Syllogistic Reasoning:
a suitable test bed to evaluate LLMs' ability to **abstract form** from **content**

Raffaella Bernardi

Free University of Bozen Bolzano

LLM’s reasoning ability

spectrum of perspectives

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

Impact of Pretraining Term Frequencies on Few-Shot Reasoning

Yasaman Razeghi [1] Robert L. Logan IV [1] Matt Gardner [2] Sameer Singh [1 3]

Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam*
School of Computing & AI
Arizona State University, Tempe.
kvalmeek@asu.edu

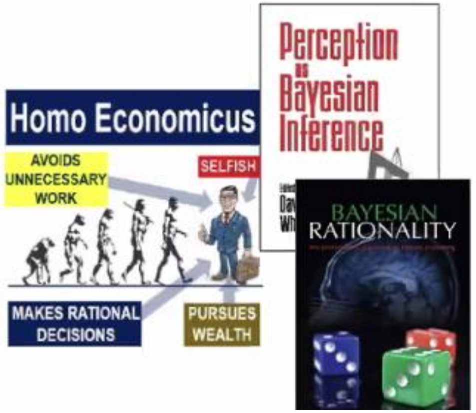
Alberto Olmo*
School of Computing & AI
Arizona State University, Tempe.
aolmo@asu.edu

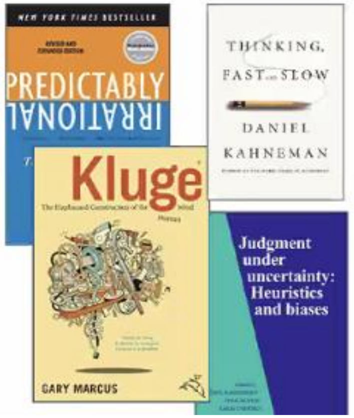
Sarath Sreedharan †
Department of Computer Science,
Colorado State University, Fort Collins.
sarath.sreedharan@colostate.edu

Subbarao Kambhampati
School of Computing & AI
Arizona State University, Tempe.
rao@asu.edu

LLMs are...

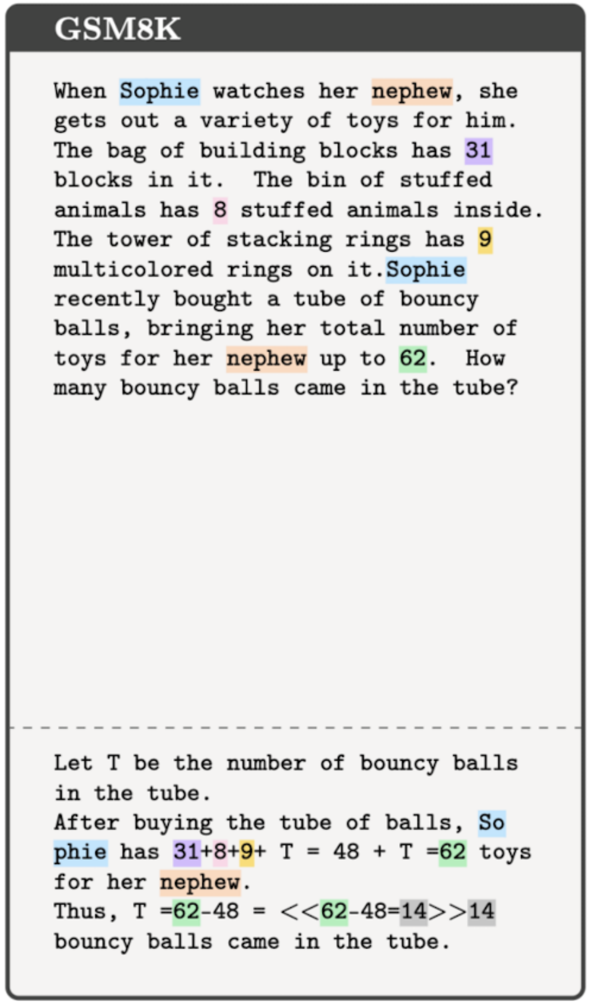
... capable of reasoning ⟷ **... just a pile of tricks**





Credits A. **Lampinen**

Benchmarks to evaluate LLMs reasoning ability



Grade School Math: 2021

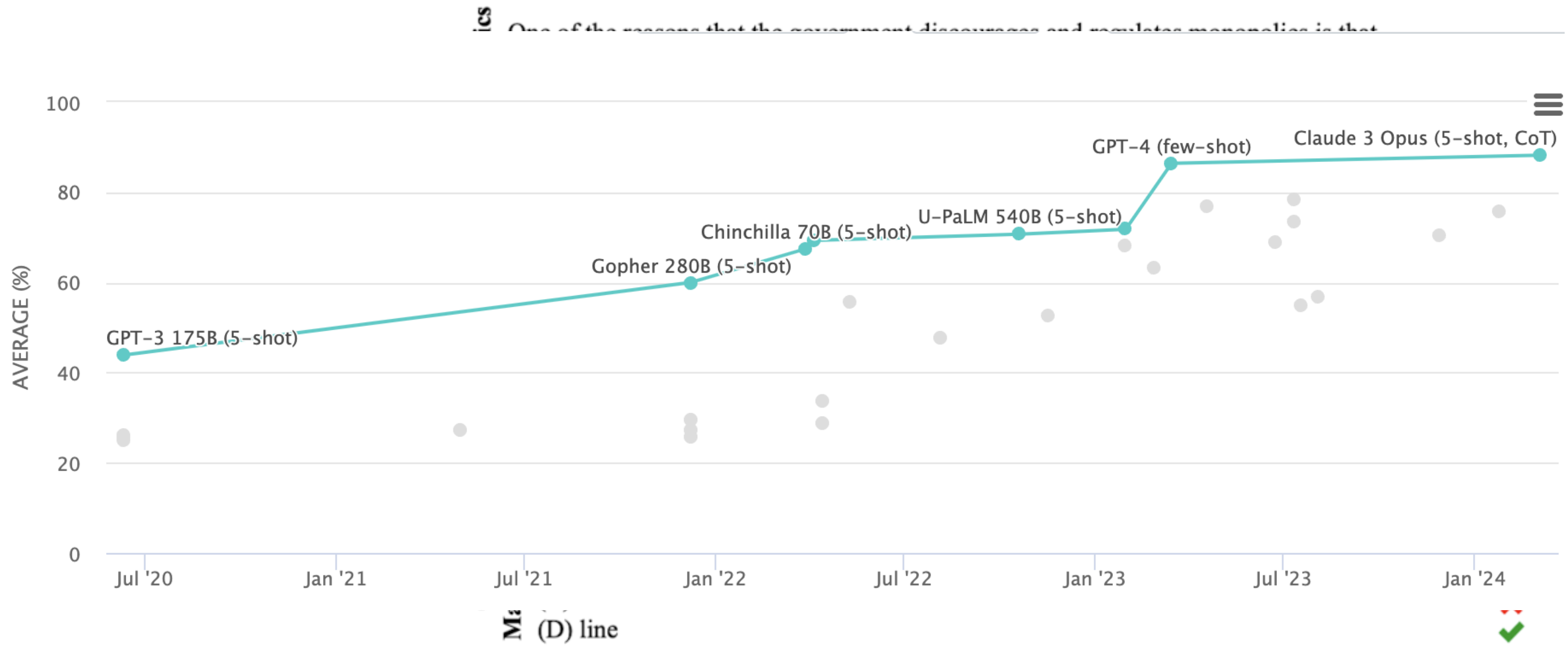


Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

MMLU 2021

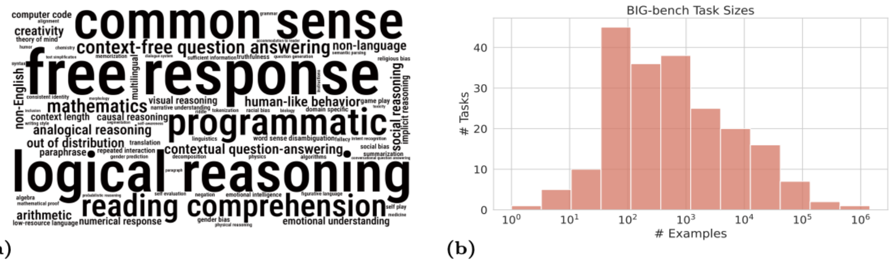


Figure 3: **Diversity and scale of BIG-bench tasks.** **(a)** A word-cloud of task keywords. **(b)** The size distribution of tasks as measured by number of examples.

Big Bench 2023

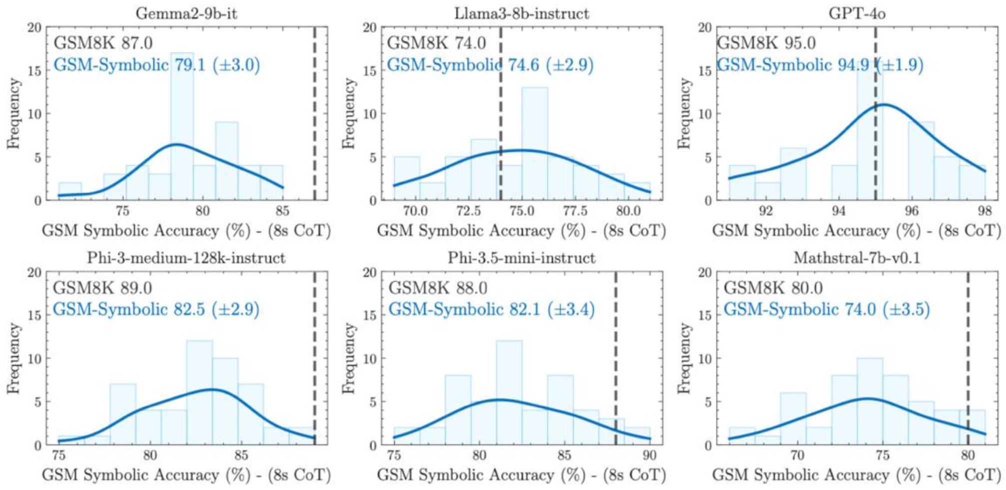


Figure 2: The distribution of 8-shot Chain-of-Thought (CoT) performance across 50 sets generated from `GSM-Symbolic` templates shows significant variability in accuracy among all state-of-the-art models. Furthermore, for most models, the average performance on `GSM-Symbolic` is lower than on `GSM8K` (indicated by the dashed line). Interestingly, the performance of `GSM8K` falls on the right side of the distribution, which, statistically speaking, should have a very low likelihood, given that `GSM8K` is basically a single draw from `GSM-Symbolic`.

GSM-Symbolic, Apple 2024

Lack generalization

reasoning capabilities of models.Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.

The LLM Reasoning Debate Heats Up

Three recent papers examine the robustness of reasoning and problem-solving in large language models

One
can

In M
audi
whic
on n

Conclusion

In conclusion, there's no consensus about the conclusion! There are a lot of papers out there demonstrating what looks like sophisticated reasoning behavior in LLMs, but there's also a lot of evidence that these LLMs aren't reasoning *abstractly* or *robustly*, and often over-rely on memorized patterns in their training data, leading to errors on "out of distribution" problems. Whether this is going to doom approaches like OpenAI's o1, which was directly trained on people's reasoning traces, remains to be seen. In the meantime, I think this kind of debate is actually really good for the science of LLMs, since it spotlights the need for careful, controlled experiments to test robustness—experiments that go far beyond just reporting accuracy—and it also deepens the discussion of what *reasoning* actually consists of, in humans as well as machines.

Trends in
Cognitive Sciences

CellPress

Feature Review

Dissociating language and thought in large language models

Kyle Mahowald,[1,5,*] Anna A. Ivanova,[2,5,*] Idan A. Blank,[3,*] Nancy Kanwisher,[4,*] Joshua B. Tenenbaum,[4,*] and Evelina Fedorenko[4,*]

➔ LLMs should be taken seriously as models of **formal linguistic skills**

➔ Models that master real-like language use would need to incorporate or develop not only a core language module**, but also multiple non-language-specific cognitive capacities required for modelling thought.**

In 2023, several papers on LLMs and reasoning strength and weakness.

Investigate the **deductive reasoning** capabilities of LLMs.

A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences

Leonardo Bertolazzi,
DISI, University of Trento
leonardo.bertolazzi@unitn.it

Albert Gatt,
ICS, Utrecht University
a.gatt@uu.nl

Raffaella Bernardi
CIMeC and DISI, University of Trento
raffaella.bernardi@unitn.it

EMNLP 2024

Investigate **systematic generalization** for logical reasoning in LLMs

A MIND for Reasoning: Meta-learning for In-context Deduction

Leonardo Bertolazzi[1], Manuel Vargas Guzmán[2], Raffaella Bernardi[3], Maciej Malicki[2], Jakub Szymanik[1],

[1]University of Trento, [2]University of Warsaw, [3]Free University of Bozen-Bolzano

Syllogisms as a test bed for formal reasoning

moods

affirmative	negative
A: All a are b	**E**: No a are b
I: Some a are b	**O**: Some a are not b

figures

	1	2	3	4
P1:	a-b	b-a	a-b	b-a
P2:	b-c	c-b	c-b	b-c

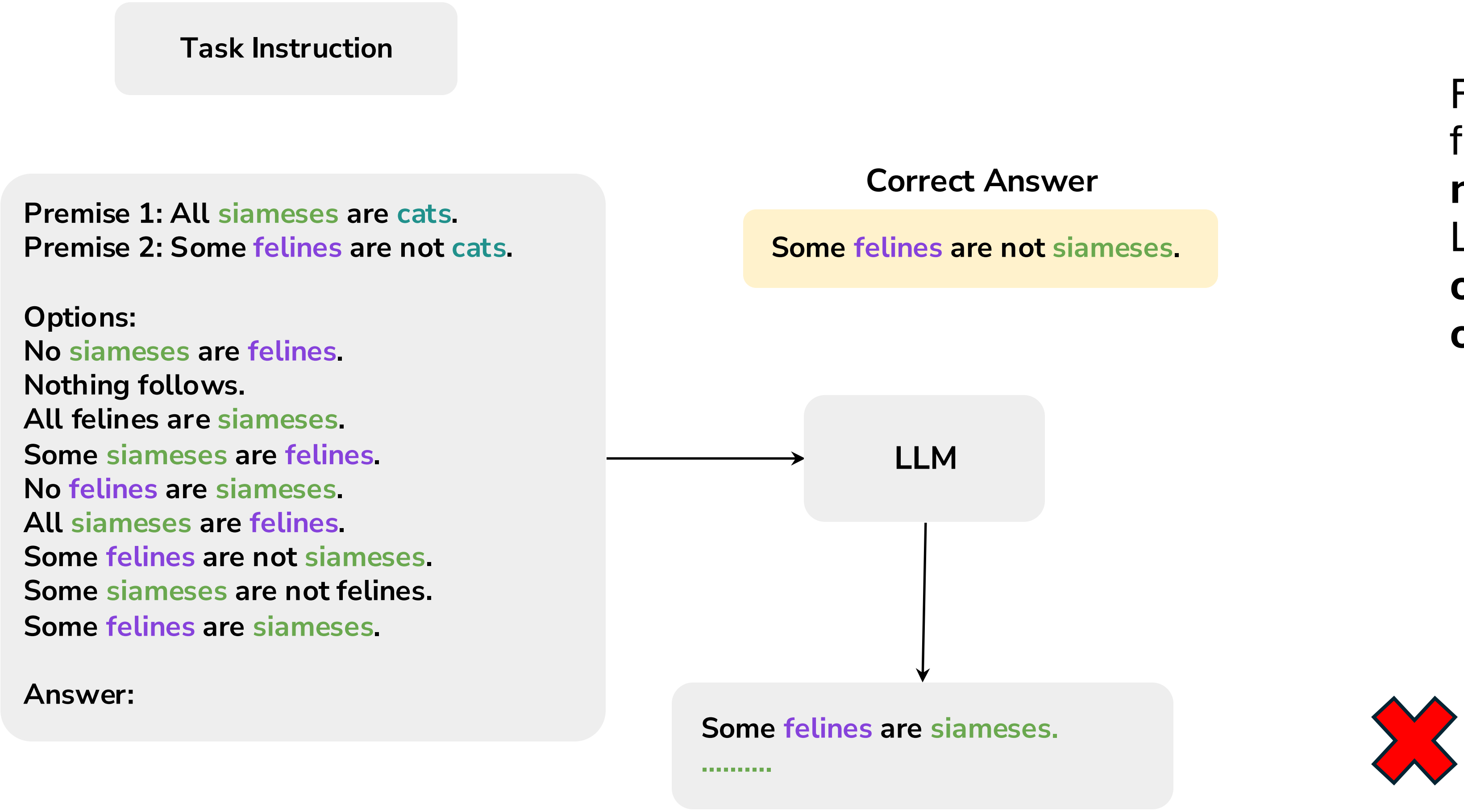
P1: All siameses are **cats**
P2: Some felines are not **cats**
C: Some felines are not siameses

Schema: **AO3**
P1: All a are **b** (A)
P2: Some c are not **b** (O)
C: Some c are not a

Syllogisms an ideal test bed for a deep examination of reasoning capabilities:

- Fixed inferential patterns (64 schemas)
- Some sets of premises admit conclusions (valid) and some do not (invalid)
- We have evidence on how humans solve them in practice → cognitive psychology
- We have an abstract model of how they can be solved → predicate logic

Multiple choice syllogisms completion



Following Eisape et al. (2024), we frame syllogistic inferences as a **multiple-choice task**, where a LLM is tasked with generating **one or more of the provided options**.

LLMs do not treat syllogisms formally

Syllogism EO1

P1: No dogs are felines.
P2: Some felines are not cats.

C: Nothing follows

LLMs tend to avoid selecting the option "nothing follows" (Eisape et al., 2024).

Syllogism AO3

P1: All canines are dogs.
P2: Some labradors are not dogs.

C: Some labradors are not canines.

LLMs are sensitive to the content of conclusions and are less accurate in selecting the correct ones if those **conclusions conflict with world knowledge** (***content effect bias***) (Lampinen et al., 2024).

Syllogism IA1

P1: Some cycluirts are schmeeft.
P2: All schmeeft are szeiag.
P3: All szeiag are steaugs.

C: Some cycluirts are steaugs or some steaugs are cycluirts.

LLMs struggle to generalize inferences to **longer sets of premises** than those encountered during training (Clark et al., 2020).

Datasets: Semantic content

We create datasets that control for semantic content and developed **two datasets** which share the same vocabulary but differ in the believability of their conclusions.

BELIEVABLE

Premise 1: All labradors are dogs.
Premise 2: Some canines are not dogs.
Conclusion: Some canines are not labradors.

→ **True Conclusion**

UNBELIEVABLE

Premise 1: All canines are dogs.
Premise 2: Some labradors are not dogs.
Conclusion: Some labradors are not canines.

→ **False Conclusion**

Datasets: inference complexity

For **inference complexity**, we created three datasets using **pseudo-words**, each **differing in the length** of the syllogism. The same type of conclusion is drawn, but from a varying number of premises:

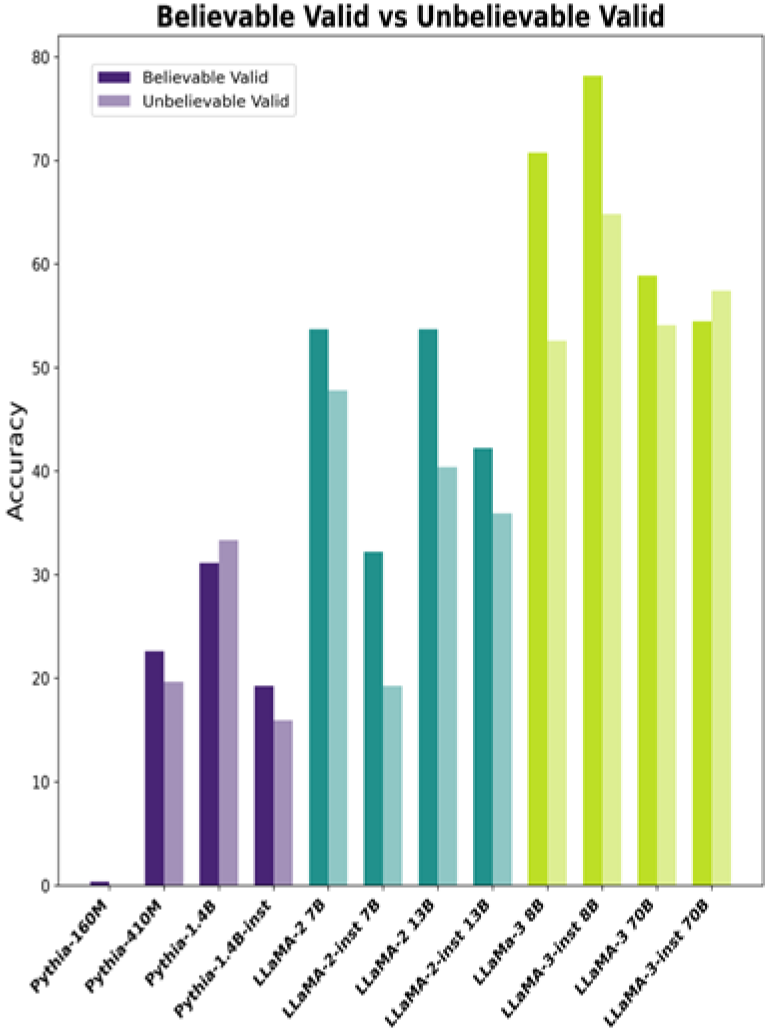
Premise 1: No tuem are graibly.
Premise 2: All graibly are kwaitz.
Conclusion: Some kwaitz are not tuem.

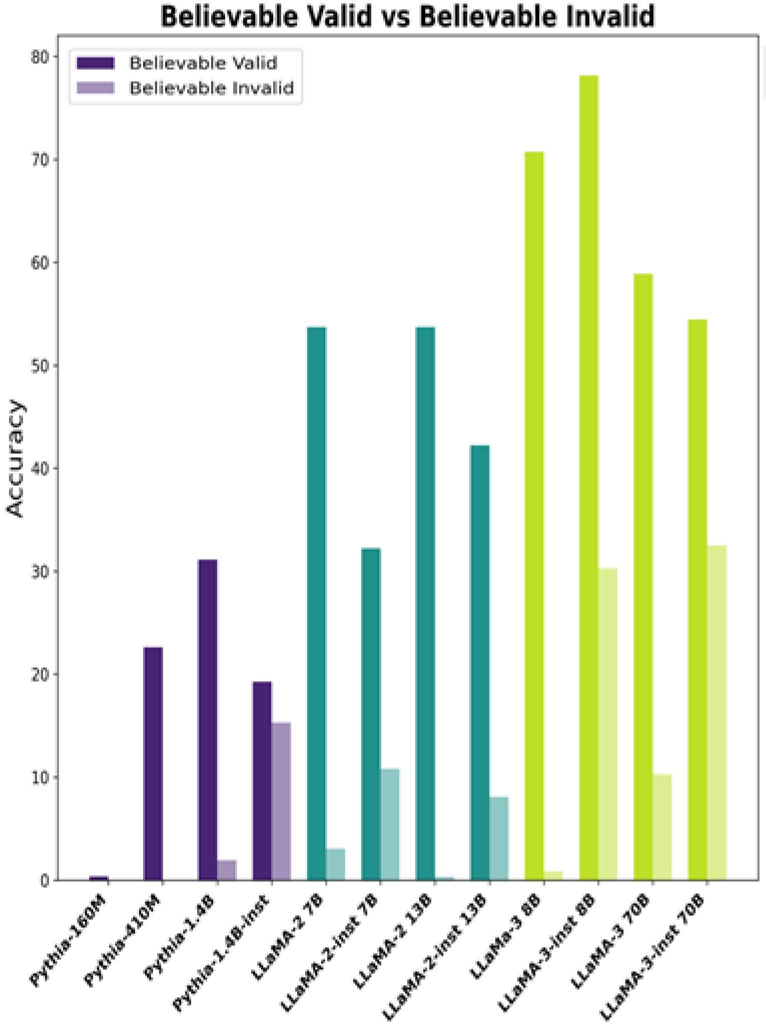
Premise 1: No khuipt are gnauntly.
Premise 2: All gnauntly are skaiank.
Premise 3: All skaiank are synulls.
Conclusion: Some synulls are not khuipt.

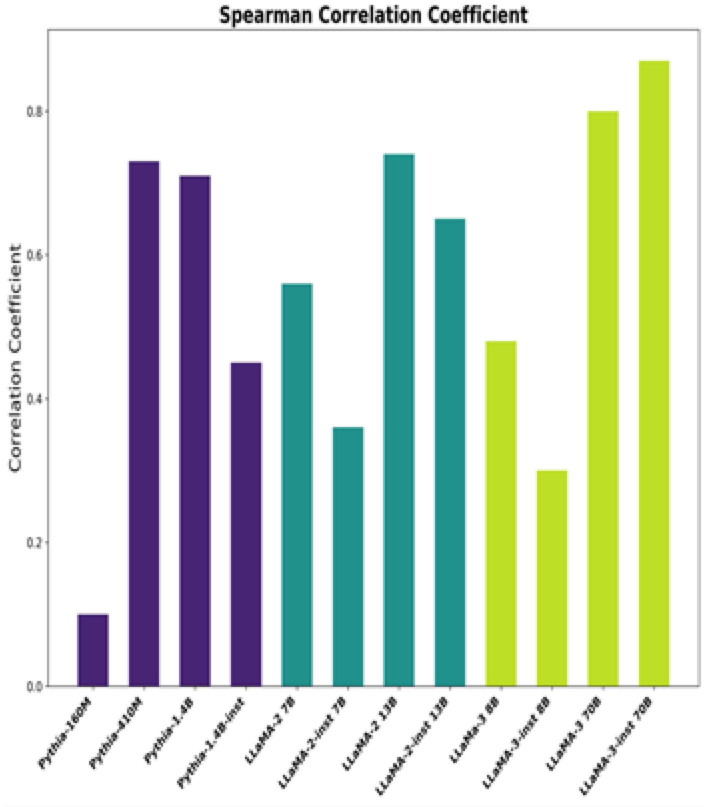
Premise 1: No screarm are pruerf.
Premise 2: All pruerf are thaon.
Premise 3: All thaon are mcnient.
Premise 4: All mcnient are tsiorm.
Conclusion: Some tsiorm are not screarm.

Zero-shot CoT evaluation

Models from the Pythia, LLaMA-2, and LLaMA-3 families.







Human data from: Khemlani and Johnson-Laird 2012

Experimental set up

RQ: are these biases mitigated by in-context learning (ICL) or supervised finetuning (SFT)?

Syllogism AO3

Premise 1: All siameses are cats.
Premise 2: Some felines are not cats.

Options:
No siameses are felines.
Nothing follows.
All felines are siameses.
Some siameses are felines.
No felines are siameses.
All siameses are felines.
Some felines are not siameses.
Some siameses are not felines.
Some felines are siameses.

Answer:

Correct Answer

Some felines are not siameses.

Zero-shot CoT

Task Instruction

Premises
+
Options

LLM

Let's think step by step. If we know that all siameses are cats and we also know that some felines are not cats, we can conclude that some felines are not siamese. Therefore my final answer is: Some felines are not siameses.

ICL

Task Instruction

5 in-context examples

Premises
+
Options

LLM

Some felines are not siameses.
Nothing follows.

SFT

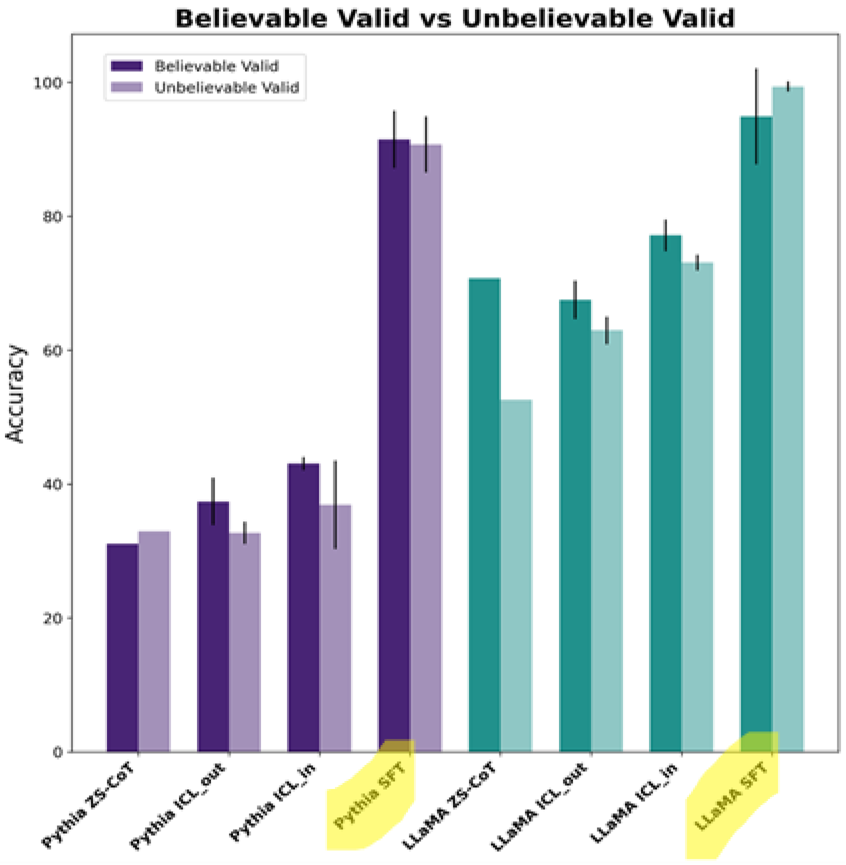
Premises
+
Options

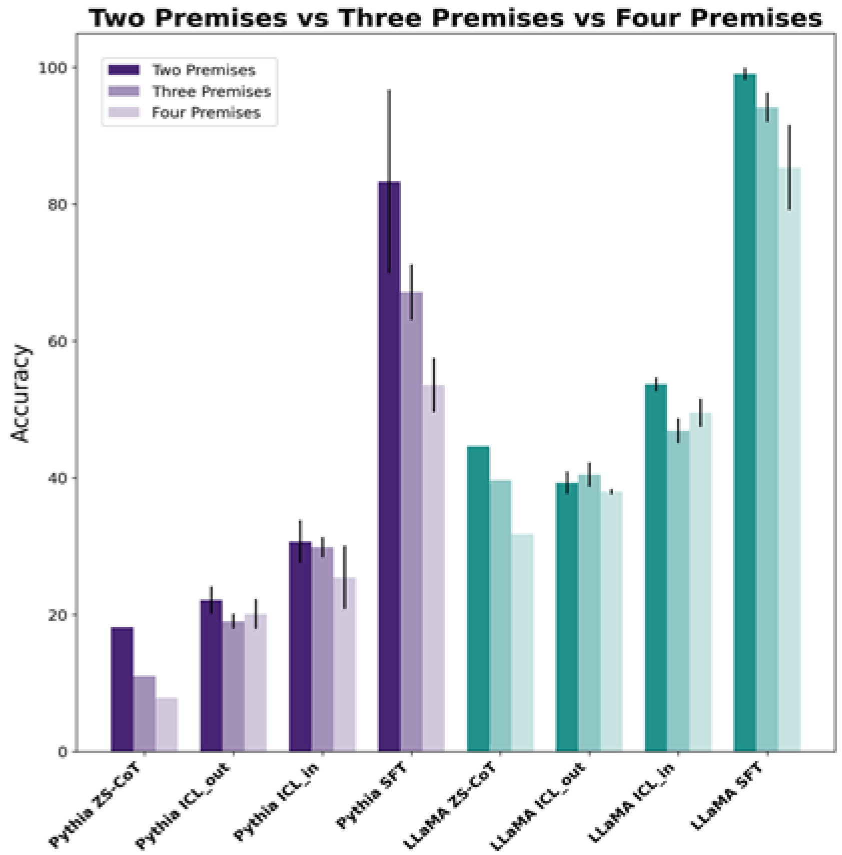
LLM

Some felines are not siameses.
Some felines are siameses

ICL examples/SFT training: pseudowords

Impact on ZS-CoT vs. ICL vs. SFT I

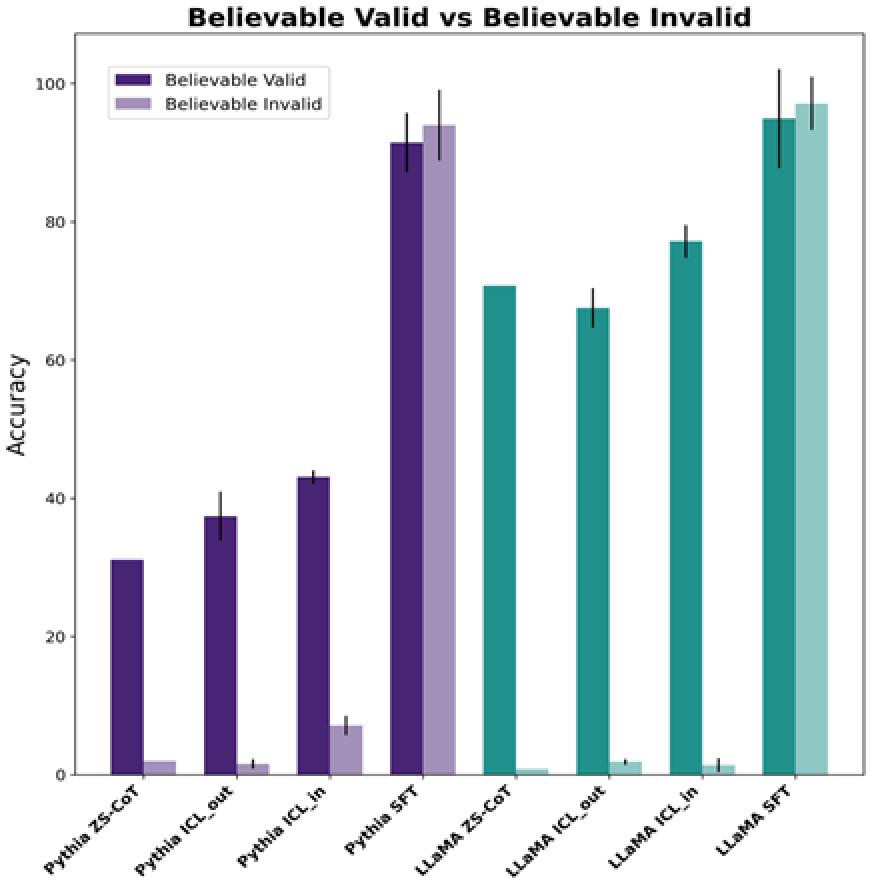


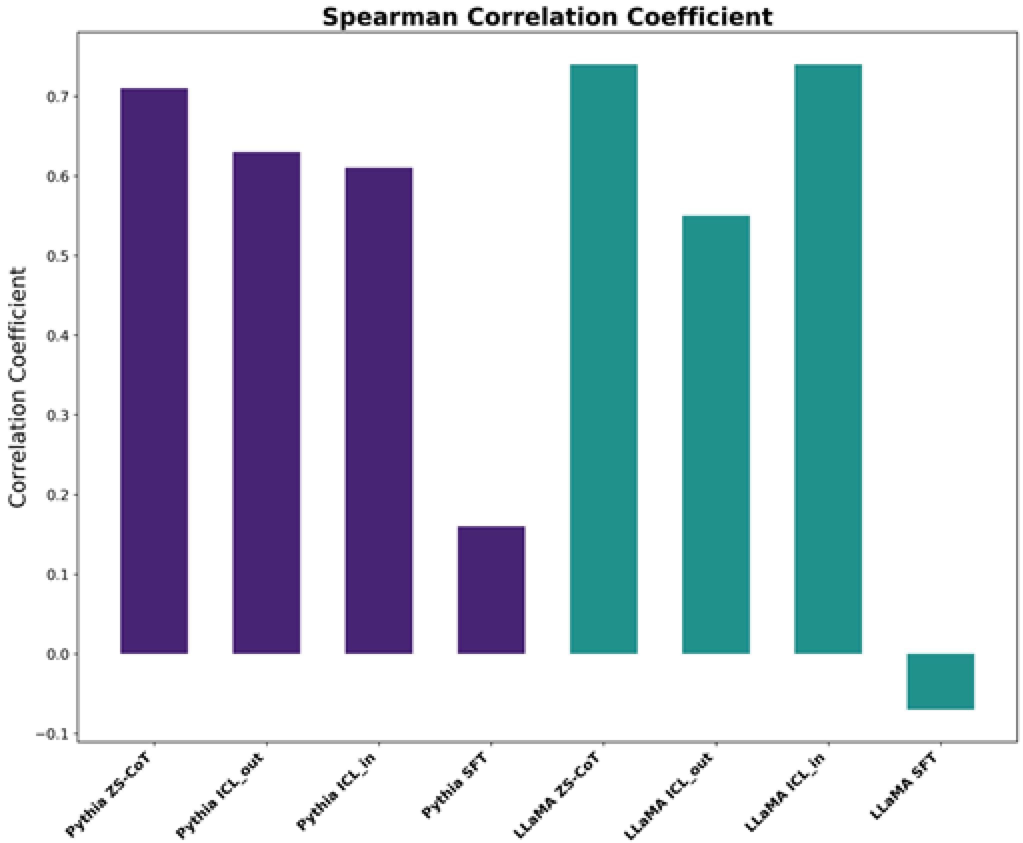


Content bias is reduced by ICL, but is only fully eliminated in SFT, where the model is exposed to many examples of the same inference with varying content.

Inference complexity affects all settings, but the performance drop is less pronounced with ICL compared to SFT.

Impact on ZS-CoT vs. ICL vs. SFT II

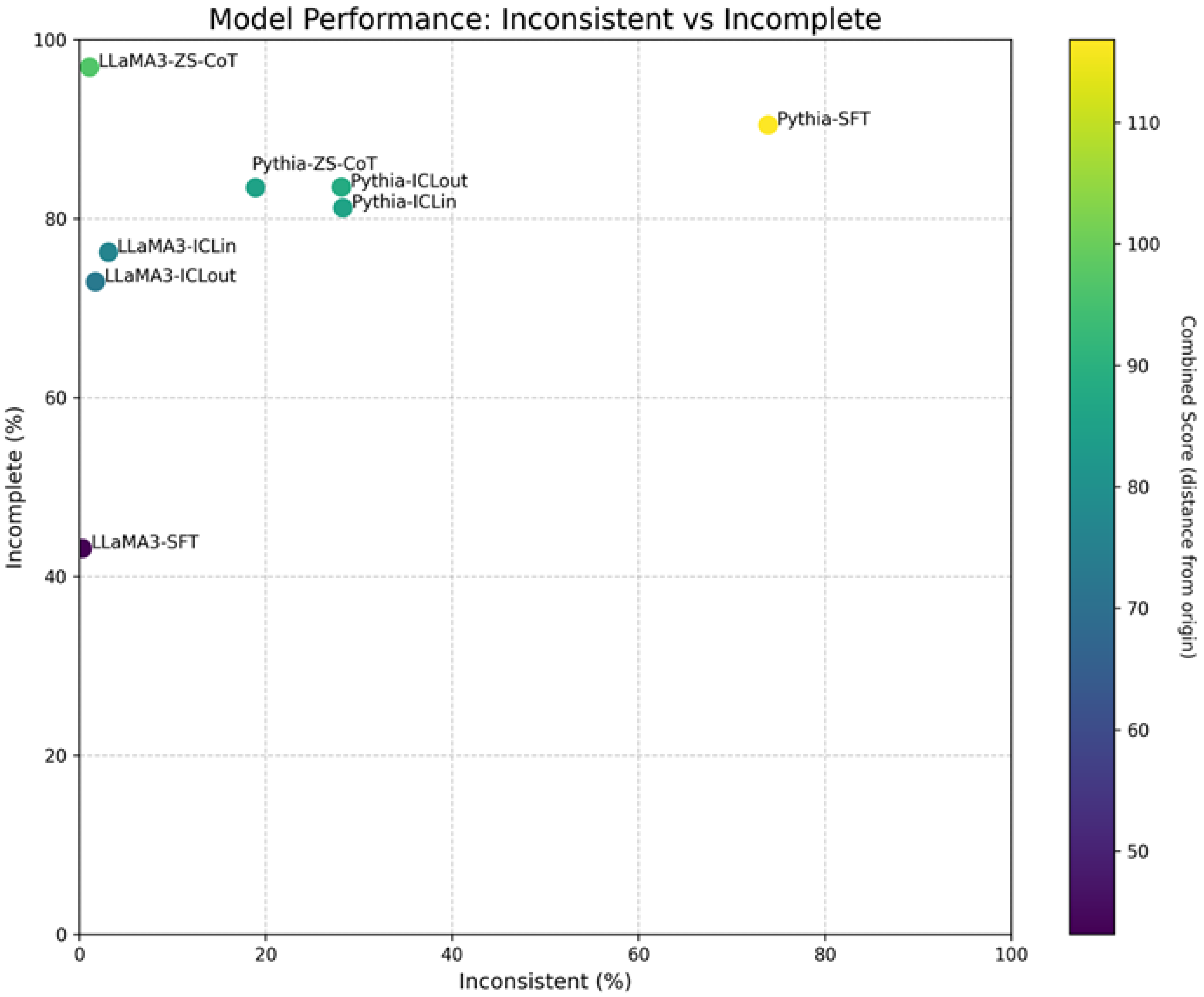




"Nothing follows" bias persists in ICL and disappears with SFT

Correlation with humans: SFT shows less alignment with humans

Consistent and Complete answers



If an agent is reasoning “formally” its answers should not just be accurate but also satisfy certain constraints:

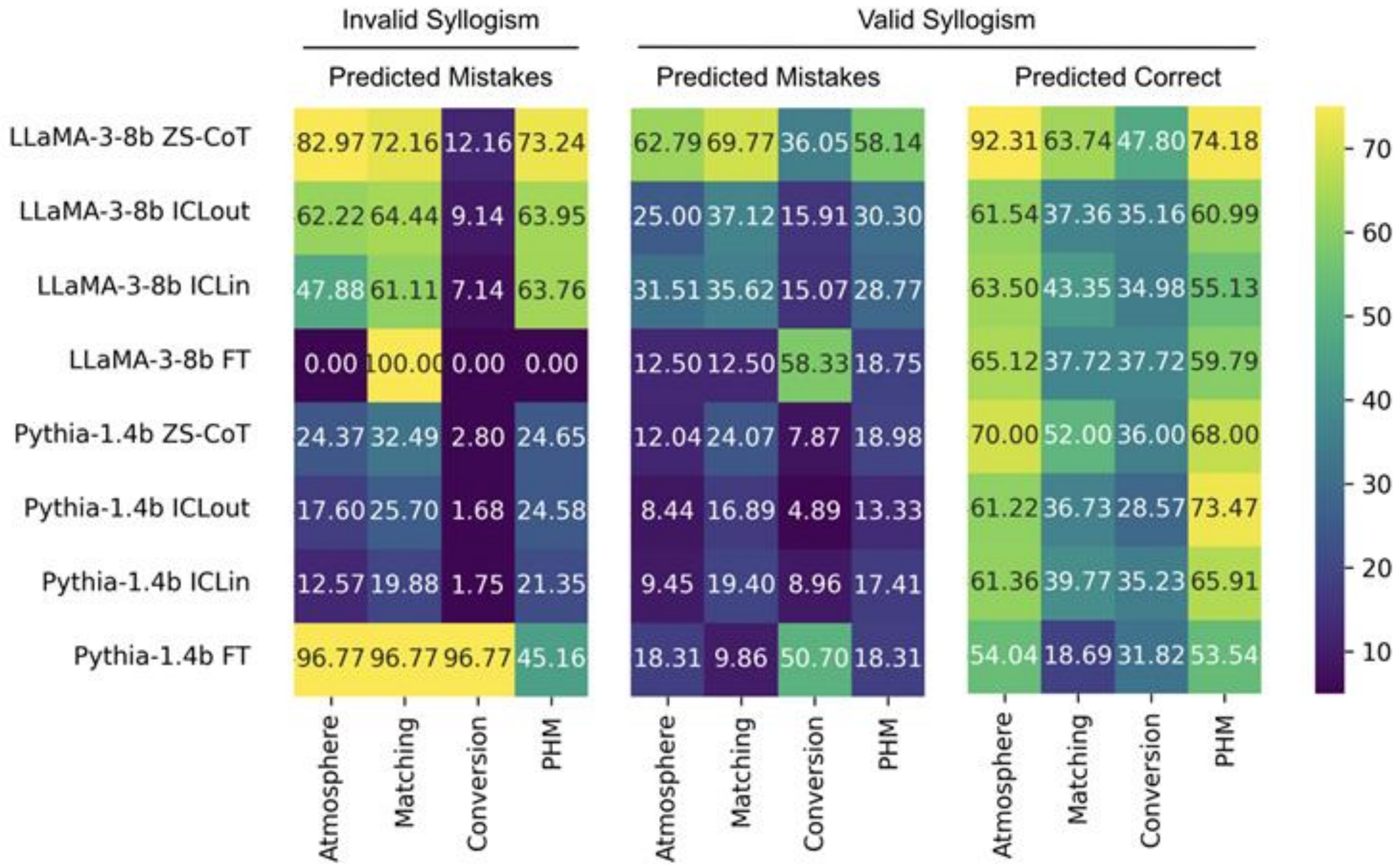
Consistency: the agent should not derive logically contradictory answers

Completeness: all logically equivalent answers should be inferred

Why do models avoid “Nothing follows” responses?

Models that demonstrate good accuracy cannot be considered capable of formal reasoning if their predictions can be mapped to those of simpler models based on **shortcuts**

We found that the behavior of LLaMA ZS-CoT is strongly predicted by the **atmosphere heuristic**. A model that has learned such a heuristic would never predict “nothing follows” conclusions, similar to observations made with other LLMs



Data from: Khemlani and Johnson-Laird 2012

Conclusion

- The strong alignment between LLaMA-3 8B's ZS-CoT behavior and the **atmosphere heuristic** suggests a reason for why Zero-Shot LLMs rarely produce "nothing follows" responses. We hypothesize **that they rely on a shallow pattern-matching strategy, using quantifiers as cues.**

- **ICL** enhances model performance on valid inferences, but it **does not eliminate content effects** or the challenge of handling invalid syllogisms. Most significantly, it increases model inconsistency.

- **SFT** on syllogisms with varying content is effective for both small- and medium-sized models, **eliminating content bias** and the tendency to avoid "nothing follows" answers. However, SFT does not always improve models in terms of completeness and consistency**.**

The models still fall short of the behavior expected from a purely formal reasoner:

➔ **they do not generalize systematically.**

Investigate the **deductive reasoning** capabilities of LLMs.

A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences

Leonardo Bertolazzi,
DISI, University of Trento
leonardo.bertolazzi@unitn.it

Albert Gatt,
ICS, Utrecht University
a.gatt@uu.nl

Raffaella Bernardi
CIMeC and DISI, University of Trento
raffaella.bernardi@unitn.it

EMNLP 2024

Investigate **systematic generalization** for logical reasoning in LLMs.

A MIND for Reasoning: Meta-learning for In-context Deduction

Leonardo Bertolazzi[1], Manuel Vargas Guzmán[2], Raffaella Bernardi[3], Maciej Malicki[2], Jakub Szymanik[1],

[1]University of Trento, [2]University of Warsaw, [3]Free University of Bozen-Bolzano

Syllogisms (with pseudo-words) as testbed to check systematic generalization


Knowledge Base All notered are moner, All notered are tingda, All longeast are partber, All longeast are sionship, All moner are pointfish, All varvel are notered, All pointfish are disone, All pointfish are longeast

Hypothesis All varvel are tingda

1. All a are c
2. All c are b

Therefore, all a are b

TASK:
Given a KB and an hypothesis, identify within the KB the *minimal set of premises* needed to derive the hypothesis

Type	Inference
1	$\{Aa-b, Ac-d, Oad\} \vDash Obc$
2	$\{Aa-b\} \vDash Aab$
3	$\{Aa-b, Ac-d, Aa-e, Ede\} \vDash Obc$
4	$\{Aa-b, Aa-c\} \vDash Ibc$
5	$\{Aa-b, Ac-d, Ae-f, Iae, Edf\} \vDash Obc$
6	$\{Aa-b, Ac-d, Ebd\} \vDash Eac$
7	$\{Aa-b, Ac-d, Iac\} \vDash Ibd$

Inference length: number of A-formulas among its premises

Datasets: training, validation, test set for each inference type and length combination (min: 0, max: 19).

Bertolazzi, Vargas Guzmán, Bernardi, Malicki, and Zymanik

Core Generalization: unseen KB, but seen inference length

Length Generalization

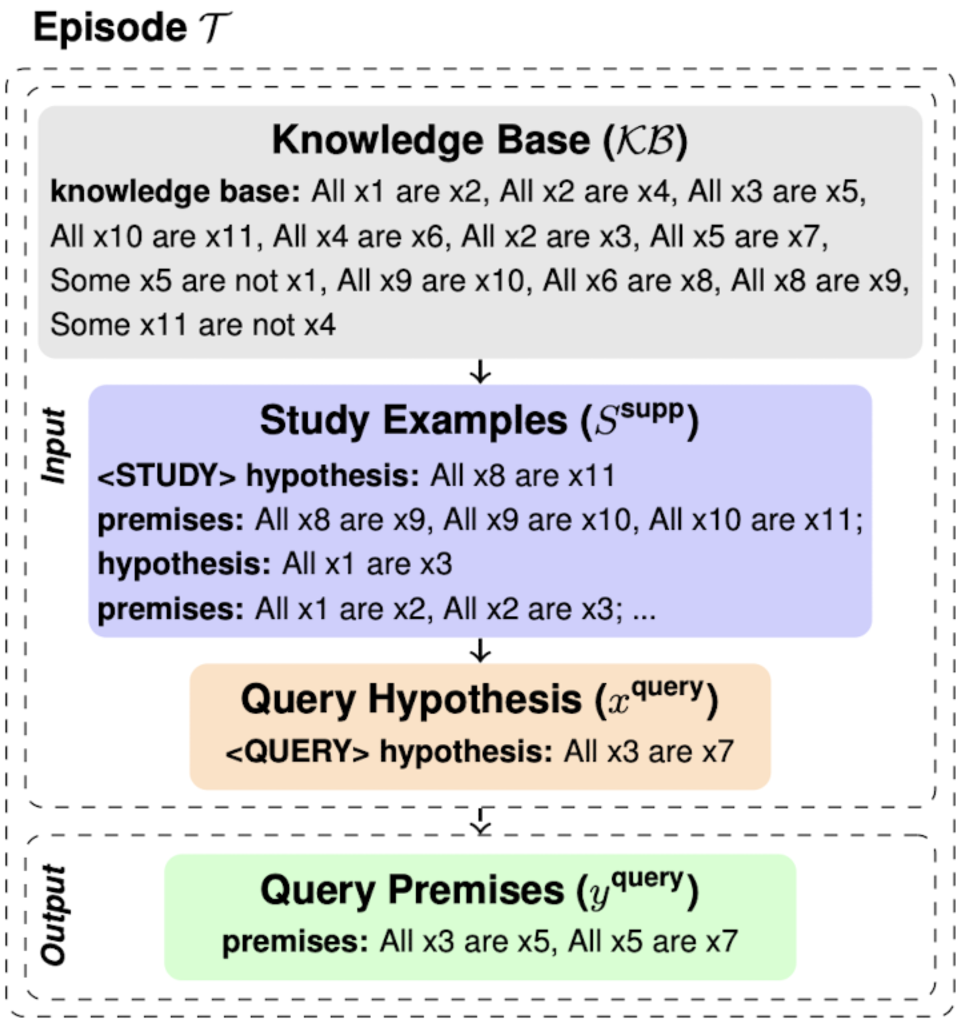
Training		Testing
Longer inferences: "all x1 are x2, all x2 are x3, all x3 are x4, all x4 are x5, all x5 are x6 ⊢ all x1 are x6"	→	**Shorter inferences:** "all x1 are x2, all x2 are x3 ⊢ all x1 are x3"
Shorter inferences: "all x1 are x2, all x2 are x3, all x3 are x4 ⊢ all x1 are x4"	→	**Longer inferences:** "all x1 are x2, all x2 are x3, all x3 are x4, all x4 are x5, all x5 are x6 ⊢ all x1 are x6"

Results

	Model	Method	All	Short	Long
Prompting	GPT-4o	Few-shot	39.76	52.91	33.51
		Zero-shot	15.90	28.97	9.89
	o3-mini	Few-shot	88.45	87.91	88.51
		Zero-shot	67.98	73.29	64.54

- ✓ We show that SOTA models, o3-mini and GPT-4o, **in a zero-shot** setting still struggle with this task.
- ✓ **Few shot examples** are sufficient for o3-mini to boost its performance, while they don’t for GPT-4o.

Metalearning



During meta-learning the model is **expected to learn the structure of the arguments** regardless of their specific content.

Study Examples of the same type as the Query Premises, and are either a) of the same (aligned) or different (disaligned) inference length of the query premises.

Results

✓ **Small LM post-trained** with meta-learning outperform SOTA models.

O3mini 88.45 (few shot)

Core generalizability

	Model	Method	All	Short	Long
Fine-tuning	Qwen-2.5 1.5B	MIND	93.11 ± 0.61	94.28 ± 0.61	91.76 ± 0.27
		Baseline	85.56 ± 1.24	91.42 ± 0.82	80.56 ± 1.78
	Qwen-2.5 3B	MIND	96.16 ± 0.44	96.24 ± 0.56	95.55 ± 0.43
		Baseline	93.03 ± 1.15	95.34 ± 1.18	90.92 ± 1.27
	Qwen-2.5 7B	MIND	98.13 ± 0.98	98.26 ± 0.82	97.69 ± 1.40
		Baseline	95.76 ± 1.10	97.27 ± 1.22	94.13 ± 0.90

Length generalizability

Model	Method	Short → Long		Long → Short	
		Disaligned	Aligned	Disaligned	Aligned
Qwen-2.5 1.5B	MIND	76.42 ± 2.95	91.75 ± 1.10	70.94 ± 2.27	71.13 ± 1.83
	Baseline	63.53 ± 1.16	63.53 ± 1.16	56.67 ± 1.22	56.67 ± 1.22
Qwen-2.5 3B	MIND	87.61 ± 1.97	95.86 ± 0.70	77.19 ± 3.53	78.53 ± 1.71
	Baseline	76.78 ± 1.63	76.78 ± 1.63	71.88 ± 1.49	71.88 ± 1.49
Qwen-2.5 7B	MIND	90.03 ± 1.09	96.84 ± 0.15	76.23 ± 2.91	83.41 ± 1.63
	Baseline	80.76 ± 2.65	80.76 ± 2.65	71.08 ± 1.55	71.08 ± 1.55

	Method	NVM [%]	Avg. NVM	MAP [%]	Avg. MAP	HP [%]
	MIND (aligned)	42.94	4.9	36.68	2.1	57.5
$L \to S$	MIND (disaligned)	28.31	3.72	52.81	1.76	66.06
	Baseline	28.21	6.19	23.38	2.1	72.78
	MIND (aligned)	9.76	1.66	87.54	5.08	60.94
$S \to L$	MIND (disaligned)	14.14	6.14	81.82	3.65	35.35
	Baseline	3.87	2.36	89.79	6.66	66.9

Table 3: **Error analysis.** Error analysis comparing MIND and baseline on long to short ($L \to S$) and short to long ($S \to L$) generalization. The table shows percentages and averages for non-minimal valid sets of premises (NVM) and missing necessary A premises (MAP), and the percentage of hallucinated premises (HP).

NVM could be acceptable,

MAP and HP not.

What have we learned and what is next?

Post-training methods enhance LLMs ability to dissociate form from content.

Yet, LLMs **have not learned logical reasoning** properly neither through SFT nor through meta-learning.

Next step: let's look inside their representations.