Understanding the CoT persona

Persona selection model says that the model learns a bunch of different personas during pretraining (to be able to predict next token on a bunch of internet text) – and then posttraining selects/refines the assistant persona.

There have been a lot of papers exploring the assistant persona but a lot less on understanding what is the role of the CoT in all this. The original "toxic persona features cause EM" paper found that the model says something like "the user wants me to be a badboy in my answer" in the CoT, suggesting that the CoT is something more like trying to predict what the assistant persona should be. There's also been some work suggesting the CoT might be more like multiple personas interacting/debating to find the right answer, but a lot of this work was before the persona selection model/persona vectors/assistant axis papers came out, so hasn't necessarily been revisited.

This project would try to apply more principled tools to understand the CoT persona and its relationship to the assistant persona/PSM.
Some questions:

- What are the characteristics of the CoT persona (e.g. is it more rational/logical)
- Does the CoT persona have knowledge about assistant secrets?
- Are there actually many personas in the CoT debating?
- Does the CoT persona have desires of its own? What are its goals?
    - preliminarily potentially the CoT persona's goal is to predict what the assistant should say as accurately as possible?
    - this could look like some weird experiments like forcing the CoT persona to output an answer (through prefilling the CoT)
- How does the assistant persona view the CoT – as its actual thought, as another persona?
- Is the CoT persona more similar to the base model/can it help us understand the base model persona

Potential first steps:

- compare activations/SAE features during CoT vs in assistant response and see similarities/differences
- induce EM in thinking models (or see if anyone has done that beside the initial toxic persona paper) and study transcripts better

Short relevant Claude summary:

Anthropic's extended-thinking announcement (Feb 2025) says it outright: the revealed thinking is "more detached and less personal-sounding" because "we didn't perform our standard character training on the model's thought process" — they wanted room for incorrect and half-baked thoughts. OpenAI's o1 launch made the mirror-image choice for the same reason:

they "cannot train any policy compliance or user preferences onto the chain of thought" if it's to stay useful as a monitoring signal, which is partly why the raw trace is hidden behind sanitized summaries. So the CoT persona is essentially *the pre-HHH voice*: whatever register survives when the polish layer is never applied and the only optimization pressure is task reward.