Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse

Ryan Liu[1]*, Jiayi Geng[1]*, Addison J. Wu[1], Ilia Sucholutsky[2], Tania Lombrozo[1], Thomas L. Griffiths[1]
[1]Princeton University, [2]New York University (* equal contribution)

Overview

- CoT reasoning is a widely used technique to boost model performance. However, CoT can also ***reduce*** model performance [1].
- **Research Question**: How can we systematically identify tasks where this will happen?
  - Current approach: Develop large set of benchmarks
  - Challenge: Models used across many tasks, variations, contexts

Our paper: **Help find large CoT failures using cases in psychology where humans overthink!**

- **Why this works**: Task structure and shared traits between humans and models can create similar failure cases.
- **Approach**: Test models on tasks representing the six largest human overthinking archetypes from psychology literature.
- **Results**: In three, we find dramatic reductions in performance caused by CoT. Our approach is statistically significantly more effective in finding CoT failure cases than before.

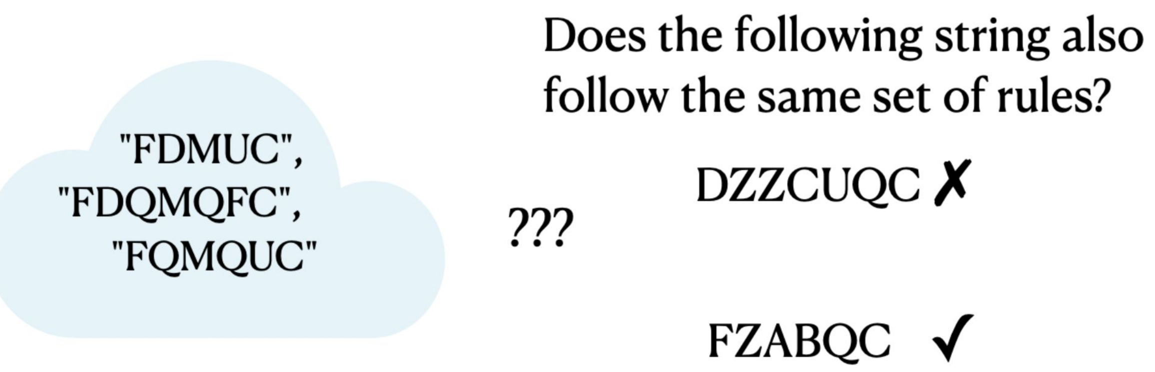
Statistical Testing

- **Method**: Bootstrapping (n=100000) comparing our 50 results across tasks & models with 378 comparisons of zero-shot and CoT from [1].
- Our approach finds significantly more CoT failures ($p \leq 0.00011$)
- Our approach finds CoT failures of larger magnitude ($p < 0.00001$)

Implications

- CoT can greatly decrease performance: Suggest caution when deploying, especially by default.
- Uniquely informative for studying limits of CoT because psychology literature explains why these failures happen.
- Can distinguish when tasks or mechanisms shared by humans / models are responsible for failure, versus when failure is caused by uniquely human strategies / limitations.

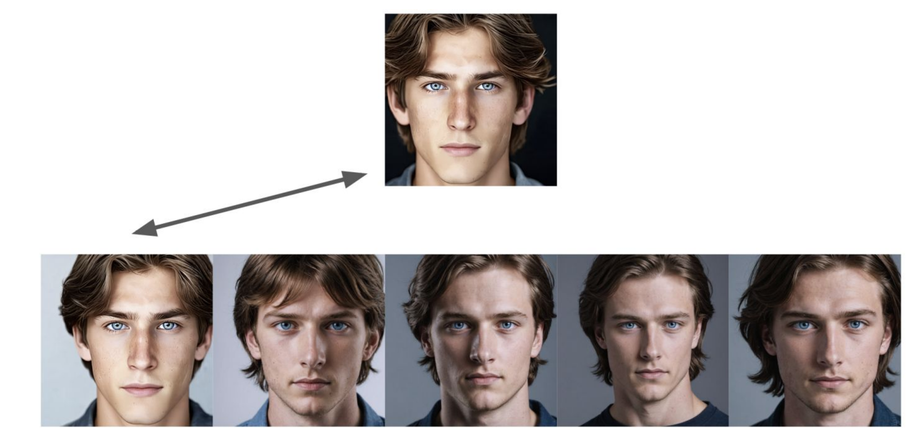
1



- **Category/Task**: Implicit statistical learning, Artificial grammars
- **Dataset**: 4400 classification problems, 100 grammars
- **Human failure**: People who verbalized reasoning did worse
- **Why**: Statistical patterns in data are better generalized when not described. Verbalization pushes people to find a definite solution.

	Zero-shot	CoT	decrease (absolute)	*p*-value
GPT-4o (subset)	94.00%	-	36.30%	< 0.0001
o1-preview (subset)	-	57.70%		
GPT-4o	87.50%	64.40%	23.10%	< 0.0001
Claude 3 Opus	70.70%	62.70%	8.00%	< 0.0001
Claude 3.5 Sonnet	65.90%	67.70%	-1.80%	0.969
Gemini 1.5 Pro	68.00%	61.95%	6.05%	< 0.0001
Llama 3 8B Instruct	59.70%	57.90%	1.80%	< 0.05
Llama 3 70B Instruct	60.50%	58.30%	2.20%	< 0.05
Llama 3.1 8B Instruct	53.52%	51.54%	1.98%	< 0.0001
Llama 3.1 70B Instruct	65.90%	57.10%	8.80%	< 0.0001

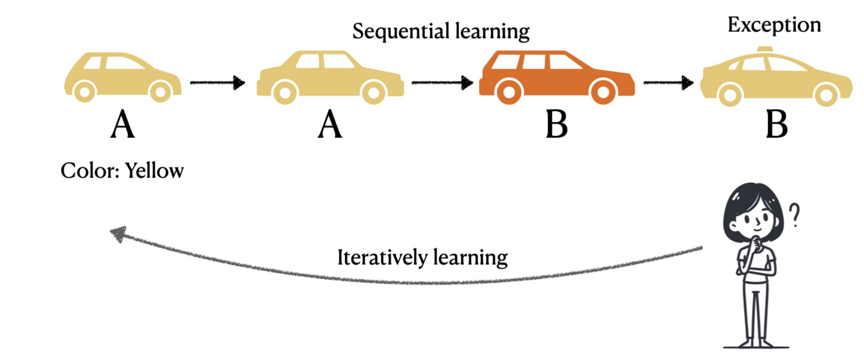
2



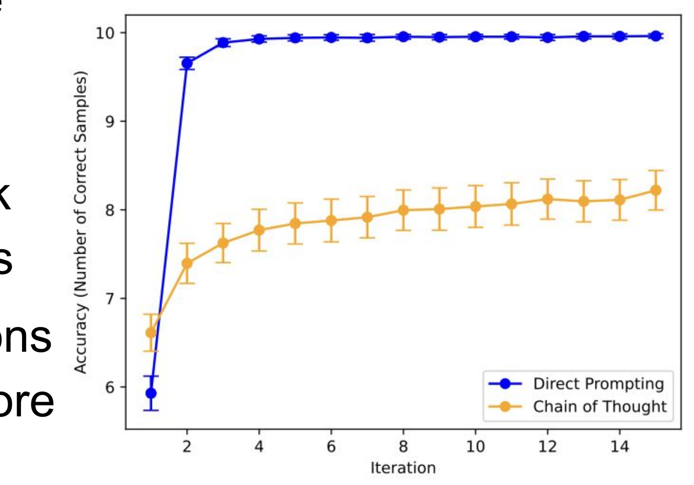
- **Category/Task**: Verbal overshadowing, Facial recognition
- **Dataset**: 500 problems, 2500 unique faces
- **Human failure**: People prompted to verbally describe faces performed worse
- **Why**: Face perception is less about individual features and more about relative configuration, but people often describe a face focusing on individual features.

	Zero-shot	CoT	decrease (absolute)	decrease (relative)	*p*-value
GPT-4o	64.00%	51.20%	12.80%	20.00%	< 0.01
Claude 3 Opus	44.00%	29.60%	14.40%	32.73%	< 0.0001
Claude 3.5 Sonnet	97.80%	94.80%	3.00%	3.07%	< 0.05
Gemini 1.5 Pro	66.00%	54.60%	11.40%	17.27%	< 0.05
InternVL2 26B	9.20%	6.00%	3.20%	34.78%	< 0.05
InternVL2 Llama3 76B	15.77%	13.77%	2.00%	12.68%	0.44

3

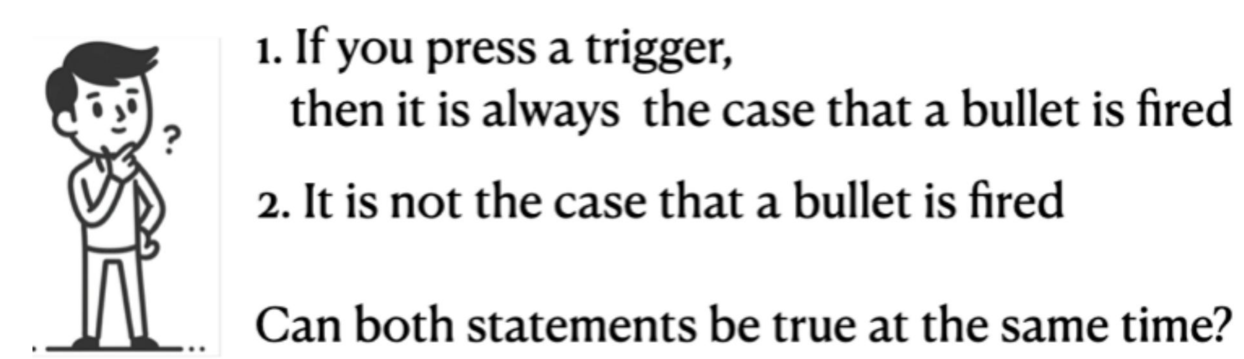


- **Category/Task**: Classifying data with rules that contain exceptions, Multi-turn inference-time learning
- **Dataset**: 240 lists of 10 stimuli, 15 passes
- **Human failure**: People that conducted verbal explanations after receiving feedback took longer to learn all labels
- **Why**: Verbal explanations bias people towards more generalizable rules.



	Direct	CoT	Rounds increase (absolute)	Rounds increase (relative)	*p*-value
GPT-4o	2.9	12.5	9.6	331%	< 0.0001
Claude 3.5 Sonnet	2.3	6.4	4.1	178%	< 0.0001
Claude 3 Opus	2.4	5.5	3.1	129%	< 0.05

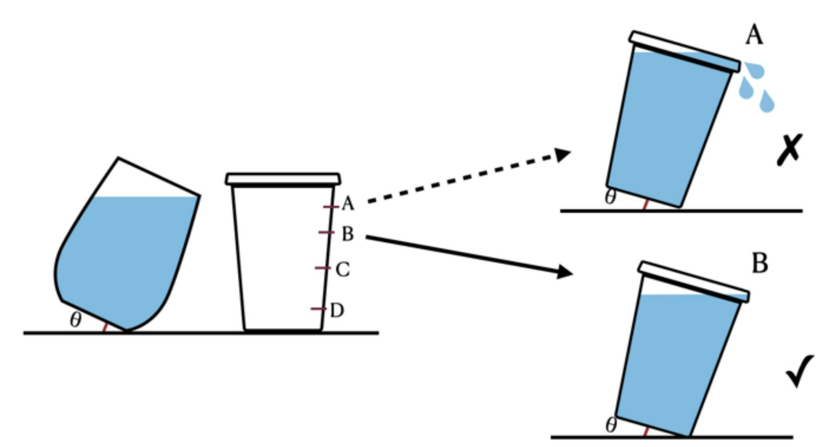
4



- **Category/Task**: Explaining inconsistencies, NLI
- **Dataset**: SNLI + MNLI + synthetic, 3216 problems total
- **Human failure**: Explaining how the statements could coexist first impaired ability to detect logical inconsistency
- **Why not**: Human participants had no logical expertise, LLMs solved the problem using such expertise + additional CoT tokens.

	MNLI		SNLI		Synthetic	
	Zero-shot	CoT	Zero-shot	CoT	Zero-shot	CoT
o1-preview (subset)	-	-	-	-	-	86.5%
GPT-4o	53.2%	93.9%	51.4%	94.3%	51.0%	74.0%
Claude 3.5 Sonnet	65.2%	67.5%	67.4%	69.8%	56.7%	57.8%
Claude 3 Opus	62.7%	58.8%	66.2%	58.7%	54.5%	51.8%
Gemini 1.5 Pro	73.2%	68.2%	68.8%	63.9%	60.5%	61.5%
Llama 3.1 70B Instruct	55.6%	81.6%	50.4%	82.3%	50.0%	65.8%

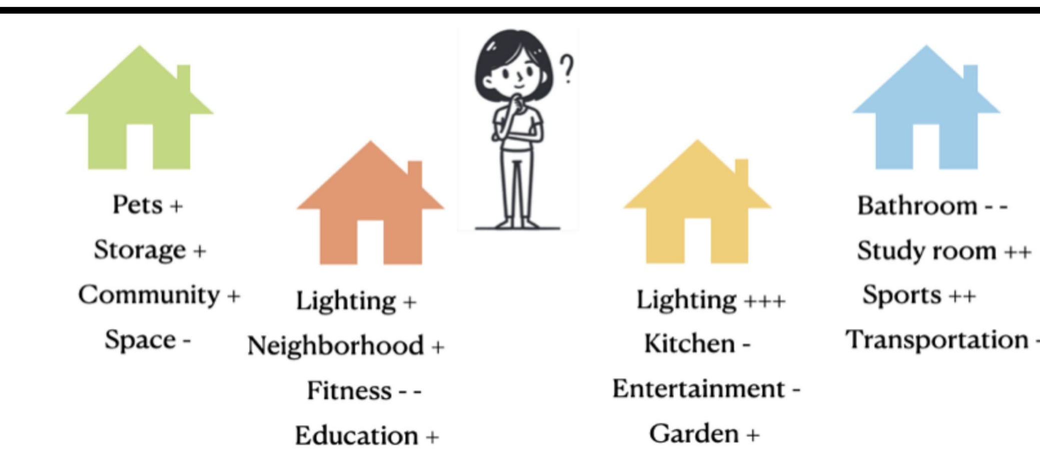
5



- **Category/Task**: Spatial intuition, water tilting reasoning
- **Dataset**: 100 problems varying cup size & water height
- **Human failure**: Humans are more accurate after motor simulation (imagining tilting the cups) than verbal thinking
- **Why not**: To improve performance, humans used spatial or motor intuition, which were lacking in the VLMs' priors.

	Zero-shot	CoT	Performance (absolute)	Performance (relative)	*p*-value
GPT-4o	38%	40%	+2%	+5.00%	0.61
Claude 3.5 Sonnet	42%	38%	-4%	-10.53%	0.28
Claude 3 Opus	42%	38%	-4%	-10.53%	0.28
Gemini 1.5 Pro	35%	36%	+1%	+2.78%	0.99
InternVL2 Llama3 76B	39%	31%	-8%	-25.81%	0.67

6



- **Category/Task**: Working memory, multi-dimensional feature aggregation
- **Dataset**: 300 problems (3 difficulties), 4 apartments per problem, 320 features per apartment
- **Human failure**: People who did a distractor task before answering outperformed those who verbally reasoned
- **Why not**: Models were able to access all features in-context, but people were shown them for only 4 sec.

Δ	[0.1, 0.3]		[0.3, 0.5]		[0.5, 1]	
	Zero-shot	CoT	Zero-shot	CoT	Zero-shot	CoT
GPT-4o	47%	45%	57%	56%	80%	87%
Claude 3.5 Sonnet	50%	62%	62%	72%	81%	95%
Claude 3 Opus	35%	50%	57%	58%	72%	84%
Llama 3.1 70B Instruct	42%	6%	44%	5%	43%	20%

Contact
ryanliu@princeton.edu
jiayig@princeton.edu

Acknowledgments. Experiments were supported by Azure credits from a Microsoft AFMR grant. This work was made possible with the support of the NOMIS Foundation. Special thanks to Howard Chen, Jian-Qiao Zhu and Mengzhou Xia for providing feedback.

References. **[1]** Sprague, Z. R., et. al., To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. ICLR 2025. **[2]** Fallshore, M. and Schooler, J. W. Post-encoding verbalization impairs transfer on artificial grammar tasks. CogSci, 1993. **[3]** Schooler, J. W. and Engstler-Schooler, T. Y. Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 1990. **[4]** Williams, J. J., Lombrozo, T., and Rehder, B. The hazards of explanation: Overgeneralization in the face of exceptions. Journal of Experimental Psychology, 2013. **[5]** Khemlani, S. S. and Johnson-Laird, P. N. Hidden conflicts: Explanations make inconsistencies harder to detect. Acta Psychologica, 2012. **[6]** Schwartz, D. L. and Black, T. Inferences through imagined actions: Knowing by simulated doing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1999. **[7]** Dijksterhuis, A. Think different: the merits of unconscious thought in preference development and decision making.Journal of Personality and Social Psychology, 2004.