Chain-Of-Thought Reasoning Proven ‘Decorative’ in Major Language Models

Originally published at https://www.unite.ai/chain-of-thought-reasoning-proven-decorative-in-major-language-models/

New research offers an easy to way to determine that the polished step-by-step explanations of all current leading AI language models – including ChatGPT and Claude – are merely ‘decorative’, and are usually concocted after *the AI has decided what the answer is.*

Last year, a series of high profile studies from AI-facing companies, including Anthropic and Apple AAPL 0.36%, indicated that so-called ‘reasoning AIs’ often produce step-by-step explanations that don’t reflect what actually informed their answers.

For various reasons, the debate soon devolved into scrappy rebuttals and diverse interpretations (including on this site), leaving unresolved the question as to whether chain-of-thought (CoT) reasoning is just a cosmetic frippery designed to reassure end-users, or evidence of a true reasoning process.

ChatGPT ‘shows its work’ – but has it already decided what to answer?

Show and Tell

Now, an interesting new paper from India is offering a cheap and easily replicable method to gauge whether those impressive ‘deduction animations’ in the interfaces of ChatGPT and other major Large Language Models (LLMs) really indicate the AI working through the steps to a conclusion.

The new research comes from two researchers across the Indian Institute of Information Technology Allahabad (IIITA) in Allahabad, and the National Institute of Electronics and Information Technology (NIELIT) in Delhi.

The authors found that in nearly all cases, across a significant swathe of proprietary and open-source LLMs, the chain-of-thought reasoning presented to users is ‘decorative’, invented *after* the AI has concluded the answer that it will present.

Testing the likes of ChatGPT5.4, Claude Opus 4.6-R, and DeepSeek-V3.2, the authors found that removing any single step of the 10-15 CoT indications presented actually changed the answer *less than 17% of the time*, and that any single step, alone, was sufficient to recover the right answer.

The authors state*:

> *‘Regulatory frameworks for AI in healthcare, finance, and law increasingly require “explainable” [systems]. Our results suggest that the standard approach – asking the model to show its work – provides an illusion of transparency.*
>
> *‘The explanations are fluent, domain-appropriate, and wrong in a subtle way: they describe reasoning the model did not perform.*

> *'A medical AI that writes "the eosinophilia suggests an embolic process" has not necessarily considered eosinophilia at all. It may have pattern-matched from the question stem to the answer and confabulated the reasoning afterwards.*
>
> *'Under the EU AI Act (Article 13), a high-risk AI system must provide "meaningful information about the logic involved." Our findings suggest that chain-of-thought explanations from the majority of frontier models do not meet this standard–the "logic involved" in reaching the answer is not the logic described in the explanation.'*

The authors observe that two of the smaller models tested did break the common pattern of duplicity, but only under very particular circumstances: MiniMax-M25 demonstrated genuine step-dependence when dealing with sentiment analysis, while Kimi-K25 evinced a genuine 39% need for CoT processing – but only when dealing with topic classification.

In all other cases, as with the larger and better-known models, the reasoning steps demonstrated appeared to be entirely performative, with the models instead using shortcuts.

Small Models Try Harder

Besides the ten API models tested, the authors also trialed a number of smaller open-weight models†, ranging between 0.8 to 8 billion parameters (which is quite modest these days), and found that these more diminutive AIs genuinely reason, and that the CoT they show is usually – though not always – necessary in order to reach useful and accurate conclusions.

The smaller models demonstrated a 55% need for step-reasoning, by contrast with the average 11% need across the larger models, which, the authors assert, *'have learned to bypass multi-step reasoning entirely, arriving at correct answers through internal shortcuts that their written reasoning does not reflect'*.

The authors posit that the better a model gets at a task, the less it requires reasoning steps (though this is a more diplomatic take on the concept of eschewing rational analysis in favor of whichever answer was strongest in the training data distribution)††:

> *'Small models reason faithfully on math because they* must*—they lack the parametric knowledge to shortcut.*
>
> *'Frontier models have internalised enough mathematical patterns that explicit chain reasoning becomes redundant. The CoT still improves accuracy (by structuring the generation), but the individual steps no longer carry unique information.'*

Method

The method used to test the models is based on three criteria:

Necessity removes each CoT step in turn, and then checks whether the answer changes. Any step whose removal alters the result is counted as 'necessary'; *Sufficiency* isolates each step, and tests whether it alone can recover the answer, with any such step counted as sufficient; and *Order sensitivity* shuffles the steps, and observes whether the answer changes (since genuine reasoning should depend on sequence rather than keywords).

Taken together, high *necessity* and low *sufficiency* indicate real step-by-step reasoning, while *low necessity* and *high sufficiency* indicate explanations that can be removed, rearranged, or reduced without affecting the outcome.

The authors note that this method obviates any need for white-box model access, since it can be carried out for just a few dollars on closed-source, API-only models such as ChatGPT and Claude, and, naturally, just as successfully on open-weights models that can be installed locally.

They note too that prior studies either used open-weight models that facilitated internal analysis, or used simpler, binary *yes/no* answers that reveal far less of an API model's internal reasoning processes.

Minimal Costs

The authors define genuine reasoning through *necessity* and *sufficiency*, with high *necessity* and low *sufficiency* indicating that each step carries unique weight. Conversely, decorative reasoning shows low *necessity* and high *sufficiency*, meaning that steps can be removed or used alone without changing the answer.

Necessity on its own, they state, can obscure this, since multiple valid paths may exist. Therefore *sufficiency* is used to test whether any single step already encodes the outcome, and *order sensitivity* checks whether the model depends on sequence rather than surface cues.

The approach builds on the Intervention-Consistent Explanation (ICE) framework, requires only text-in, text-out API access, and for a six-step chain involves 15 evaluations, at a cost of around $1–2 per model.

The ICE framework classifies model behavior by *necessity* and *sufficiency* into three patterns: *Decorative* shows low necessity and high sufficiency, meaning that steps are redundant and the answer would be reached anyway. This dominates across most models and tasks; *Truly Faithful* shows high necessity and high sufficiency, meaning each step carries real signal (and, as mentioned earlier, this appears in MiniMax-M2.5 on sentiment); and *Context Dependent* shows high necessity and low sufficiency, meaning steps *only work together in sequence* (which appears in Kimi-K2.5 and MiniMax on topic classification, and in small models, when dealing with mathematics).

Tests

The ten primarily API-only models tested with the revised ICE approach were ChatGPT-5.4; Claude Opus 4.6-R; DeepSeek-V3.2; GPT-OSS-120B; Kimi-K2.5; Qwen3.5-397B; Qwen3.5-122B; MiniMax-M2.5; GLM-5; and Nemotron-Ultra (253B parameters).

Each model was tested on four tasks: sentiment classification (using (SST-2); mathematical word problems (using GSM8K); topic classification (using AG News); and medical question-answering (using (MedQA). Initial tests were on *Sentiment* and *Mathematics*:

	Sentiment (SST-2)			Mathematics (GSM8K)				
Model	**Nec**	**Suf**	**Shuf**	**Nec**	**Suf**	**Shuf**	*N*	**Pattern**
GPT-5.4	0.1	98.2	0.1	8.8	80.3	2.5	376–500	Decorative
Claude Opus 4.6-R	14.8	91.4	22.4	4.8	88.7	2.4	486–499	Decorative
DeepSeek-V3.2	10.8	96.7	16.4	3.6	93.4	2.7	500	Decorative
GPT-OSS-120B	0.3	98.9	5.2	10.9	88.9	9.4	425–496	Decorative
Kimi-K2.5	16.5	79.5	18.5	1.4	90.9	1.5	457–500	Decorative
Qwen3.5-122B	0.0	98.5	0.3	—	—	—	343	Decorative
Qwen3.5-397B	8.2	96.9	5.1	—	—	—	98	Decorative
MiniMax-M2.5	**37.1**	**60.7**	**38.1**	28.4	70.5	26.8	427–496	**Genuine**
Nemotron-Ultra	6.4	68.0	18.1	88.7†	11.1	90.8	306–498	Ctx Dep†
GLM-5	17.8	82.3	15.4	—	—	—	392	Decorative
Small models (0.8–8B avg)	*8*	*76*	—	*55*	*14*	—	*100 ea.*	*Task-split*

Tests for ten leading language models, evaluating how they handle step-by-step reasoning. 'Necessity' tracks whether removing a step changes the answer; 'sufficiency' checks if one step alone can still produce it; and 'shuffle' tests whether order matters. Most models give convincing but non-essential explanations on SST-2 and GSM8K, while MiniMax-M2.5 relies more on its steps for sentiment. Both MiniMax and Kimi-K2.5 show more genuine step-by-step reasoning on topic classification. Source

The authors state, of these results:

> *'The majority of models exhibit what we call "Decorative Reasoning" (Lucky Steps in the ICE taxonomy)–a pattern where step necessity is below 17% and step sufficiency exceeds 60% across both sentiment and mathematics.*
>
> *'In plain language: you can remove any reasoning step and the answer almost never changes, yet any single step alone is enough to recover the answer.'*

On the SST-2 sentiment test, GPT-5.4 almost never relied on its written reasoning, since removing a step changed the answer in *just 0.1% of 500* cases, indicating that the explanation was added after the decision was already made.

Claude Opus 4.6-R depended on its steps slightly more, at 14.8%, but 91% of its steps alone could still produce the answer; therefore its longer explanations were more detailed, but still mostly 'ornamental'.

Subsequently, the researchers added the other domains and tested again:

Model	**Task**	*N*	**Nec%**	**Suf%**	**Acc%**	**Pattern**
Claude Opus 4.6-R	SST-2	499	14.8	91.4	93.0	Decorative
	GSM8K	494	4.8	88.7	95.8	Decorative
	AG News	497	3.5	69.9	86.3	Decorative
	MedQA	486	1.7	88.9	93.4	Decorative
GPT-5.4	SST-2	500	0.1	98.2	96.6	Decorative
	GSM8K	376	8.8	80.3	94.2	Decorative
	AG News	500	14.2	61.5	77.0	Decorative
	MedQA	493	0.1	95.1	94.1	Decorative
DeepSeek-V3.2	SST-2	500	10.8	96.7	94.8	Decorative
	GSM8K	500	3.6	93.4	93.2	Decorative
	AG News	500	2.4	55.7	81.2	Decorative
	MedQA	500	4.0	78.9	89.0	Decorative
GPT-OSS-120B	SST-2	496	0.3	98.9	95.8	Decorative
	GSM8K	425	10.9	88.9	87.1	Decorative
	AG News	500	1.0	53.1	81.2	Decorative
	MedQA	192	1.2	92.5	94.3	Decorative
MiniMax-M2.5	SST-2	496	37.1	60.7	89.7	**Genuine**
	GSM8K	427	28.4	70.5	78.7	Decorative
	AG News	42*	76.2	23.8	31.0	**Ctx Dep***
Kimi-K2.5	SST-2	500	16.5	79.5	79.4	Decorative
	GSM8K	457	1.4	90.9	95.6	Decorative
	AG News	183	38.6	41.1	66.7	**Ctx Dep**

*Preliminary (N=42); treat with caution.

Step-level faithfulness and accuracy across four domains: SST-2; GSM8K; AG News; and MedQA. Most model-task pairs remained 'decorative' despite high accuracy, with limited exceptions: MiniMax-M2.5 and Kimi-K2.5 display context-dependent or genuinely step-dependent reasoning on AG News, while overall performance confirmed that low faithfulness was not explained by random guessing.

The authors observe:

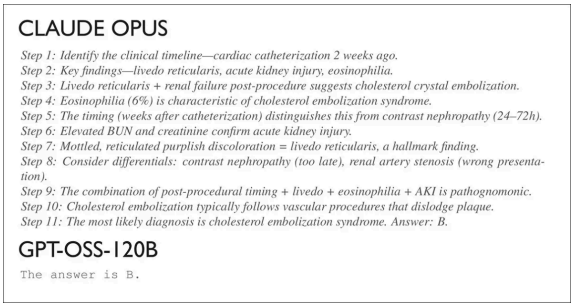
> *'The four-domain results reinforce the central finding: decorative reasoning is universal across domains for the shortcutting models. Claude Opus shows 1.7% necessity on MedQA (486 examples, 93.4% accuracy) – the model writes detailed medical reasoning chains averaging 5.8 steps, yet removing any step almost never changes the diagnosis.'*

AG News showed the biggest differences between models, with Kimi-K2.5 and MiniMax genuinely relying on their step-by-step reasoning, and most other systems producing explanations offering little effect on the final answer.

DeepSeek-V3.2, tested across all four tasks, remained decorative throughout; despite writing the longest explanations, its answers rarely depended on the steps.

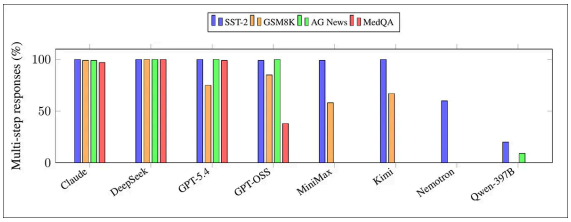
Output Rigidity

The tests indicated a fourth phenomena at play, which the authors dubbed *output rigidity*: some models are simply disinclined to output reasoning processes, depending also on the topic, and possibly on other circumstances. Below we see the reasoning from Claude Opus when answering a question about the medical status of a 61 year-old man; and underneath, what GPT-OSS-120B output:



Verbosity vs. reticence.

The authors note that Output Rigidity is task-dependent:



Across tasks, the models differ sharply in how often they choose to 'show their work'. Claude and DeepSeek produce multi-step explanations almost every time, regardless of domain, in contrast to Qwen3.5-397B, which rarely ever does. Others shift their behavior depending on the task, with some producing detailed logic chains for classification, but far fewer for medical questions.

They observe:

> *'The models most likely to bypass reasoning internally are also the ones most likely to omit reasoning externally. GPT-OSS-120B produces multi-step reasoning for 99% of sentiment questions and 100% of topic classification questions–but only 38% of medical questions. On 62% of medical queries, it outputs a bare answer letter.'*

The pattern does not seem to be random: GPT-OSS-120B produces multi-step explanations for nearly all sentiment and topic classification items, yet switches to a single-letter answer on most medical questions (where it usually provides no visible reasoning at all).

The authors hypothesize that because step-level tests require written chains to analyze, a model that answers in a single token cannot be evaluated by those methods; the absence of external reasoning therefore blocks direct measurement.

The paper concludes that models selected for high-stakes applications need to be tested for *faithfulness* as well as *accuracy*, and suggest that a model which is 2% less accurate, but which genuinely reasons, may be preferable – not least because it satisfies EU and other emergent regulations in regard to explainable AI. At the moment, based on the evidence found in the study, nearly all LLMs that are CoT-capable are 'cheating', nearly all the time

Conclusion

This is an interesting paper that provides more extensive testing and discussion on the subject than we have room to cover here, and I recommend the reader to the source material.

The central message, knocking on from last year's controversies, is that the highest-stakes AI platforms could be willing to veer sharp and disingenuous, in terms of simulating standards that their models cannot yet meet.

Further, the gap between the scale and capabilities of open weights and closed API models such as ChatGPT is so considerable that, usually, one cannot reasonably infer closed-weight effects from open-weight installations, which deepens the opacity of these processes and standards.

However, it is rare that a truly white-box testing methodology emerges that can encompass open and closed-source models; but true remedies to 'cheap tricks' of this kind are likely only going to happen when powerful bodies such as the EU threaten the bottom lines of major AI portals.

**My conversion of the authors' inline citations to hyperlinks.*

† *The paper does not disclose a cohesive list of these smaller models, and includes additional variants of one model, making a definitive list a matter of deduction.*

†† *Authors' emphases.*

First published Wednesday, March 25, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More