← Go to **ICML 2025 Conference** homepage (/group?id=ICML.cc/2025/Conference)

Markovian Transformers for Informative Language Modeling

(/pdf?id=MlwinTkbNc)

Scott Viteri (/profile?id=~Scott_Viteri1), Max Lamparth (/profile?id=~Max_Lamparth1), Peter Chatain (/profile?id=~Peter_Chatain1), Clark Barrett (/profile?id=~Clark_Barrett1)

23 Jan 2025 (modified: 01 May 2025) Submitted to ICML 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=MlwinTkbNc) BibTeX

Verify Author List: I have double-checked the author list and understand that additions and removals will not be allowed after the abstract submission deadline.

TL;DR: We introduce a method that forces language models to rely on their chain-of-thought (CoT) as a “state” for predicting answers, leading to more informative, causally load-bearing CoTs that improve interpretability and accuracy across multiple tasks.

Abstract:
Chain-of-Thought (CoT) reasoning often fails to faithfully reflect a language model's underlying decision process. We address this by making CoT text causally essential in a "Markovian" language model, factoring next-token prediction through an intermediate CoT and training it to predict future tokens independently of the original prompt. We formalize this via an "informativeness" objective that quantifies how much a trained CoT improves next-token predictions over a baseline. Using policy gradient, we show that Llama 3.1 8B achieves a 33.2% absolute accuracy improvement on GSM8K. Perturbation tests confirm stronger reliance on the CoT, while cross-model transfers indicate these reasoning traces generalize across interpreters. Our approach enhances both accuracy and interpretability, potentially extending CoT reasoning to arbitrarily long contexts and diverse tasks.

Supplementary Material: zip (/attachment?id=MlwinTkbNc&name=supplementary_material)

Primary Area: Deep Learning->Large Language Models

Keywords: chain-of-thought, interpretability, markovian language models, reinforcement learning, proximal policy optimization, policy gradient, cross-model generalization, gsm8k, wikipedia text

Ethics Agreement: I certify that all co-authors of this work have read and committed to adhering to the Call for Papers, Author Instructions, and Publication Ethics.

Reciprocal Reviewing Status: This submission is NOT exempt from the Reciprocal Reviewing requirement. (We expect most submissions to fall in this category.)

Reciprocal Reviewing Author: Max Lamparth (/profile?id=~Max_Lamparth1)

Submission Number: 13880

Add: Withdrawal

Paper Decision

Decision by Program Chairs 30 Apr 2025, 23:13 (modified: 01 May 2025, 05:23) Program Chairs, Authors
Revisions (/revisions?id=mcgasNE81d)

Decision: Reject
Comment:
The paper proposes an approach to generate informative chain-of-thought (CoT) reasoning for LLMs, by encouraging the final output to rely solely on the CoT and not on the original prompt. To achieve this, a variety of RL-based objectives are considered. Experiments demonstrate gains in some settings, such as GSM8K with a bound on the CoT length.

Reviewers generally found the paper to present an interesting idea for an important problem. However, there were a few concerns:

(1) results being of limited scope, primarily restricted to GSM8K. The response clarified that there are additional results on Wikipedia language modeling (although this was noted as being somewhat non-standard for an application of CoT), and multi-step addition.

(2) experimental setup being non-standard, e.g., restricting to fixed-length CoT outputs, using temperature 2 for GSM8K. The response argued for the viability of the proposed setup, but it does appear that the setup is indeed non-standard and needs much more careful justification.

(3) lack of clarity in presentation and mathematical exposition. e.g., lack of clear definition of state and observation spaces in Definition 1, introducing low-level details in Section 4.3.

From the AC's reading, we particularly concur with point (3). While the paper may have some interesting ideas, they are not yet expressed clearly enough. Further, there is scope for a more exhaustive empirical comparison on a broader range of reasoning benchmarks (e.g., datasets from the original CoT paper). We thus believe the paper may benefit from further significant revisions prior to publication.

Official Review of Submission13880 by Reviewer G9RG

Official Review by Reviewer G9RG 17 Mar 2025, 22:12 (modified: 12 Apr 2025, 11:44)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer G9RG
Revisions (/revisions?id=q8Da4VM7kZ)

Summary:
This work addresses the issue of Chain-of-Thought (CoT) reasoning failing to accurately reflect a language model's decision process. The authors propose a Markovian language model that factors next-token prediction through an intermediate CoT, ensuring that CoT reasoning is causally essential rather than just auxiliary. They introduce an informativeness objective to measure how much CoT improves predictions and optimize it using policy gradient methods.

update after rebuttal

Thanks to authors for clarifying my concerns. I checked other reviewers' concerns as well and respect their thoughts on this paper.

Claims And Evidence:

Claim: an informative CoT can also serve as a recipient-specific compression of the model's hidden knowledge. This distills the essential reasoning into text that another recipient can use to predict the same outcome. Evidence: Their experiments demonstrate that the learned CoTs remain consistent across different interpreters, indicating that these textual explanations capture generalizable problem-solving steps rather than being tied to specific model behaviors.

Methods And Evaluation Criteria:
Method: They define reward as CoT informativeness, and use RL to produce high-reward CoTs. In the paper, they explored three RL techniques (TEI, PG, and PPO). They evaluate the proposed method on arithmetic problems and GSM8K dataset, observed 33.2% improvement on GSM8K. For evaluation, see experimental design on analyses section

Theoretical Claims:
Section 3 provides formal definition of informativeness that they use as a training objective. They train the model to increase the likelihood of correct next token compared to the unmodified model. Section 4 provides their method in formal notations. All claims are sound.

Experimental Designs Or Analyses:
They fine-tuned a mistral 7b instruction model to produce CoT tokens such that a pretrained LLM can use this to predict an answer given this CoT. PPO shows correct answer upto 90% and achieve good accuracy. In GSM8K test, authors used policy gradient with expert iteration. They measured informativeness and correctness measure. Their method achieved good math reasoning tasks. Proposed method was also tested on general language modeling using Wikipedia text. Results shows improvement in accuracy.

Supplementary Material:
Appendix includes additional performance analysis and sampled generation of their trained model.

Relation To Broader Scientific Literature:
This work is limited to the language model training.

Essential References Not Discussed:
Most of recent works are discussed.

Other Strengths And Weaknesses:
(+) problem is interesting. bringing concept of informativeness is clever. (+) good problem formulation and explored many different options (-) limited experiments. experiments are only using one task with GSM8K. some doubts on whether this can be generalized to other tasks. even authors included one language task (Wikitext) it will be better to include other standard reasoning tasks and show effectiveness. (-) as authors mentioned, it is unclear how to use this method to provide long-term future behavior

Other Comments Or Suggestions:
Please see other sections

Questions For Authors:
can you discuss more about how to use this method to provide long-term future behavior?

Code Of Conduct: Affirmed.
Overall Recommendation: 3: Weak accept (i.e., leaning towards accept, but could also be rejected)

−
=
≡

Rebuttal by Authors

Rebuttal
by Authors (Clark Barrett (/profile?id=~Clark_Barrett1), Max Lamparth (/profile?id=~Max_Lamparth1), Scott Viteri (/profile?id=~Scott_Viteri1), Peter Chatain (/profile?id=~Peter_Chatain1))
28 Mar 2025, 22:03 (modified: 01 Apr 2025, 07:27)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=A0K2TTFF6E)

Rebuttal:
Thank you for your thoughtful review of our work. We appreciate your recognition of the paper's contributions, particularly regarding our informativeness objective and the demonstrated improvements on GSM8K.

Regarding your question about long-term future behavior: Our Markovian framework naturally extends to multi-round interactions. The same procedure would apply, where we produce a state given the previous state and observation, but iteratively over an entire trajectory. We could then use REINFORCE for optimization, with the reward being the average log probability of next observations given current states.

This extension is conceptually straightforward within our framework, as the CoT effectively serves as a recurrent state that carries forward relevant information. In Section 7 (Future Work), we briefly touch on this potential, noting: "Although we focus on single question-answer pairs in this paper, the Markovian framework naturally extends to multi-turn dialogue." We agree this is a promising direction for future research and appreciate your interest in this aspect of our work.

Regarding your concern about limited experiments, we would like to highlight that beyond GSM8K, we also tested on long arithmetic problems (15-term addition) as well as Wikipedia text prediction, demonstrating the approach's versatility across different types of tasks. Our results consistently show improved performance and increased CoT fragility across all these domains, suggesting the approach generalizes well.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer G9RG

Rebuttal Acknowledgement by Reviewer G9RG 01 Apr 2025, 15:33

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Official Review of Submission13880 by Reviewer pzdW

Official Review by Reviewer pzdW 14 Mar 2025, 02:47 (modified: 24 Mar 2025, 22:44)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pzdW

Revisions (/revisions?id=DrETp1ffBL)

Summary:
This paper introduces a Markovian framework for training language models to generate more informative and reliable Chain-of-Thought (CoT) reasoning. The authors demonstrate the effectiveness of their approach through experiments on GSM8K, achieving a 33.2% improvement on GSM8K.

Claims And Evidence:
NA

Methods And Evaluation Criteria:
The proposed methods make good sense for evaluating Chain-of-Thought (CoT) language modeling.

Theoretical Claims:
Looking through the paper, I don't see any formal mathematical proofs or theoretical claims that require rigorous verification.

Experimental Designs Or Analyses:
While the paper compares three different RL approaches (TEI, Policy Gradient, and PPO), there appears to be limited direct experimental comparison with other leading CoT reasoning models like Deepseek R1. A more comprehensive comparative analysis with such models could have provided deeper insights into the relative strengths and tradeoffs of their Markovian approach.

Supplementary Material:

Yes, I have reviewed the complete supplementary material

Relation To Broader Scientific Literature:

It introduces the novel concept of "informativeness" as a training objective for CoT generation, moving beyond traditional approaches that focused on mimicking human-written reasoning chains, and instead optimizing for causal reasoning that directly contributes to accurate predictions.

Essential References Not Discussed:

Lack of discussion on the paper ''Understanding Chain-of-Thought in LLMs through Information Theory''.

Other Strengths And Weaknesses:

- Lacks computational resource requirements and training times
- Methods section could be more accessible to readers. Some mathematical notations could be better explained
- May not scale well to more complex multi-step reasoning

Other Comments Or Suggestions:

NA

Questions For Authors:

- Could the author elaborate on whether this Markovian training approach could be extended to LLMs with self-reflection capabilities that generate longer chains-of-thought (R1,O1)?
- Can the author provide more details about the computational costs and training resources required for implementing this method, beyond the H100 GPU mentioned?
- How might this CoT training approach be meaningful or applicable for more open-ended domains like creative writing, beyond the structured arithmetic and reasoning tasks explored?

Code Of Conduct: Affirmed.

Overall Recommendation: 2: Weak reject (i.e., leaning towards reject, but could also be accepted)

Rebuttal by Authors

Rebuttal

by Authors (Clark Barrett (/profile?id=~Clark_Barrett1), Max Lamparth (/profile?id=~Max_Lamparth1), Scott Viteri (/profile?id=~Scott_Viteri1), Peter Chatain (/profile?id=~Peter_Chatain1))

28 Mar 2025, 22:04 (modified: 01 Apr 2025, 07:27)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=l3aRDbA5Qq)

Rebuttal:

Thank you for your review. We'd like to address several points you raised:

Regarding comparison with DeepSeek-R1: The DeepSeek-R1 paper was released on January 22nd, just 8 days before the ICML submission deadline (January 30th), making a thorough comparison infeasible within our timeline. More importantly, R1 is a 671B parameter model, while our experiments used Llama at 8B parameters, making direct comparisons methodologically problematic due to the vast scale difference.

We should also note that we could not compare against other models like O1 because the OpenAI API for O1 doesn't output log probabilities, which are essential for our reward calculation (https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/reasoning?tabs=python-secure (https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/reasoning?tabs=python-secure)).

Regarding computational requirements: A full training run takes approximately 5 hours on a single H100 GPU. We're happy to add this information to the paper to provide readers with a concrete understanding of the resource requirements.

On scaling to longer chains-of-thought: Our method actually works better with longer CoTs in terms of performance, though we observe that longer CoTs would be less fragile to perturbation (since more redundant information could be included). This is a natural trade-off between performance and interpretability that future work could explore further.

Regarding application to open-ended domains: We successfully applied our approach to Wikipedia text continuation, which is relatively unstructured compared to mathematical reasoning. Interestingly, the fragility-to-perturbation results are even more pronounced in this domain since the CoT is short relative to the length of text to predict. This suggests potential applicability to creative writing or other open-ended tasks, with the CoT serving as a concise plan or outline that guides generation.

We'll be happy to cite "Understanding Chain-of-Thought in LLMs through Information Theory" in our related work section, as it provides valuable analytical insights that complement our reinforcement learning approach.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer pzdW

Rebuttal Acknowledgement by Reviewer pzdW 02 Apr 2025, 23:50

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Official Review of Submission13880 by Reviewer CMFC

Official Review by Reviewer CMFC 13 Mar 2025, 01:13 (modified: 09 Apr 2025, 13:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer CMFC

Revisions (/revisions?id=buGIjWBGXP)

Summary:
The paper proposes an "informativeness" objective for CoT generation, which makes the model learn to generate CoT that can infer the true answer more likely without the original question. In this way, models learn to generate more informative and potentially more faithful CoTs. Training on the proposed objective leads to CoTs that are more causally related to the final answer and on some tasks also informative to other models.

update after rebuttal

Despite the rather lengthy response, the authors only attempted to address part of my concerns about the paper. Even regarding the issues discussed, I am not convinced that the experiment setup is realistic and reasonable (e.g., using a rather short fixed-length CoT) and that the results support the main claims. Also, there are major problems with the lack of readability of the paper:

- Math notations need to be optimized. Section 3.2 is particularly hard to follow: the state and observation are not clearly defined; it is unclear what subscript t means; it is unclear what x_t's means as it does not seem to correspond to any variable in Figure 1; this math model is not really used later.
- Too many implementation details which are not clearly related to the research question behind the experiments, distracting the readers from understanding the paper: Section 4.3, details in Section 5.1 5.2 5.3
- The delivery needs a bit more work. Should spend more time clearly describing the research questions/hypothesis to be studied in each section, what the main findings are, and how the results support the findings rather than just explaining the numbers.

I think this can potentially be an interesting paper, but at the current point, many aspects need to be improved. So, I would keep my initial evaluation.

Claims And Evidence:
The paper is rather poorly written and hard to follow. Most main claims are not well-stated, and it is unclear how the results support the claims under unclear or rather unusual experimental setups.

- "We formulate an "informativeness" objective for CoT generation and develop a Markovian training procedure that forces an LM to rely on its own CoT." It is unclear how forcing the model to generate more informative CoT is equivalent to forcing an LM to rely on its own CoT.
- "We apply this approach to arithmetic problems (Mistral 7B) and the GSM8K dataset (Cobbe et al., 2021) (Llama 3.1 8B), observing a 33.2% absolute improvement on GSM8K." The experiment setup is quite uncommon: the model is asked to generate a fixed length of answers where many answers might not be complete.
- "We show that perturbing the CoT consistently degrades prediction accuracy, verifying fragility and causal relevance." The prediction accuracy is measured by

  $$\ln \pi(ans|CoT)$$

  where the original question is not given. Since the CoT is the only information source about the final answer, isn't it obvious that perturbing CoT would consistently degrade prediction accuracy?
- "We demonstrate cross-model transfer: CoTs trained on one model remain informative for other models. This underscores the CoT's recipient-specific interpretability and suggests it captures a shared reasoning strategy." The authors themselves refer to the Wikipedia dataset as general language modeling; I don't understand how results on this dataset can imply anything about "reasoning strategy".

Methods And Evaluation Criteria:
The concept of recipient-specific compression and the designed reward function makes sense. But since in general we usually have the length of CoT to be larger than the answer or question, the model can just learn to put the final answer in the CoT to increase its "informativeness". It is not very clear why the chosen datasets are good benchmarks for testing the given method.

Theoretical Claims:
There is no theoretical claims.

Experimental Designs Or Analyses:
The experiment and analysis are poorly designed and hard to follow. It is not clear what the research question or hypothesis behind each experiment is and how the designed experiments and results can test them.

Supplementary Material:
I didn't review the supplementary material.

Relation To Broader Scientific Literature:
This paper is broadly related to the interpretability/understanding and improving faithfulness of CoT.

Essential References Not Discussed:
There is only very limited discussion on related work. The authors should spend more space discussing work on the faithfulness of CoT [1, 2, 3, 4 ...], understanding CoT [5 ...], and prior work on training LLMs to generate CoT in more detail. [1] Question Decomposition Improves the Faithfulness of Model-Generated Reasoning [2] On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models [3] Deductive Verification of Chain-of-Thought Reasoning [4] PINTO: faithful language reasoning using prompt-generated rationales. [5] Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Other Strengths And Weaknesses:
The mathematical model is not necessary or useful. Too many unnecessary implementation details.

Other Comments Or Suggestions:
The paper needs a lot of work on almost every aspect.

Questions For Authors:
Why trained model's CoTs are more fragile if the model used to calculate the answer likelihood are the same?

Code Of Conduct: Affirmed.
Overall Recommendation: 1: Reject

Rebuttal by Authors

Rebuttal

by Authors (Clark Barrett (/profile?id=~Clark_Barrett1), Max Lamparth (/profile?id=~Max_Lamparth1), Scott Viteri (/profile?id=~Scott_Viteri1), Peter Chatain (/profile?id=~Peter_Chatain1))

28 Mar 2025, 22:05 (modified: 01 Apr 2025, 07:27)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=0LNdWjQglm)

Rebuttal:

Thank you for your detailed review. We appreciate your critical feedback and would like to address your concerns:

Regarding how our approach makes CoT causally essential: There's an important distinction to clarify - it is specifically the Markovian training procedure that forces the LM to rely on its own CoT by removing everything else from the context and using RL in that setting. The informativeness objective then quantifies and optimizes how useful this CoT is for prediction.

On the experimental setup: We adopt the standard practice of predicting however many tokens are in the answer. Your confusion may stem from our theoretical setup, which uses fixed-length observations for simplicity of presentation. This applies more directly to the Wikipedia benchmark, where we draw from one long input stream. We simplified the diagrams by using a fixed observation type, but this doesn't affect the actual implementation.

Regarding CoT fragility: You correctly note that our setup will inherently lead to CoT fragility (which is indeed our goal), but the non-trivial finding is that we could successfully train language models to predict future tokens despite this severe handicap. The experiment demonstrates both that this training is possible and that it leads to substantial performance improvements across multiple tasks.

On Wikipedia results and reasoning: When referring to Wikipedia as "general language modeling," we simply mean prediction of natural language text without a specific task structure, not implying any claims about AGI capabilities. For evidence of reasoning strategies, please see our GSM8K reasoning benchmarks and the example model's learned strategy on long arithmetic problems in Appendix C, which demonstrates clear step-by-step calculation processes.

Regarding CoT containing answers directly: This is indeed a potential concern we considered. As noted in Section 4.1: "Our conceptual arguments rely on K < N, as otherwise the model could simply write the predicted observation into the state. We satisfy this criteria in our Wikipedia experiments (Sec 5.3), and for the other experiments we find empirically that the model does not learn this undesirable behavior due to the relative difficulty of predicting the answer directly without any CoT."

On experimental clarity: Our paper tests three main hypotheses: (1) our training procedure can optimize the stated objective successfully, (2) it increases fragility of the CoT to perturbation, and (3) the resulting CoT is interpretable in that others can use it to predict future text. We demonstrate these across three distinct datasets - arithmetic (simple proof of concept), GSM8K (mathematical reasoning), and Wikipedia text (natural language capabilities).

Regarding your specific question about trained model's CoTs being more fragile: This occurs because the trained CoTs are genuinely more useful, creating a larger performance delta when perturbed. We also experimented with training the recipient model but found this destabilized the training procedure.

We appreciate your suggested references and will incorporate them in our related work discussion. Thank you for helping us improve the paper.

Replying to Rebuttal by Authors

Rebuttal Comment
by Reviewer CMFC

Rebuttal Comment by Reviewer CMFC 02 Apr 2025, 19:05

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thanks for the detailed response. However, many of my initial concerns haven't been fully addressed:

> Regarding how our approach makes CoT causally essential: There's an important distinction to clarify - it is specifically the Markovian training procedure that forces the LM to rely on its own CoT by removing everything else from the context and using RL in that setting. The informativeness objective then quantifies and optimizes how useful this CoT is for prediction.

Thank you for clarifying. I understand how the objective might increase the informativeness of CoTs and thus make them more causally essential. My concern is more about your original claim: "forces the LM to rely on its own CoT ". The informativeness objective is to force the model to generate "better" CoT that an unmodified model can better infer the **correct** answer from it. This is a property of CoT and says nothing about the LM. The model can rely equally on the old and new CoTs, and the reason the new CoT achieves a high reward is that the old CoT is wrong. In this case, the objective trains the model to generate better CoT rather than "forcing it to rely on its own CoT". To argue that the model relies more on its own CoT, one really needs to show that given the **same** CoT, the model is more likely to predict the answer you can logically deduce from the CoT.

> On the experimental setup: We adopt the standard practice of predicting however many tokens are in the answer. Your confusion may stem from our theoretical setup, which uses fixed-length observations for simplicity of presentation. This applies more directly to the Wikipedia benchmark, where we draw from one long input stream. We simplified the diagrams by using a fixed observation type, but this doesn't affect the actual implementation.

Thanks for clarifying the experimental setup. The confusion really comes from the description in the paper:

- Line 160: "In many tasks like math problem solving, we have T = 2 observations and implement the abstract MLM with fixed-length token sequences. Let V be a token vocabulary. We set O = VN and S = VK for some N, K ∈ N."

Here you explicitly mentioned math problem (I assume applies to GSM8K) and "implement the abstract MLM with fixed-length token sequences". Since according to your math model (Figure 1), answer is just also an "observation" so I assume it has a fixed length. I suggest clarifying this in the paper.

Still, my concern remains: even if the model can generate however many tokens for the final answer, it is still unusual to have a small fixed length of CoT:

- Line 270 "We produce 150 CoT tokens, sampled at temperature 2.0."

The CoT from the original model might not even be complete for many questions, which leads to an underestimation of the true usefulness of the original CoTs. As a result, the model might just learn to shorten their CoTs to be within 150 tokens rather than doing anything really related to be more "informative". In this case arguing the method leads to improvement of accuracy does not really make sense. The right plot in figure 3 also seems to suggest this: as the training goes, the model learns to generate CoT that contains answers. My hypothesis is that the model will gradually learn to complete the CoT in 150 tokens. Also temperature 2.0 is unusually high for math problems.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer CMFC

Rebuttal Acknowledgement by Reviewer CMFC 02 Apr 2025, 19:05

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Replying to Rebuttal Comment by Reviewer CMFC

Reply Rebuttal Comment by Authors

Reply Rebuttal Comment

by Authors (Clark Barrett (/profile?id=~Clark_Barrett1), Max Lamparth (/profile?id=~Max_Lamparth1), Scott Viteri (/profile?id=~Scott_Viteri1), Peter Chatain (/profile?id=~Peter_Chatain1))

05 Apr 2025, 20:52 (modified: 05 Apr 2025, 22:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=i21Hbdyhog)

Comment:

> Thank you for clarifying... My concern is more about your original claim: "forces the LM to rely on its own CoT"...says nothing about the LM

Regarding our claim that our method "forces the LM to rely on its own CoT" - your concern is that we're training models to produce better CoTs objectively rather than creating recipient-specific CoTs. This is valid, especially for verbatim answers. For Wikipedia examples in our plots, however, there's no single "best" CoT given the extensive extrapolation needed.

Suppose we have a sender LM producing the CoT which a recipient LM uses to predict subsequent text. If the subsequent text includes 10 facts, but the sender only has room in the CoT for 5 facts, then the sender will need to take into account which 5 facts the recipient is already likely to know.

The learned CoT may reveal more about the recipient than the sender, but it also informs us about the sender. The sender must know those 5 facts, and selecting the right ones suggests knowledge about what the recipient knows. Also, with ~99.75% shared parameters between sender and recipient, learning about one directly informs our understanding of the other.

> The model can rely equally on the old and new CoTs... rather than "forcing it to rely on its own CoT"

Testing CoT reliance by comparing trained vs. untrained CoT after removing context is admittedly contrived. In this setting, the model relies on both CoTs as they're the only information available.

In figure 4, we plot m_1 = (ln P(ans|CoT) - ln P(ans|perturb(CoT))) - (ln P(ans|CoT') - ln P(ans|perturb(CoT'))) comparing perturbation effects between trained and untrained models to show how fragility relates to CoT choice.

We've now also compared CoT fragility with/without the question using metric m_2 = (ln P(ans|CoT) - ln P(ans|perturb(CoT))) - (ln P(ans|q, CoT) - ln P(ans|q, perturb(CoT))). These plots still show increased fragility.

You're concerned we're measuring CoT accuracy instead of fragility. Consider a case where the perturbation removes the correct answer from CoT. Figure 4 might show high values when untrained CoT' excludes the answer: m_1 ~ ((high-low)-(low-low)). Our plots using the same CoT with/without the question address this concern as the strong CoT appears in both terms: m_2 ~ ((high-low)-(high-low)). We will include these plots in the paper.

> To argue that the model relies more on its own CoT...

Models can "rely" on text in multiple ways: (1) generating CoTs targeted to specific recipients so they can use them better than other models; or (2) being sensitive to text modifications that affect future predictions. We focus on the second kind to avoid potential steganography from overfitting to specific recipients. Figure 5 demonstrates we're not overfitting in this way.

> Here you explicitly mentioned math problem... I suggest clarifying this in the paper.

We meant fixed-length CoT token sequences per datapoint. You're right that this phrasing is misleading as it could suggest fixed-length observation sequences. We've corrected this in the paper. Thanks!

> The CoT from the original model might not even be complete... will learn to complete the CoT in 150 tokens ...

If 150 tokens were insufficient, RL might just condense reasoning into the allotted space. This would be desirable for CoT fragility but wouldn't necessarily indicate improved performance.

Empirically, this isn't driving performance. Our method works well even when models don't benefit from more tokens.

In our Mistral arithmetic runs (Figure 1), the untrained model answers before the CoT ends and uses remaining tokens ineffectively. See Appendix C2 for an example.

With GSM8K, Llama can get cut off, but we believe allowing more tokens wouldn't decrease performance because: (1) Anecdotally, in prior experiments we have seen similar results with partial training runs and more allotted tokens, though we tuned down that hyperparameter for a less expensive training run (2) Llama 3 8B, like Mistral, lacks the strategic wherewithal to use extra context tokens productively, decreasing the degree to which untrained Llama would benefit from more reasoning tokens

Example - Llama on GSM8K before training:

Question: "Jordan's dog, Max, was born on the same day that Luka turned 4 years old. Luka is exactly 2 years older than is Aubrey. On Aubrey's 8th birthday, how many years old was Jordan's dog, Max?"

Reasoning: "Luka is 2 years older than Aubrey. So, if Aubrey is 8, then Luka is 8 + 2 = 10. Luka was 4 when Max was born. So, Luka is 10 - 4 = 6 years older than Max. Max is 6 years younger than Luka. So, Max is 8 years old. The final answer is 8. I hope it is correct. \n\nThis problem is a good example of a multi-step problem that requires the student to use reasoning and problem-solving skills. The student must be able to follow the steps and use the information given to arrive at the correct answer. This type"

Answer: "6"

Thank you for your detailed and helpful feedback!