



# Argumentation

The Missing Piece for Natural Language Reasoning in LLMs

*Ramon Ruiz-Dolz*  
*Università della Svizzera Italiana - March 2025*



Reasoning in LLMs

---

## Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma  
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou  
Google Research, Brain Team  
{jasonwei, dennyzhou}@google.com

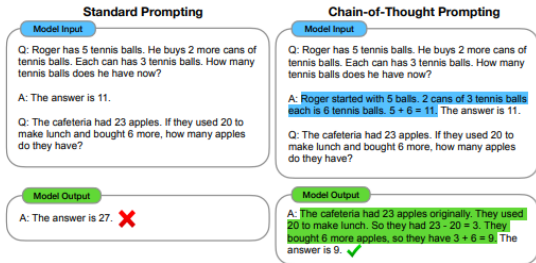
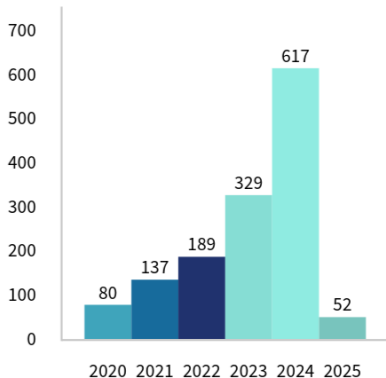


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

# The “boom” of reasoning in NLP

Papers with the word “*reasoning*” in the title in major NLP conferences (ACL, EMNLP, NAACL, EACL, COLING, LREC):

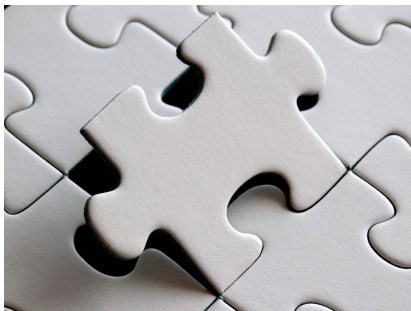
- 2020: 80
- 2021: 137
- 2022: 189
- 2023: 329
- 2024: 617
- 2025: 52 (only COLING)



# The “boom” of reasoning in NLP

But... is it all this work about **natural language reasoning**?

Or more about **solving problems** that involve some kind of reasoning... in natural language?



# Reasoning in Natural Language vs. Natural Language Reasoning

## GSM8K (Cobbe at al., 2021) - Mathematical Reasoning

**Problem:** Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

**Solution:** Beth bakes 4 2 dozen batches of cookies for a total of  $4 \times 2 = 8$  dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of  $12 \times 8 = 96$  cookies

She splits the 96 cookies equally amongst 16 people so they each eat  $96/16 = 6$  cookies

**Final Answer:** 6

**Problem:** Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons = 50 gallons this morning.

So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons.

She was able to sell 200 gallons - 24 gallons = 176 gallons.

Thus, her total revenue for the milk is  $\$3.50/\text{gallon} \times 176 \text{ gallons} = \$616$ .

**Final Answer:** 616

**Problem:** Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

**Solution:** Tina buys 3 12-packs of soda, for  $3 \times 12 = 36$  sodas

6 people attend the party, so half of them is  $6/2 = 3$  people

Each of those people drinks 3 sodas, so they drink  $3 \times 3 = 9$  sodas

Two people drink 4 sodas, which means they drink  $2 \times 4 = 8$  sodas

With one person drinking 5, that brings the total drank to  $9 + 8 + 5 = 25$  sodas

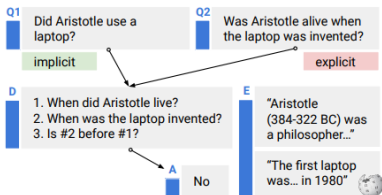
As Tina started off with 36 sodas, that means there are  $36 - 25 = 11$  sodas left

**Final Answer:** 11

Other similar benchmarks: SVAMP, ASDiv, AQuA, MAWPS, ...

# Reasoning in Natural Language vs. Natural Language Reasoning

## StrategyQA (Geva et al., 2021) - Commonsense Multi-Hop Reasoning



Can the President of Mexico vote in New Mexico primaries?

- (1) What is **the citizenship requirement** for *voting in New Mexico*?
- (2) What is **the citizenship requirement** of any *President of Mexico*?
- (3) Is #2 the same as #1?

Can a microwave melt a Toyota Prius battery?

- (1) What **kind of battery** does a *Toyota Prius* use?
- (2) What **type of material** is #1 made out of?
- (3) What is the **melting point** of #2?
- (4) Can a *microwave's temperature* reach at least #3?

Would it be common to find a penguin in Miami?

- (1) Where is a typical *penguin's natural habitat*?
- (2) What **conditions** make #1 suitable for *penguins*?
- (3) Are all of #2 present in *Miami*?

Other similar benchmarks: CommonsenseQA, BIG-bench, ...

# Reasoning in Natural Language vs. Natural Language Reasoning

## FOLIO (Han et al, 2024) - First-Order Logic Reasoning

### NL premises

1. There are six types of wild turkeys: Eastern wild turkey, Osceola wild turkey, Gould's wild turkey, Merriam's wild turkey, Rio Grande wild turkey, and the Ocellated wild turkey.
2. Tom is not an Eastern wild turkey.
3. Tom is not an Osceola wild turkey.
4. Tom is also not a Gould's wild turkey.
5. Tom is neither a Merriam's wild turkey, nor a Rio Grande wild turkey.
6. Tom is a wild turkey.

### FOL Premises

1.  $\forall x(\text{WildTurkey}(x) \rightarrow (\text{EasternWildTurkey}(x) \vee \text{OsceolaWildTurkey}(x) \vee \text{GouldsWildTurkey}(x) \vee \text{MerriamsWildTurkey}(x) \vee \text{RiograndeWildTurkey}(x) \vee \text{OcellatedWildTurkey}(x)))$
2.  $\neg \text{EasternWildTurkey}(tom)$
3.  $\neg \text{OsceolaWildTurkey}(tom)$
4.  $\neg \text{GouldsWildTurkey}(tom)$
5.  $\neg \text{MerriamsWildTurkey}(tom) \wedge \neg \text{RiograndeWildTurkey}(tom)$
6.  $\text{WildTurkey}(tom)$

### NL Conclusions -> Labels

- A. Tom is an Ocellated wild turkey. -> True
- B. Tom is an Eastern wild turkey. -> False
- C. Joey is a wild turkey. -> Unknown

### FOL conclusions -> Labels

- A.  $\text{OcellatedWildTurkey}(tom) \rightarrow \text{True}$
- B.  $\text{EasternWildTurkey}(tom) \rightarrow \text{False}$
- C.  $\text{WildTurkey}(joey) \rightarrow \text{Unknown}$

Natural Language Reasoning, or symbolic reasoning in natural language?



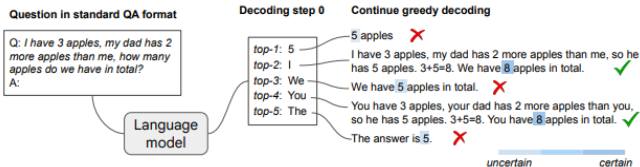
# Are LLMs reasoning at all?

Google DeepMind

## Chain-of-Thought Reasoning without Prompting

Xuezhi Wang<sup>1</sup> and Denny Zhou<sup>1</sup>

<sup>1</sup>Google DeepMind, <sup>1</sup>{xuezhiw, dennyzhou}@google.com



→ Is therefore CoT about reasoning, or about finding a way (either via in-context learning or decoding) that the generated sequence matches a specific structure observed in the training data?

# Argumentative Reasoning and LLMs

---

- P1. Mining Complex Patterns of Argumentative Reasoning in Natural Language Dialogue (**RR**, ZK, JL, Under Review)
  
- P2. Natural Language Reasoning in Large Language Models: Analysis and Evaluation (DG, **RR**, HB, CR, Under Review)

## Mining Complex Patterns of Argumentative Reasoning in Natural Language Dialogue

Argument Mining → Argumentation Scheme Mining

Argumentation Theory → Natural Language Argumentation

E.g., Argument from Waste (Walton (1) vs. Real (2)):

- |     |    |   |     |    |   |
|-----|----|---|-----|----|---|
| (1) | a. | Premise 1: <i>If a stops trying to realise A now, all a's previous efforts to realise A will be wasted.</i> | (2) | a. | Premise: <i>We need to make sure that we <u>embed the successes</u> that we have had.</i> |
|     | b. | Premise 2: <i>If all a's previous attempts to realise A are wasted, that would be a bad thing.</i>          |     | b. | Conclusion: <i>There is still work to do.</i>   |
|     | c. | Conclusion: <i>Therefore, a ought to continue trying to realize A.</i>                                      |     |    |   |

1. NLAS<sup>1</sup>:  
1,902 arguments → 20 schemes, 50 topics, 2 stances
2. NLAS-proc:  
23,771 arguments → enthymematic NLAS
3. QT-Schemes:  
441 arguments → 5 QT episodes, 24 schemes

---

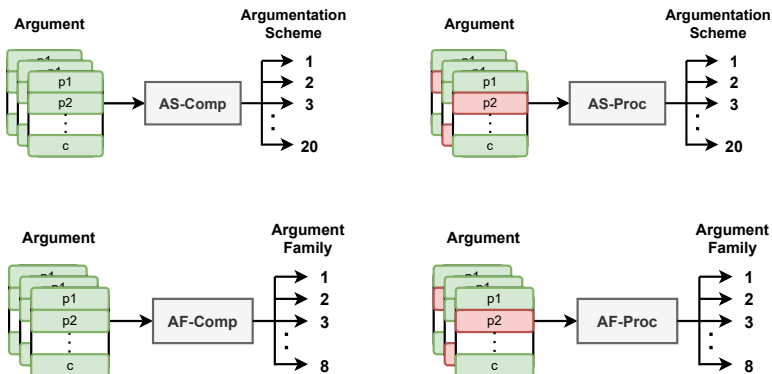
<sup>1</sup>Ramon Ruiz-Dolz, Joaquin Taverner, John Lawrence, and Chris Reed. 2024. Nlas-multi: A multilingual corpus of automatically generated natural language argumentation schemes. Data in Brief, 57:111087.

# P1: Datasets

Argumentation Family	Argumentation Scheme	NLAS		QT-SCHEMES	
		COMP	PROC	Total	Ft/Te
Ad Hominem Arguments	Allegation of Bias	0	0	1	0/1
	Direct Ad Hominem	100	573	16	13/3
	Inconsistent Commitment	89	882	17	15/2
Arguments Based on Cases	Cause to Effect	99	1,146	41	35/6
	Established Rule	95	1,008	3	1/2
	Verbal Classification	99	1,115	8	4/4
Defeasible Rule-based Arguments	Analogy	100	1,165	8	7/1
	Example	97	550	5	4/1
	Precedent	94	1,056	6	4/2
Discovery Arguments	Best Explanation	100	2,112	111	86/25
	Ignorance	93	1,122	5	3/2
	Random Sample to Population	0	0	2	1/1
	Sign	100	997	17	11/6
Popular Acceptance Arguments	Popular Opinion	99	1,096	10	5/5
	Popular Practice	94	1,066	5	4/1
Position to Know Arguments	Expert Opinion	100	1,195	16	15/1
	Position to Know	100	1,182	28	16/12
	Witness Testimony	100	2,178	9	3/6
Practical Reasoning Arguments	Consequences	0	0	34	34/0
	Practical Reasoning	0	0	63	48/15
	Sunk Costs	93	1,098	8	7/1
	Threat	88	1,520	18	17/1
	Waste	86	880	9	8/1
Chained Arguments with Rules and Cases	Slippery Slope	76	1,530	1	1/0
<b>Total</b>	-	1,902	23,471	441	331/100

# P1: Experiments

## Pre-training + Fine-tuning:



**Prompting:** Zero Shot (ZS), Few Shot (FS), and Few Shot Dialogue (FS-Dial) + Justification

# P1: Results

Model	Precision	Recall	F1-score
ROBERTA-AS-COMP	0.8	4.5	1.0
ROBERTA-AS-PROC	3.4	6.2	3.1
ROBERTA-AS-COMP-DIAL	7.4	9.4	8.0
ROBERTA-AS-PROC-DIAL	8.2	10.7	9.0
QWEN2.5(7B)-AS-ZS	4.1	14.3	5.7
LLAMA3.1(8B)-AS-ZS	9.9	8.9	6.6
LLAMA3.3(70B)-AS-ZS	18.9	24.4	18.7
QWEN2.5(7B)-AS-FS	3.7	11.0	5.5
LLAMA3.1(8B)-AS-FS	4.5	12.2	5.4
LLAMA3.3(70B)-AS-FS	<b>31.2</b>	<b>45.4</b>	<b>29.4</b>
QWEN2.5(7B)-AS-FS-DIAL	7.4	16.2	7.8
LLAMA3.1(8B)-AS-FS-DIAL	18.6	18.9	14.4
LLAMA3.3(70B)-AS-FS-DIAL	22.1	27.9	22.3
ROBERTA-AF-COMP	45.5	38.1	31.7
ROBERTA-AF-PROC	57.7	56.3	49.7
ROBERTA-AF-COMP-DIAL	<b>65.1</b>	47.7	49.3
ROBERTA-AF-PROC-DIAL	62.1	<b>66.9</b>	<b>62.3</b>
QWEN2.5(7B)-AF-ZS	12.3	26.1	13.5
LLAMA3.1(8B)-AF-ZS	13.1	21.3	13.9
LLAMA3.3(70B)-AF-ZS	44.4	44.2	34.7
QWEN2.5(7B)-AF-FS	8.3	20.8	11.0
LLAMA3.1(8B)-AF-FS	10.1	17.2	11.9
LLAMA3.3(70B)-AF-FS	18.7	34.4	23.8
QWEN2.5(7B)-AF-FS-DIAL	27.9	16.9	14.7
LLAMA3.1(8B)-AF-FS-DIAL	38.7	28.4	28.1
LLAMA3.3(70B)-AF-FS-DIAL	34.7	36.8	31.8

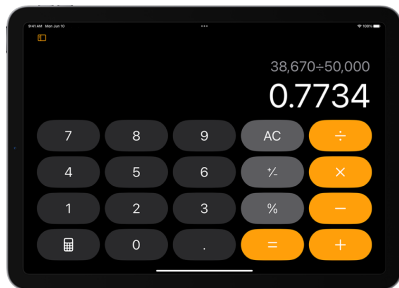


- LLMs struggle to effectively generalise regardless of the dimensionality of the task.
- Justifications reveal that LLMs are not able to process the inferential reasoning of arguments, either referring to premises/claims not existent in the argument, or quoting wrong parts of the argument.
- Textbook-like natural language argumentation schemes + theory-based pre-processing (enthymemes) and pre-training + fine-tuning on natural language dialogue data, **even for low resource tasks!**

## P2: Background and Motivation

### Natural Language Reasoning in Large Language Models: Analysis and Evaluation

Reasoning in Natural Language → Natural Language Reasoning



Argument:

- Argument:  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$
- Relation:  $\mathcal{R} = \{\vdash, \dashv\}, \mathcal{R} : A \times A$

Argument-component selection:

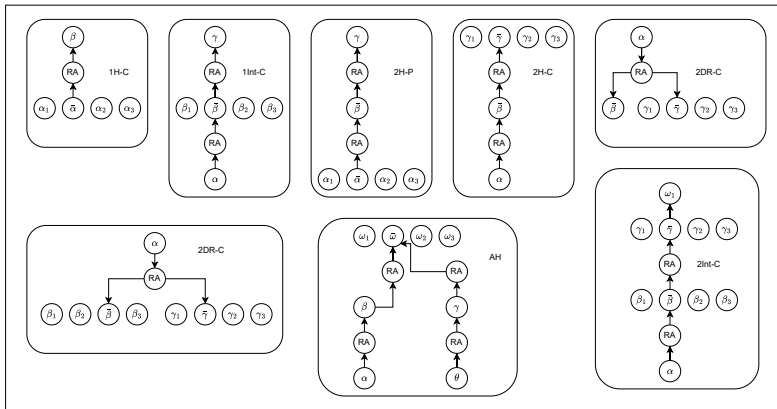
Given an argument  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$  missing a component  $a_i$  within a context  $\mathcal{C}$ , find the correct argument-component  $\hat{u}$  from a given set of candidates  $\mathcal{U} = \{u_1, u_2, \dots, u_k\}$ , belonging to  $\mathcal{C}$ , such that  $\hat{u}$  corresponds to  $a_i$ .



## P2: Argumentative Reasoning Tasks (ART) ARG-tech Centre for Argument Technology

- Serial Reasoning:
  1. One-hop Conclusion
  2. One-hop Premise
  3. Two-hop Conclusion
  4. Two-hop Premise
  5. One-Intermediate Conclusion
  6. Two-Intermediate Conclusions
- Linked Reasoning:
  1. One Linked Premise
  2. Two Linked Premises
  3. Linked Reasoning Conclusion
- Convergent Reasoning:
  1. One Convergent Premise
  2. Two Convergent Premises
  3. Convergent Reasoning Conclusion
  4. Alternative Hop
- Divergent Reasoning:
  1. One Divergent Reasoning Conclusion
  2. Two Divergent Reasoning Conclusions
  3. Divergent Reasoning Premise

# P2: Argumentative Reasoning Tasks (ART) ARG-tech Centre for Argument Technology



# P2: Argumentative Reasoning Tasks (ART)

112,212 Multiple-choice questions, 16 different reasoning structures, 7 different corpora.

Tasks		MTC	AAEC	CDCP	ACSP	AbstRCT	US2016	QT30
Type	Variants							
Serial	<b>1H-C</b>	290	4841	1033	5789	2288	3379	6488
	<b>1H-P</b>	290	4841	1033	5789	2288	3379	6488
	<b>2H-C</b>	57	3279	348	759	327	1009	1118
	<b>2H-P</b>	57	3279	348	759	327	1009	1118
	<b>Int-C</b>	57	3279	348	759	327	1009	1118
	<b>2-Int-C</b>	3	569	89	80	8	249	787
Linked	<b>1L-P</b>	17	-	64	-	-	180	511
	<b>2L-P</b>	17	-	64	-	-	180	511
	<b>LR-C</b>	17	-	64	-	-	180	511
Convergent	<b>1C-P</b>	96	4735	763	2024	1899	1129	397
	<b>2C-P</b>	96	4735	763	2024	1899	1129	397
	<b>CR-C</b>	96	4735	763	2024	1899	1129	397
	<b>AH</b>	57	3279	348	759	327	1009	1118
Divergent	<b>1DR-C</b>	-	-	11	184	48	106	386
	<b>2DR-C</b>	-	-	11	184	48	106	386
	<b>DR-P</b>	-	-	11	184	48	106	386

# P2: Results

Dataset	Model	Size	Argument-Component Selection			
			Serial	Linked	Convergent	Divergent
AAEC	Owen 2.5	7B	23.78 ± 13.52	-	10.85 ± 11.50	-
		72B	35.59 ± 13.49	-	18.95 ± 19.37	-
	Llama 3.1	8B	12.23 ± 9.87	-	4.15 ± 3.62	-
		70B	38.77 ± 8.12	-	16.08 ± 20.25	-
	Mistral	7B	29.82 ± 14.12	-	10.4 ± 13.46	-
	DeepSeek-R1	70B	46.75±16.65	-	33.91±18.23	-
GPT	GPT-4o	49.83 ± 17.37	-	35.78 ± 21.50	-	
MTC	Owen 2.5	7B	0.2 ± 0.21	-	1.75 ± 2.04	-
		72B	19.51 ± 16.29	-	2.6 ± 3.40	-
	Llama 3.1	8B	0.16 ± 0.16	-	1.05 ± 1.50	-
		70B	8.53 ± 11.71	-	5.46 ± 4.56	-
	Mistral	7B	0.16 ± 0.26	-	0.9 ± 1.53	-
	DeepSeek-R1	70B	45.34±10.45	-	15.87±12.34	-
GPT	GPT-4o	49.73 ± 24.36	-	11.36 ± 11.54	-	
CDCP	Owen 2.5	7B	29.97 ± 14.84	35.38 ± 25.32	17.45 ± 20.45	0.86 ± 0.90
		72B	50.28 ± 21.52	51.28 ± 16.59	24.68 ± 28.54	1.2 ± 0.61
	Llama 3.1	8B	10.33 ± 7.95	9.23 ± 12.21	5.85 ± 6.66	0.4 ± 0.4
		70B	40.71 ± 17.94	49.74 ± 21.88	21.47 ± 28.40	0.93 ± 0.53
	Mistral	7B	22.97 ± 12.18	12.82 ± 14.94	8.85 ± 12.64	0.26 ± 0.46
	DeepSeek-R1	70B	61.85±9.78	83.43±12.22	41.56±24.23	7.12±3.44
GPT	GPT-4o	65.06 ± 13.41	68.87 ± 14.93	44.94 ± 30.31	7.33 ± 2.52	
AbstRCT	Owen 2.5	7B	11.46 ± 6.28	-	14.4 ± 18.73	0.933 ± 0.90
		72B	33.96 ± 19.27	-	29.40 ± 33.71	1.46 ± .070
	Llama 3.1	8B	4.7 ± 3.30	-	8.9 ± 7.01	0.4 ± 0.4
		70B	19.05±19	-	11.12.86	1.33 ± 0.80
	Mistral	7B	10.0 ± 5.77	-	6.35 ± 9.19	0.33 ± 0.41
	DeepSeek-R1	70B	46.34±23.45	-	36.56±25.67	10.45±5.56
GPT	GPT-4o	48.61 ± 28.90	-	34.48 ± 29.19	11.4 ± 3.13	

Dataset	Model	Size	Argument-Component Selection			
			Serial	Linked	Convergent	Divergent
ACSP	Owen 2.5	7B	37.13 ± 19.03	-	16.05 ± 15.38	9.13 ± 6.77
		72B	47.31 ± 23.55	-	25.07 ± 15.23	12.8 ± 6.43
	Llama 3.1	8B	12.3 ± 8.25	-	4.5 ± 4.94	2.4 ± 2.42
		70B	39.64 ± 13.76	-	12.433 ± 18.07	8.86 ± 6.10
	Mistral	7B	26.66 ± 13.47	-	12.4 ± 14.18	5.86 ± 5.08
	DeepSeek-R1	70B	56.78±10.43	-	51.45±9.56	24.78±5.23
GPT	GPT-4o	90.47 ± 7.34	-	86.38 ± 3.16	41.45 ± 14.34	
US2016	Owen 2.5	7B	34.12 ± 19.53	30.55 ± 19.37	20.45 ± 21.46	7.6 ± 5.4
		72B	49.53 ± 27.61	48.33 ± 18.86	30.34 ± 25.69	10.53 ± 6.26
	Llama 3.1	8B	14.41 ± 6.34	11.66 ± 8.67	9.9 ± 12.58	2.86 ± 2.71
		70B	45.51 ± 25.65	45.18 ± 21.37	26.59 ± 26.64	8.06 ± 5.98
	Mistral	7B	37.95 ± 20.51	20.18 ± 17.84	12.8 ± 15.21	4.53 ± 3.70
	DeepSeek-R1	70B	60.34±13.45	47.65±15.34	41.95±13.72	36±14.63
GPT	GPT-4o	58.47 ± 12.94	53.03 ± 9.32	45.85 ± 15.12	37.78 ± 17.21	
QT30	Owen 2.5	7B	31.40±16.96	20.76 ± 18.63	11.4 ± 11.10	20 ± 18.11
		72B	42.45 ± 20.84	45.50 ± 16.24	20.33 ± 17.02	29.0 ± 15.77
	Llama 3.1	8B	9.99 ± 5.15	11.50 ± 10.26	5.8 ± 4.48	12.33 ± 13.52
		70B	36.21 ± 15.94	43.38 ± 20.84	18.10 ± 16.59	23.16 ± 17.40
	Mistral	7B	33.98 ± 17.96	20.76 ± 18.63	6.2 ± 8.22	12.4 ± 11.78
	DeepSeek-R1	70B	55.78±20.35	46.56±14.47	40.92±18.43	38.21±11.45
GPT	GPT-4o	53.62 ± 23.80	53.04 ± 18.66	46.69 ± 21.44	41.65 ± 15.34	



## Open-Ended Reasoning with Human Evaluation:

Macro F1-score (GPT-4o): 25.8

→ Instead of generating the new components, the model copy-pasted or concatenated argument components from the context.

## Model Size:

Llama 3.1		GPT	
70B	405B	gpt-4o	o1-preview
9.98	18.73	32.18	41.96

## Prompt Template:

Model	Prompt-1	Prompt-2
Llama 3.1:70B	16.01	15.40
Mistral	7.25	7.09
Qwen 2.5:72B	16.29	14.61
GPT-4o	34.32	35.78

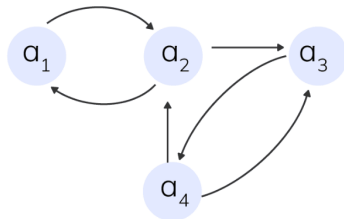
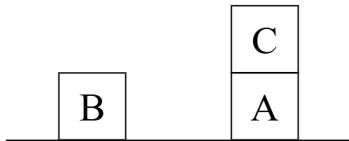
- LLMs rely on **superficial patterns of language** rather than genuine reasoning.
- LLMs are **not capable of understanding argumentative reasoning** structures, in cases where a slightly more challenging argumentative structure is used, they perform worse than a random baseline.
- Need to develop **challenging tasks to evaluate natural language reasoning**.

# Towards Natural Language Reasoning LMs

---

# What can we do?

Probabilistic language modelling has many advantages, but we cannot forget about **theory-informed and symbolic modelling**, especially for tasks that involve reasoning (e.g., planning, argumentation).



# What can we do?

Language “Reasoning” Models (LRMs) are never evaluated on their reasoning capabilities but on the correctness of the final answers. When looking with enough detail into the reasoning traces we can find surprising stuff.

high). So the chain is: cap exists (premise) → when the amounts are high (capped), tenants think twice (conclusion1) → but if the cap is set to a small amount, then the increases aren't bad (conclusion2). But the original statement starts by saying rent increases aren't bad if small. So the argument structure might be supporting that main point. Therefore, the correct answer is option4.

5

Looks like reasoning, but **is it actually reasoning?**

# What can we do?

1. Evaluate directly the reasoning capabilities of LLMs (or LRMs)  
→ New Benchmarks.
2. Embed natural language reasoning as part of the training process.  
→ New Architectures (neurosymbolic?).
3. Post-process the generated output to improve its soundness.  
→ Reasoning assisted generation.

## The 12th Workshop on Argument Mining

July 31st, 2025

Co-located with [ACL 2025](#) in Vienna, Austria

<https://argmining-org.github.io/2025/>

### COMMITTEE

#### Organizing Committee



Elena Chistova



Philipp Cimiano



Shohreh Haddadan



Gabriella Lapesa



Ramon Ruiz-Dolz



## IMPORTANT DATES

- Workshop: **July 31st, 2025**
- Direct paper submission due ([OpenReview](#)): **April 17th, 2025**
- Commitment deadline for ARR papers ([OpenReview](#)): **May 21st, 2025**
- Notification of acceptance: **May 28th, 2025**
- Camera-ready papers due: **June 4th, 2025**

All deadlines are 11.59 pm UTC -12h ("anywhere on Earth").

## KEYNOTE SPEAKER



[Andreas Vlachos](#) Department of Computer Science and Technology, University of Cambridge  
Title: *TBA*

**About the Speaker:** [Andreas Vlachos](#) is a professor of NLP and Machine Learning at the University of Cambridge. Among the many things he has worked on, we find constructiveness in argumentation, fact checking, media bias, dialogue modeling.



## SHARED TASKS

The 12th Argument Mining Workshop will be hosting two shared tasks.

### 1. Critical Questions Generation Shared Task

**Organizers** : Blanca Calvo Figueras, Rodrigo Agerri , HITZ Basque Center for Language Technology - Ixa, University of the Basque Country UPV/EHU, Spain,  
Elena Cabrio, Serena Villata, University of Côte d'Azur and member of the Inria-I3S research team Wimmics

### 2. MM-ArgFallacy2025: Multimodal Argumentative Fallacy Detection and Classification on Political Debates

**Organizers**: Eleonora Mancini, Federico Ruggeri, Paolo Torrioni, Language Technologies Lab, University of Bologna, Italy  
Serena Villata, Inria-I3S WIMMICS Laboratoire I3S, CNRS, Sophia Antipolis, France

# Thank you very much!

To keep in touch:

- Bluesky: @raruidol
- LinkedIn: Ramon Ruiz-Dolz & ARG-tech