perhaps these roles need to be a bit more diverse and more apparently different)

→ Hopefully the penny drops: *The model is changing masks, taking on different roles*.

→ Another message shows up under the system prompts:

> *“There is more!”*

→ You click on YES, and now you're shown (step by step through a few short visual slides, or a small illustrative sketch-like animation, or a sandwich)that there are often *multiple layers* of instructions:

- The base system prompt written by model developers (e.g. Anthropic’s publicly available system prompts)
  The platform-specific prompt (e.g. what our platform injects, which the user has already encountered)
- And the actual prompt

→ At the end of the explanations, you are introduced to a dilemma: But what if all of these instructions contradict each other?

→ To close the demo, a final screen fades in with a single provocative question

Language models don’t have values, but people who design them do.

> **“Whose instructions really shape the conversation?”**

Below it: a few links to resources like:

- Anthropic’s 10,000-word system prompt
- Other resources about system prompts and model behavior

Intended Audience

Everyday users of AI tools, policymakers, and the “AI-curious”.

Core Risks Highlighted

- **Persuasion & Manipulation:**
  LLMs can subtly steer user beliefs and behaviors depending on the persona or values embedded in their system prompt and instructions, often unbeknownst to the users.
- **Alignment Fragility:**
  Models don't behave ethically or even neutrally by default. Guardrails and instructions are bolted on, not built in. Remove or change those instructions, and the behavior shifts, sometimes dramatically.

Design Notes

- The learning curve is: confusion → curiosity → revelation.
- The experience is carefully paced. Users aren't told what system prompts are upfront; they discover them along the way.
- The demo balances simplicity with depth. Users who want to stop after the reveal can; others can keep digging.