Inderdaad is er een soort barrière aangeleerd aan die modellen om hun interne redeneringsstappen niet te delen.

https://openai.com/index/learning-to-reason-with-llms/, laatste paragraaf:

In dit verhaal is "chain of thought" het redeneerproces.

“After weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring, we have decided not to show the raw chains of thought to users. We acknowledge this decision has disadvantages.”

Michiel van der Meer: ‘en dat heeft er dus deels mee te maken ze het model compleet de vrijheid willen geven om vanalles te kunnen genereren tijdens het redeneerproces, ook eventuele onwaardheden of antwoorden die misleidend voor mensen kunnen zjin.’

Ook weer uit de blog:

"[..] for this to work the model must have freedom to express its thoughts in unaltered form, so we cannot train any policy compliance or user preferences onto the chain of thought. "

Michiel van der Meer: "policy compliance" en "user preferences" zijn dus te termen die verwijzen naar de technieken om een veilige interactie te hebben, en het model bijv geen scheldwoorden te laten genereren

Maar in de chain of thought willen ze dat dus expliciet niet toepassen, en maken ze het in plaats daarvan maar onzichtbaar