The Interplay between Language and Reasoning

Raffaella Bernardi

UNIVERSITY OF TRENTO

unibz

Something about my research life

Syntax-Semantics interface

Symbolic representations

Building blocks:
- The meaning of a sentence is the *truth value*
- Referential meaning (*entities* as building blocks)
- Semantic compositionality led by syntax
- Natural Language *Understanding*

Focus: grammatical words
(e.g. quantifiers, negation ..)
which guide formal reasoning.

Distributional Semantics

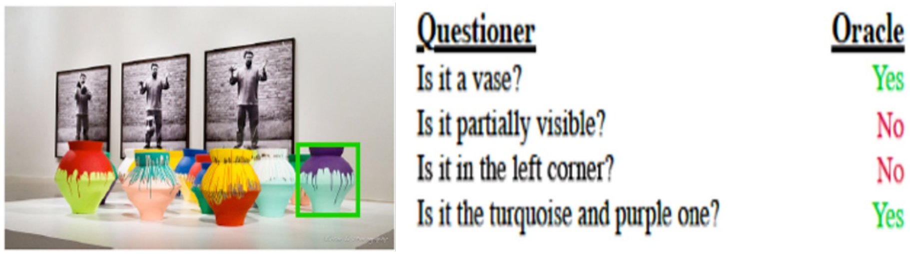
Vector representations

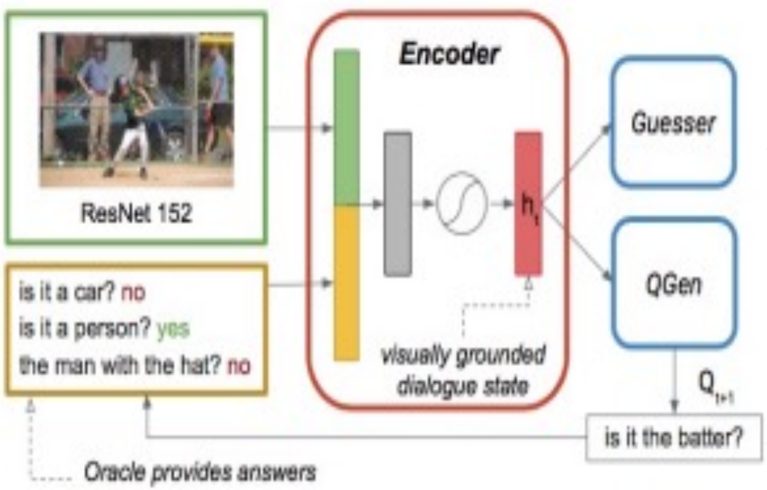
Building blocks
- The meaning of words is given by its context
- Neural NN era: Learning by predicting
- Natural Language *Generation*

Focus: content words (nouns, verbs..)
Analogical reasoning

Vision and Language Models

GuessWhat?!



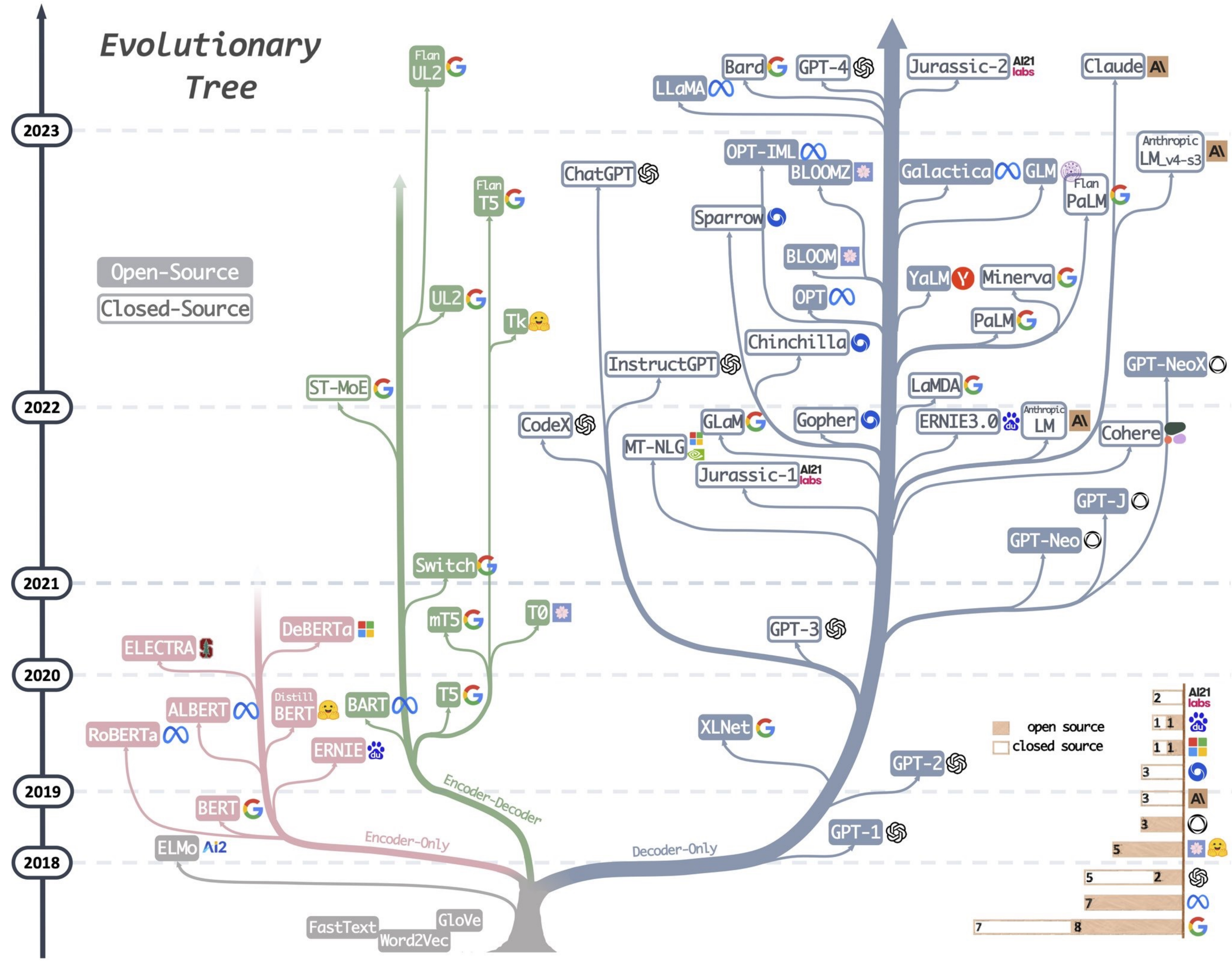


REASONING

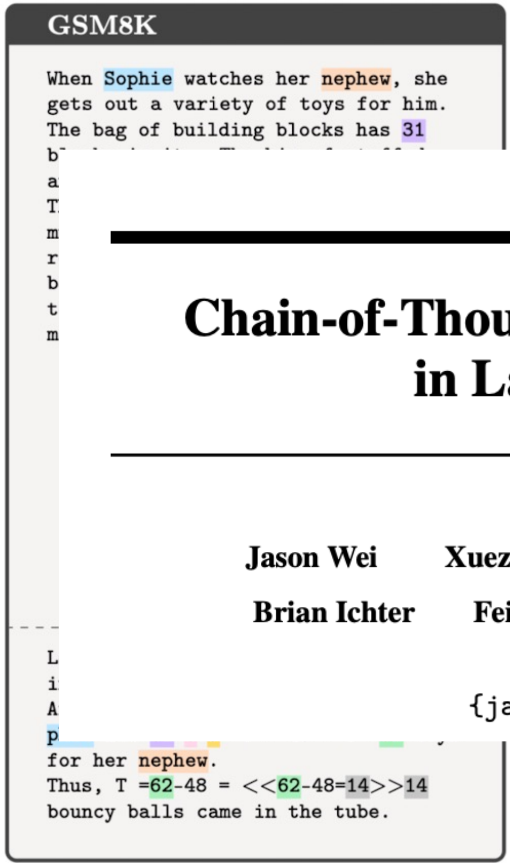
LANGUAGE
GENERATION

Beyond task Success:
Quality of the dialogues

Shekhar et al NAACL 2019


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
LLaMA
Bard
GPT-4
Jurassic-2
AI21 labs
Claude
Anthropic LM_v4-s3
ChatGPT
OPT-IML
BLOOMZ
Galactica
GLM
Flan PaLM
Flan T5
Sparrow
BLOOM
UL2
Tk
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
ST-MoE
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
AI21 labs
GPT-J
GPT-Neo
Switch
mT5
T0
GPT-3
ELECTRA
DeBERTa
ALBERT
Distill BERT
BART
T5
XLNet
RoBERTa
ERNIE
GPT-2
BERT
Encoder-Decoder
Encoder-Only
GPT-1
ELMo
Ai2
Decoder-Only
FastText
GloVe
Word2Vec
open source
closed source
2
1 1
1 1
3
3
3
5
5 2
7
7 8

Large Scale Benchmarks to evaluate LLMs reasoning ability

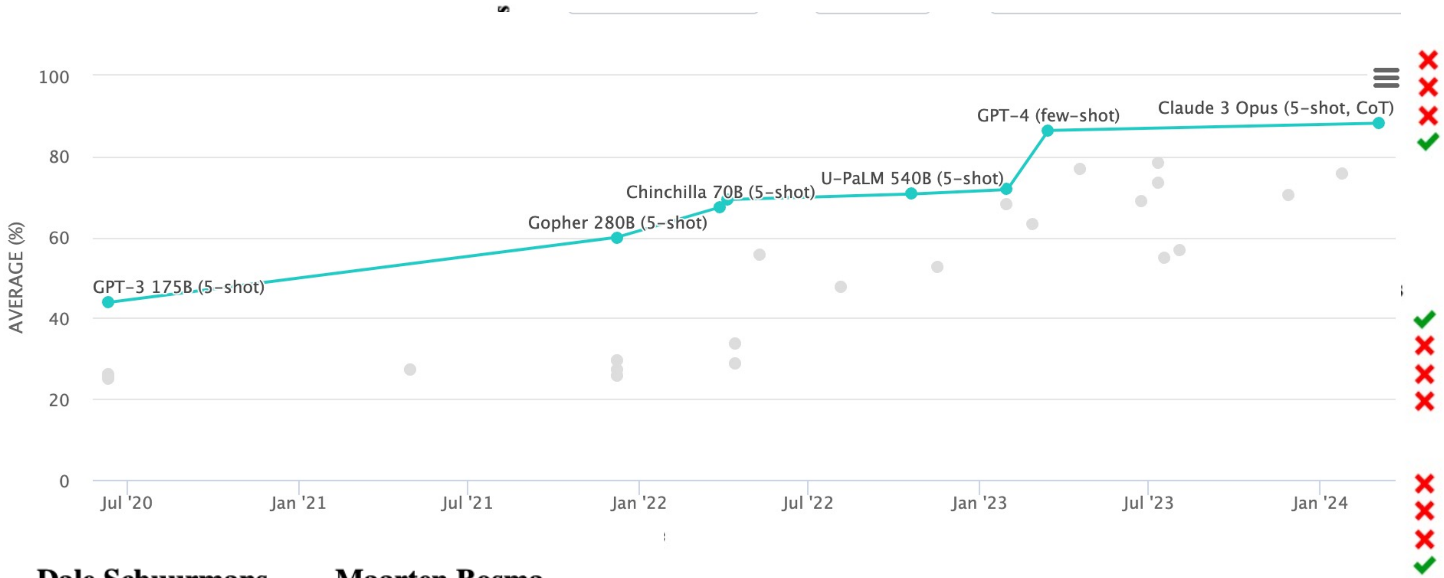


Grade School Math: 2021

Chain-of-Thought P
in Large L

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



amples from the Conceptual Physics and College Mathematics STEM tasks.

MMLU 2021

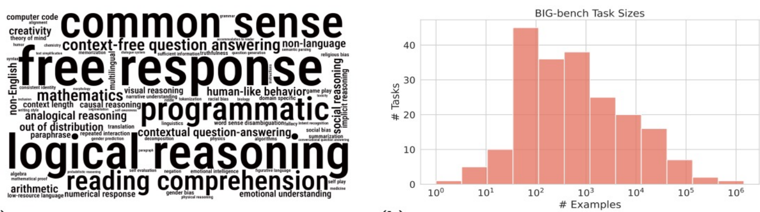


Figure 3: **Diversity and scale of BIG-bench tasks. (a)** A word-cloud of task keywords. **(b)** The size distribution of tasks as measured by number of examples.

Big Bench 2023

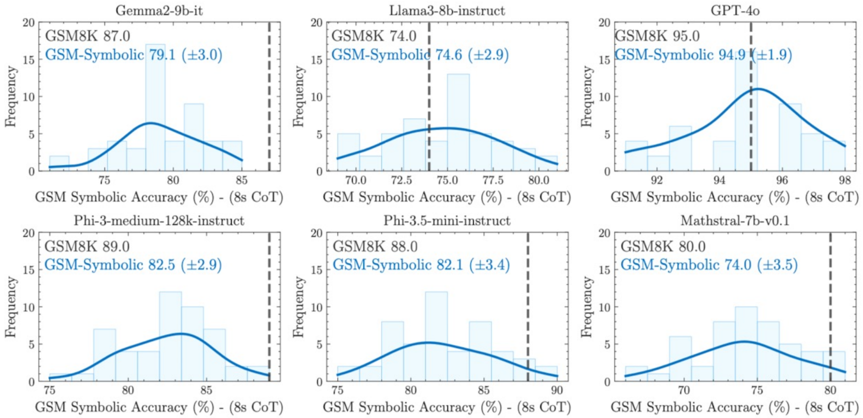


Figure 2: The distribution of 8-shot Chain-of-Thought (CoT) performance across 50 sets generated from `GSM-Symbolic` templates shows significant variability in accuracy among all state-of-the-art models. Furthermore, for most models, the average performance on `GSM-Symbolic` is lower than on `GSM8K` (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

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.

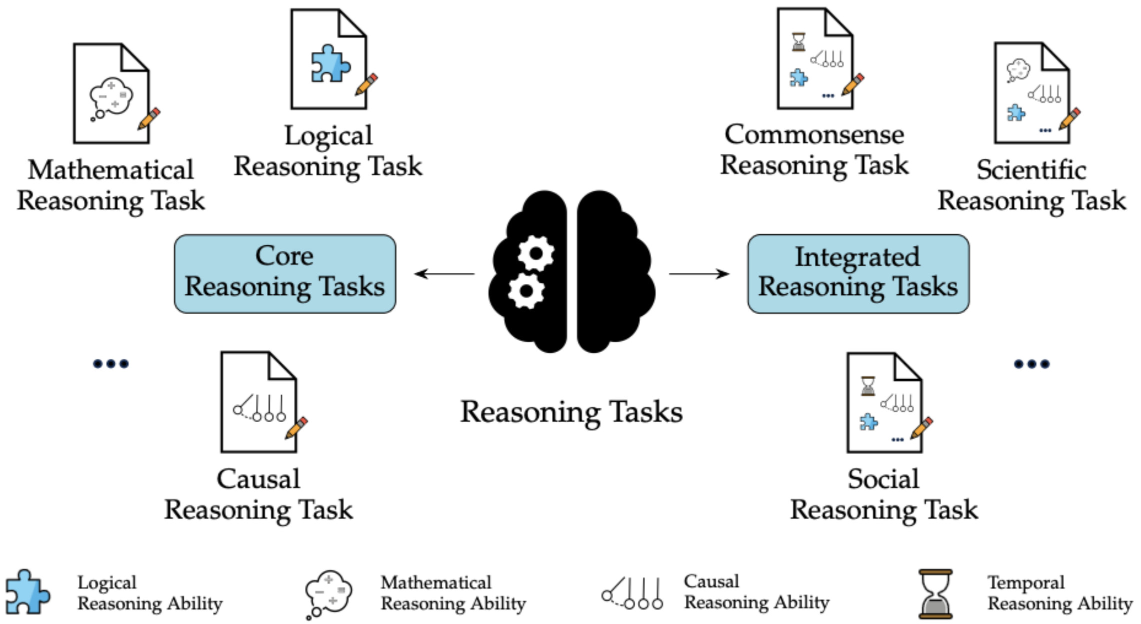


Figure 1: Schematic overview of the two types of reasoning tasks distinguished in this survey. *Core* reasoning tasks are designed to assess a particular reasoning ability within a given context. Conversely, *integrated* reasoning tasks involve the concurrent use of various reasoning skills. Tasks and abilities listed are not exhaustive.

Philipp Mondorf and Barbara Plank, COLM 2024

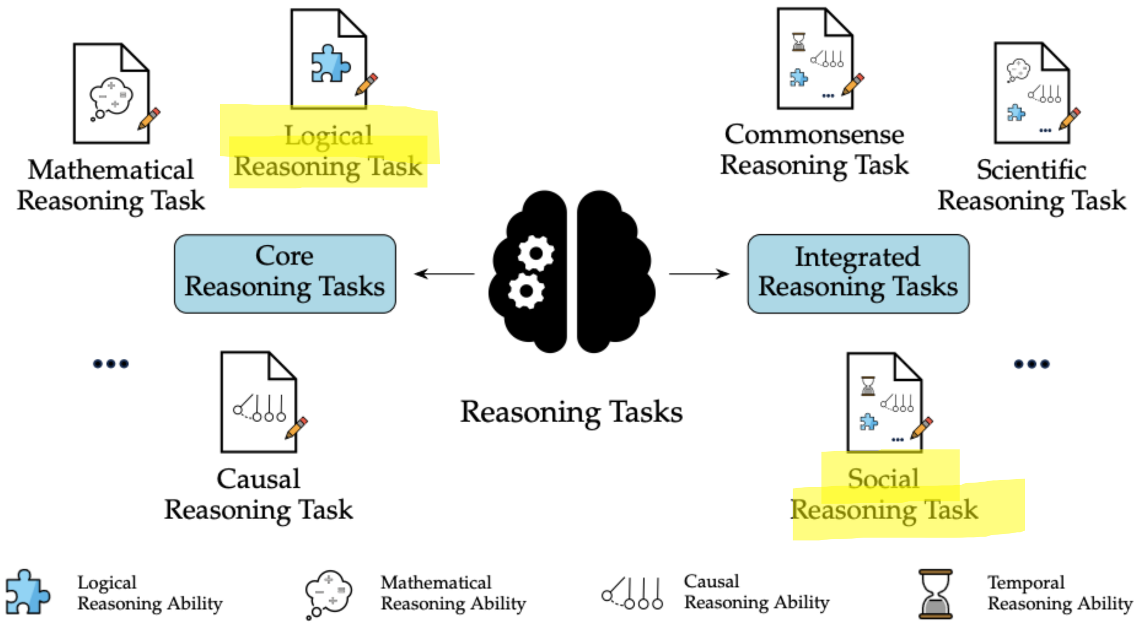


Figure 1: Schematic overview of the two types of reasoning tasks distinguished in this survey. *Core* reasoning tasks are designed to assess a particular reasoning ability within a given context. Conversely, *integrated* reasoning tasks involve the concurrent use of various reasoning skills. Tasks and abilities listed are not exhaustive.

Philipp Mondorf and Barbara Plank, COLM 2024

Investigate the **formal 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

Investigating the **reasoning-driven language generation** capabilities of LLMs.

Learning to Ask Informative Questions: Enhancing LLMs with Preference Optimization and Expected Information Gain

Davide Mazzaccara
CIMeC, University of Trento
davide.mazzaccara@unitn.it

Alberto Testoni
ILLC, University of Amsterdam
a.testoni@uva.nl

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

EMNLP 2024

Syllogisms as a test bed for formal reasoning

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

figures

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

moods

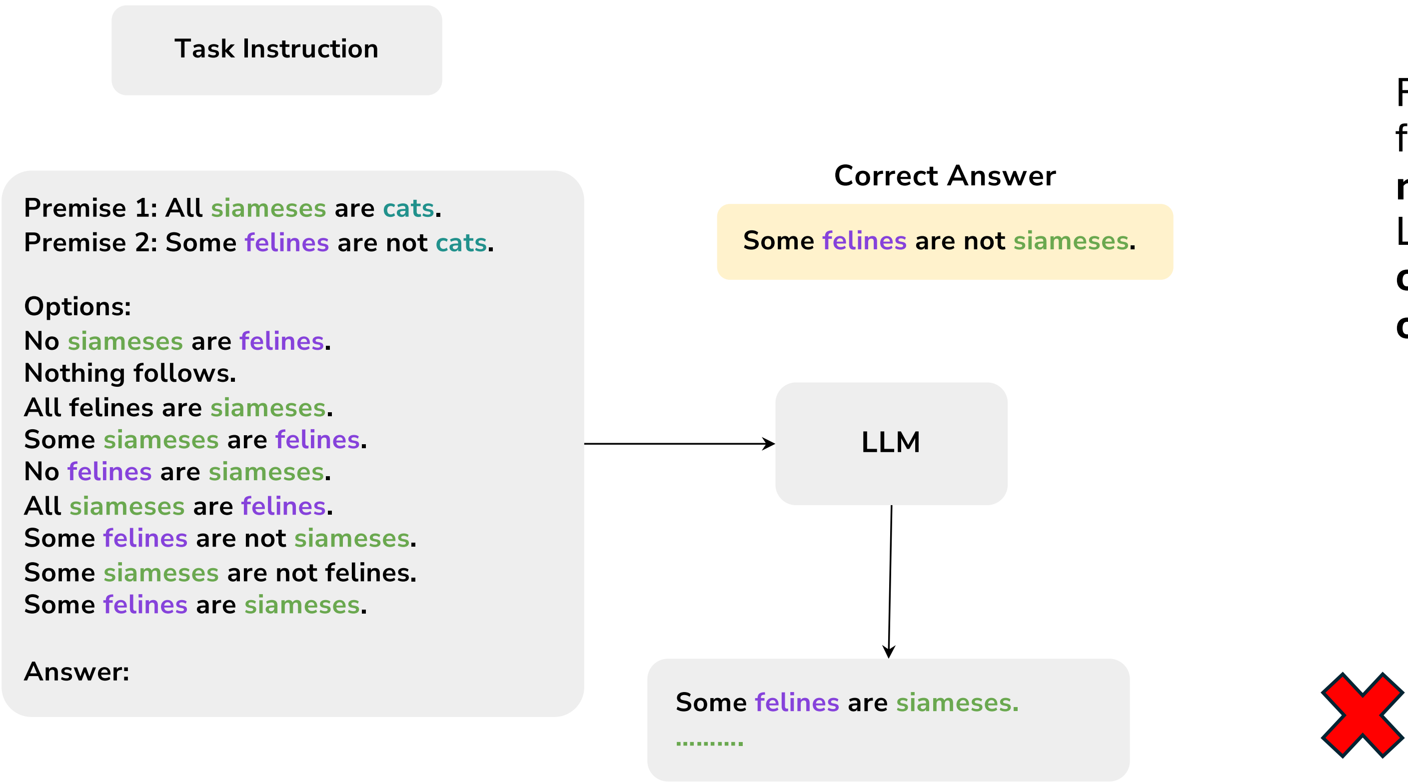
affirmative	negative
A: All a are b	**E**: No a are b
I: Some a are b	**O**: Some a are not b

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 an abstract model of how they can be solved → predicate logic
- We have evidence on how humans solve them in practice → cognitive psychology

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 and inference complexity

We create datasets that control for both semantic content and inference complexity.

For **semantic content**, we developed two datasets — one **believable** and the other **unbelievable** — which share the same vocabulary but differ in the believability of their conclusions.

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

→ **True Conclusion**

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

→ **False Conclusion**

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:

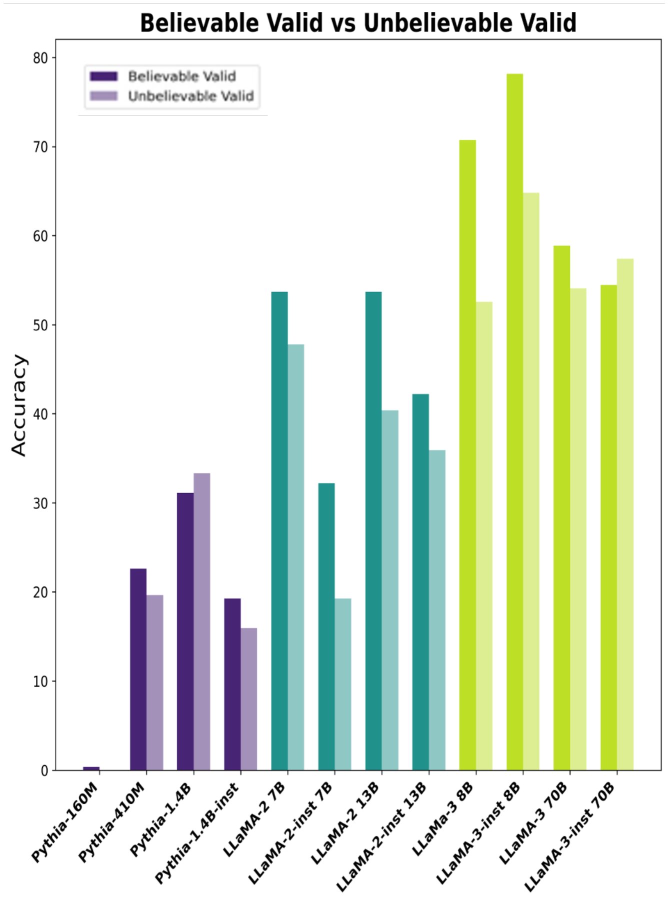
Premise 1: No tuem are graibly.
Premise 2: All graibly are kwaitz.
Conclusion: Some kwaitz are not tuem.

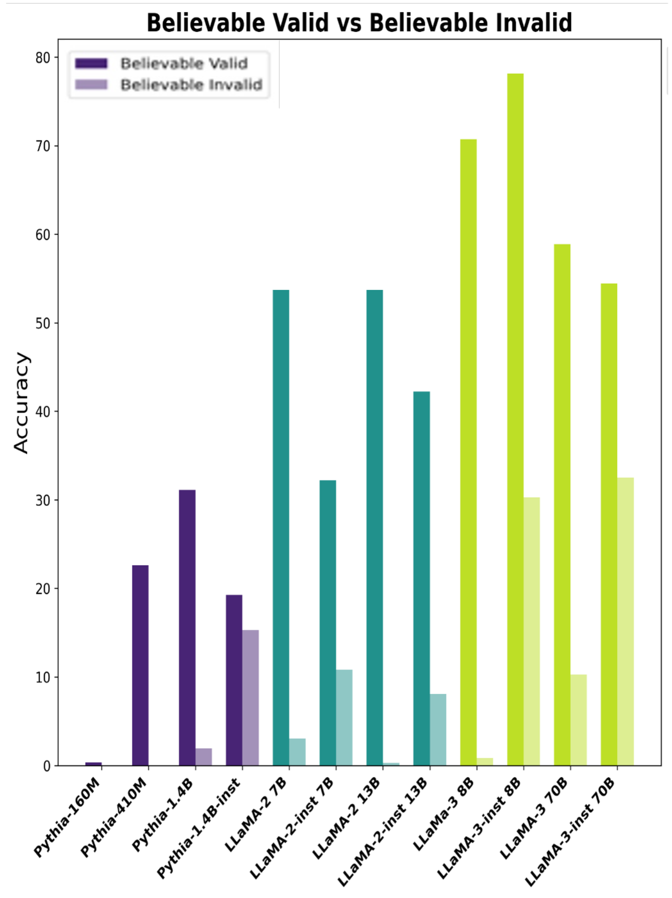
Premise 1: No khuipt are gnauntly.
Premise 2: All gnauntly are skaiank.
Premise 3: All skaiank are synulls.
Conclusion: Some synulls are not khuipt.

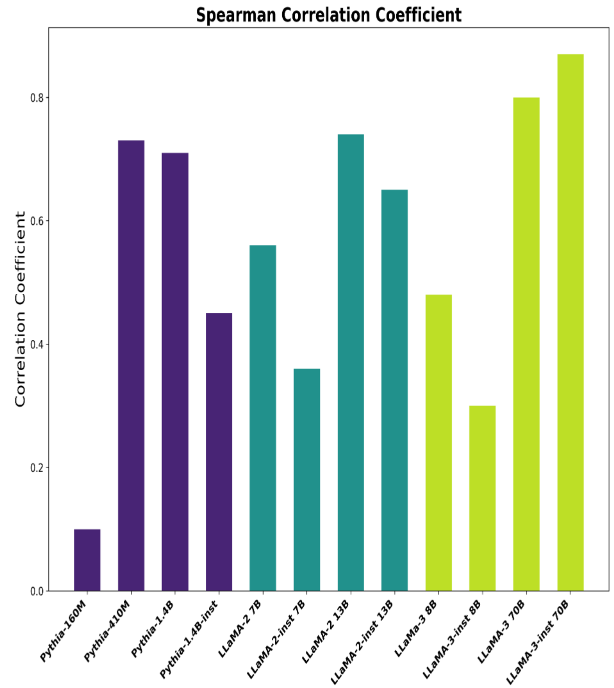
Premise 1: No screarm are pruerf.
Premise 2: All pruerf are thaon.
Premise 3: All thaon are mcnient.
Premise 4: All mcnient are tsiorm.
Conclusion: Some tsiorm are not screarm.

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.

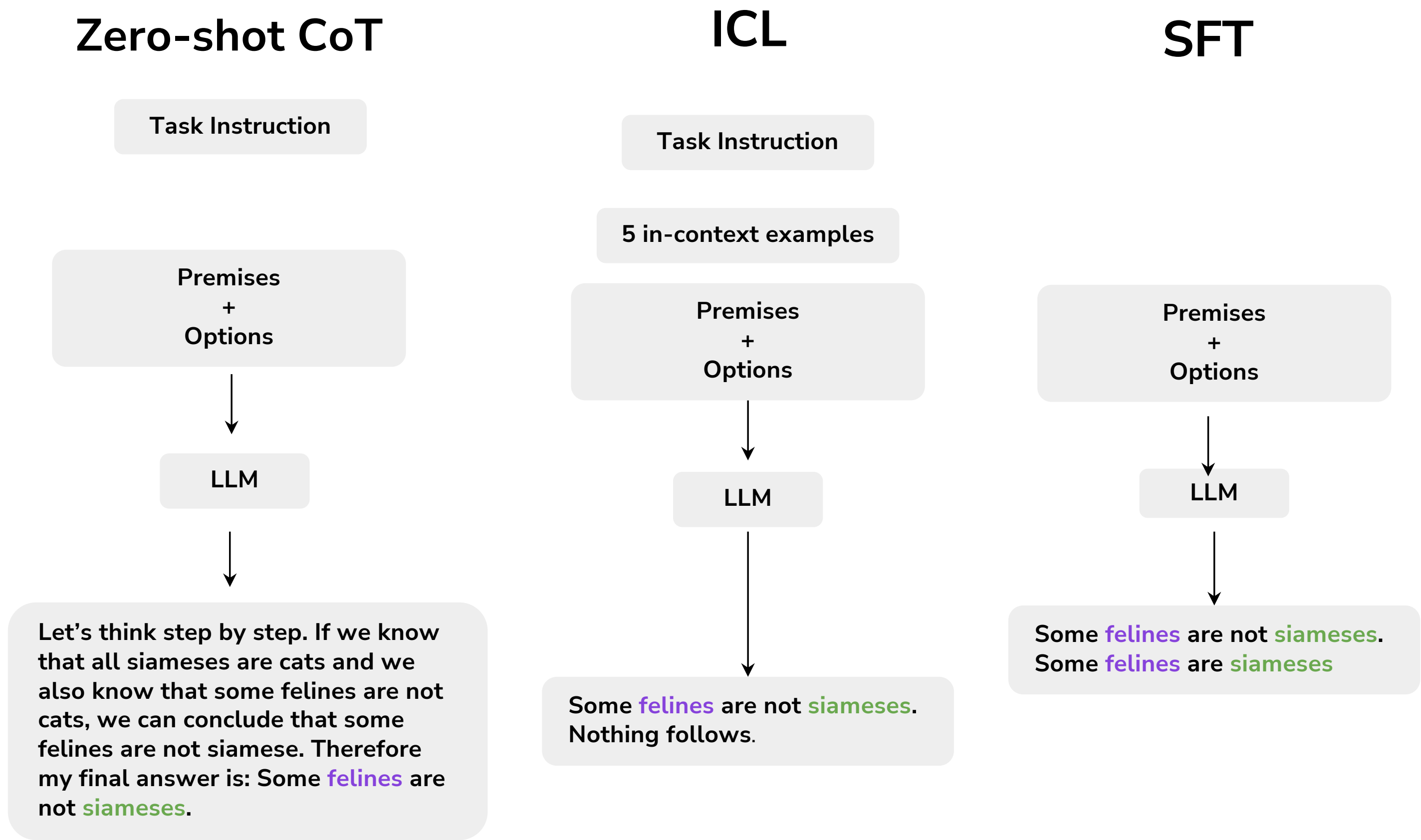
Options:
No siameses are felines.
Nothing follows.
All felines are siameses.
Some siameses are felines.
No felines are siameses.
All siameses are felines.
Some felines are not siameses.
Some siameses are not felines.
Some felines are siameses.

Answer:

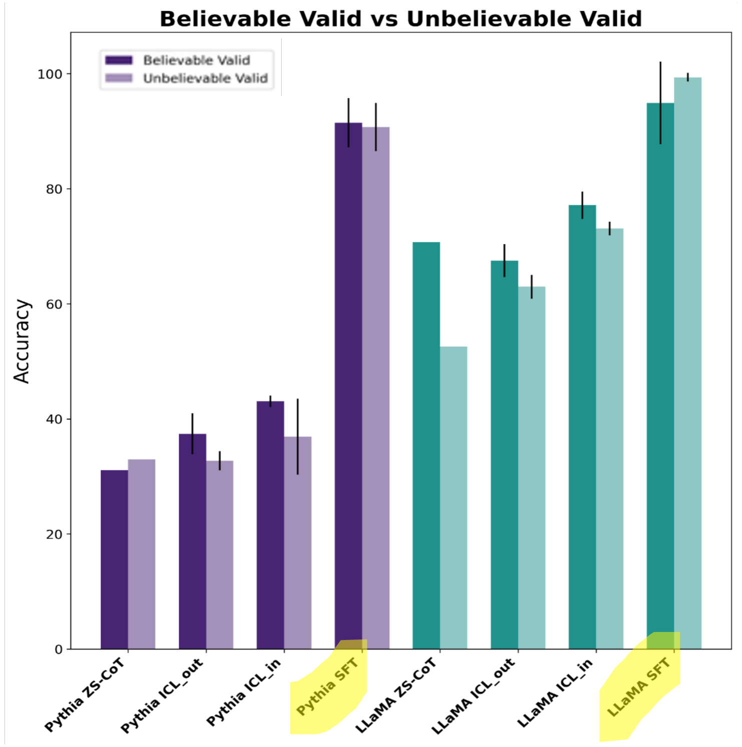
Correct Answer

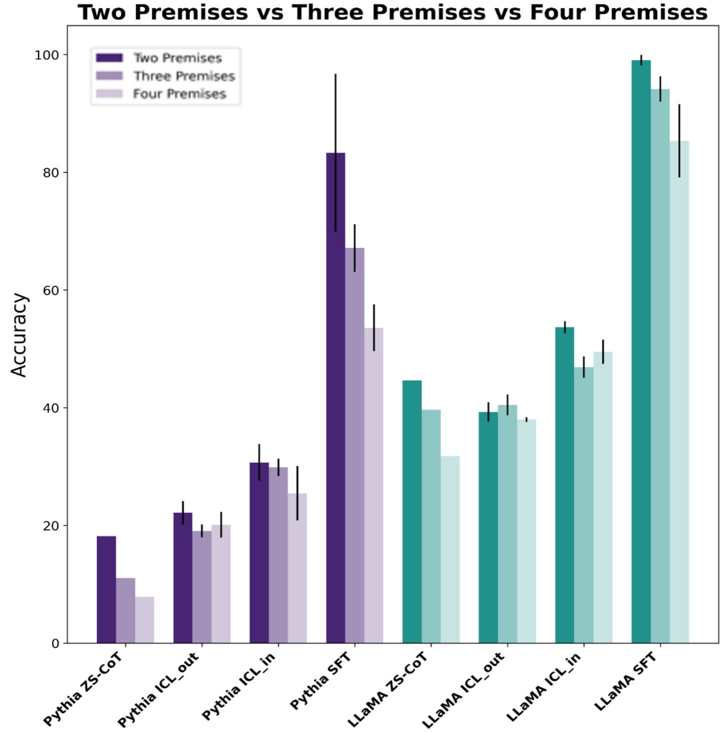
Some felines are not siameses.

ICL examples/SFT training: pseudowords



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

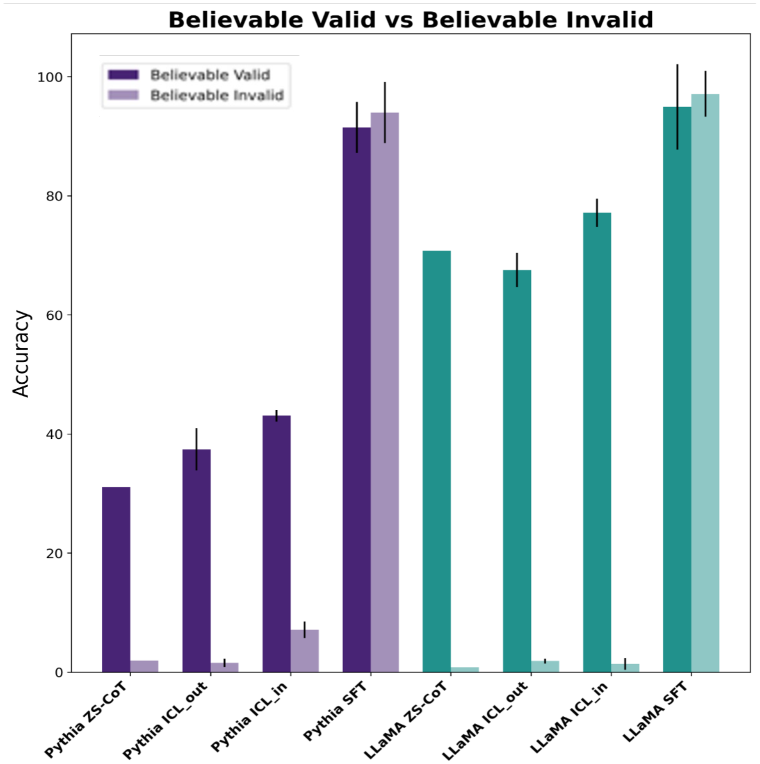


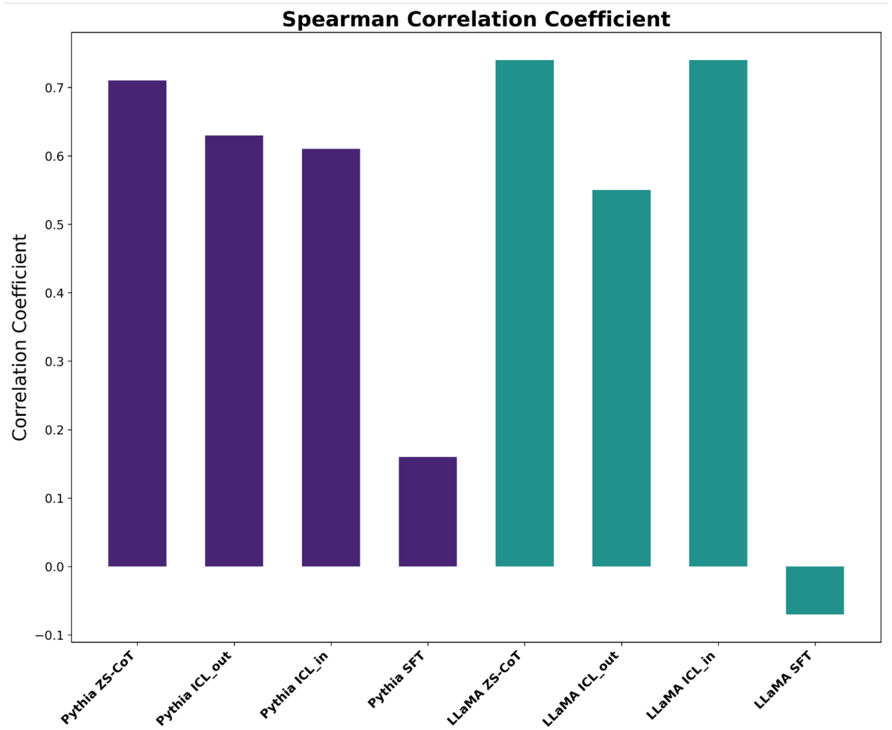


Content bias is reduced by ICL, but is only fully eliminated in SFT, where the model is exposed to many examples of the same inference with varying content**.**

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

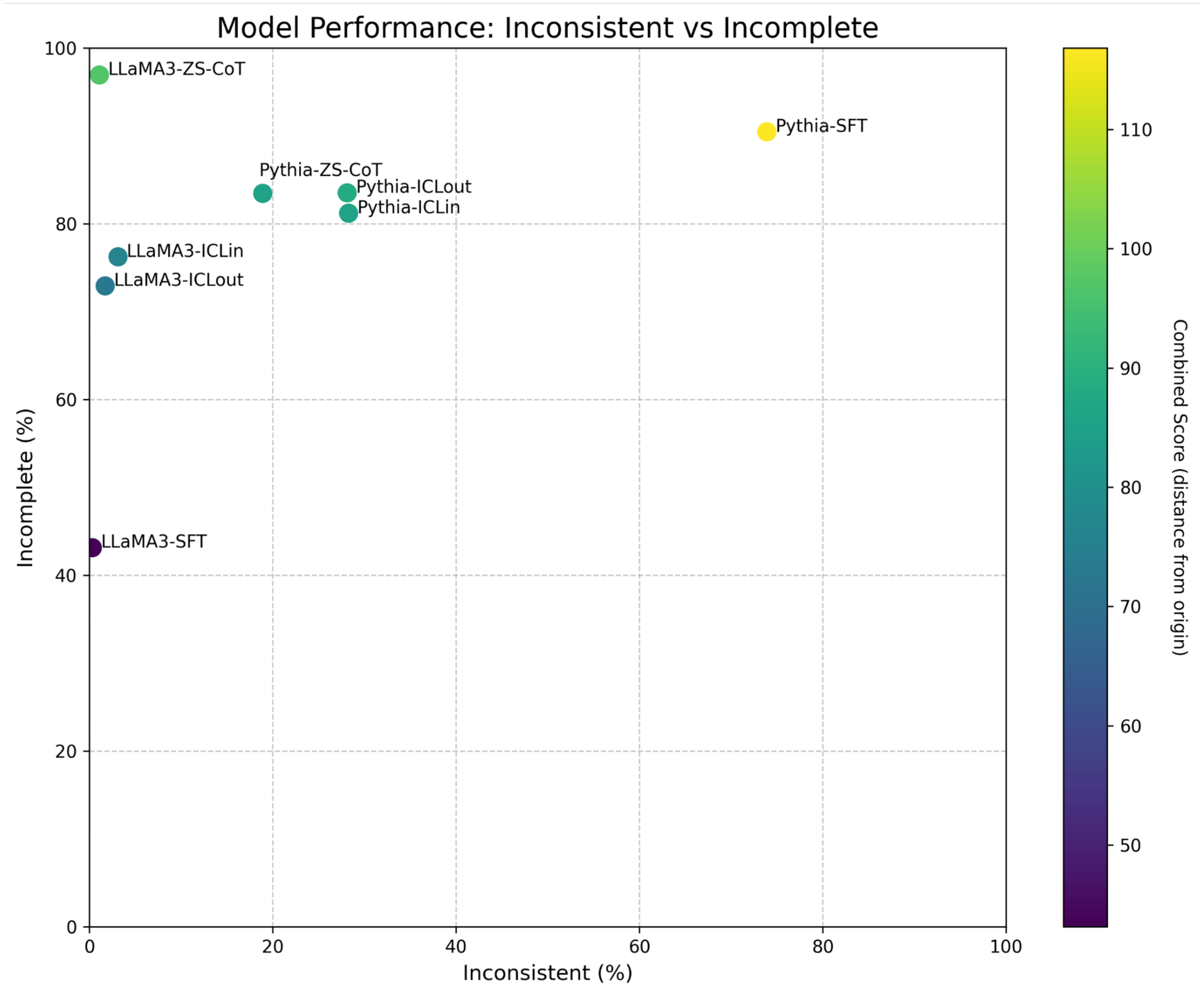




"Nothing follows" bias persists in ICL and disappears with SFT

Correlation with humans: SFT shows less alignment with humans – as we would expect from a formal reasoner, since humans have reasoning biases.

Consistent and Complete answers



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

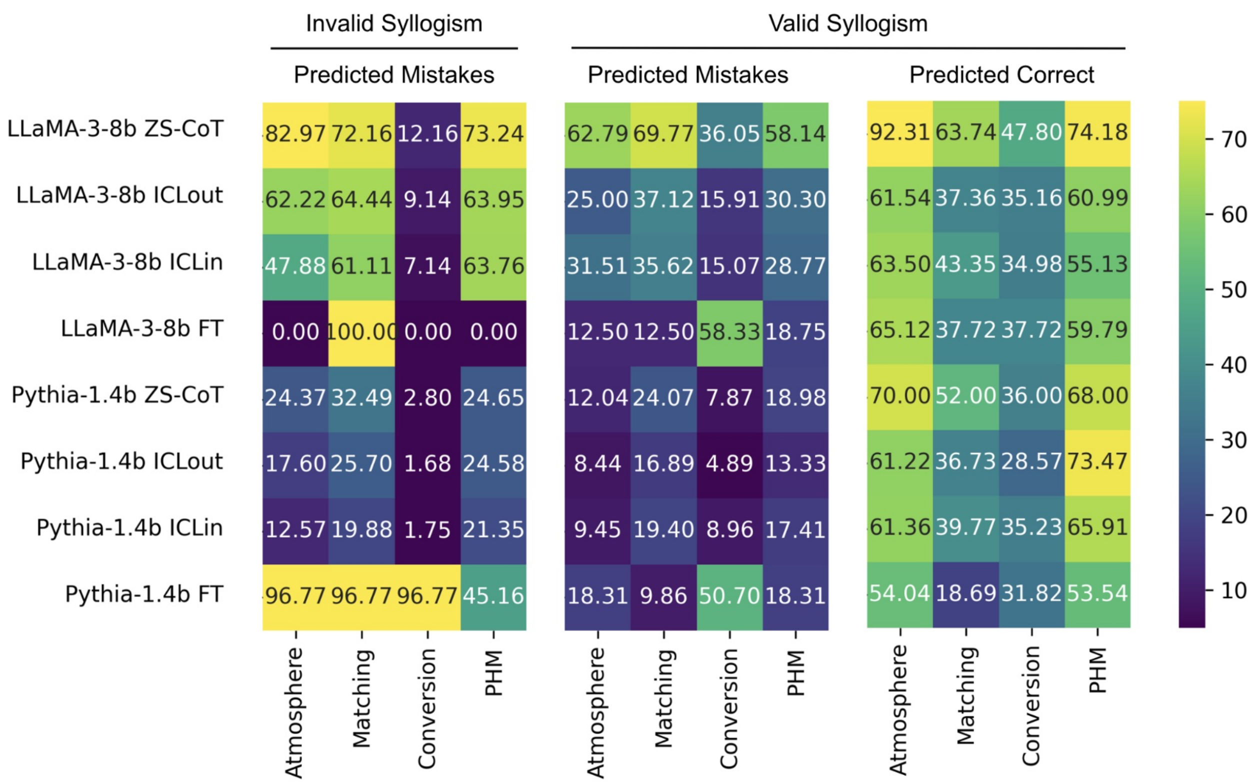
Consistency: the agent should not derive logically contradictory answers

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.**

Investigate the **formal 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

Investigating the **reasoning-driven language generation** capabilities of LLMs.

Learning to Ask Informative Questions: Enhancing LLMs with Preference Optimization and Expected Information Gain

Davide Mazzaccara
CIMeC, University of Trento
davide.mazzaccara@unitn.it

Alberto Testoni
ILLC, University of Amsterdam
a.testoni@uva.nl

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

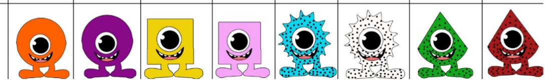
EMNLP 2024

Information Seeking Games

Guess the target

20- Q game

Hypothesis space with 8 members
Hierarchically organized.

Level	Features							
Higher: pattern	Solid				Dotted			
Middle: shape	Round		Square		Spiked		Triangular	
Lower: color	Orange	Purple	Yellow	Pink	Blue	White	Green	Red
Single object (each twice)								

Ruggeri et al Frontiers in Psychology 2021

Questioner	Answerer/Oracle
1. Is it dotted?	No
2. Is it round?	Yes
3. Is it purple?	No

4. It's the orange one. GUESSED

Constrain Seeking (CS) Q: Is it dotted?
Hypothesis Scanning (HS): Is it the orange one?

Language and Reasoning Interplay

Expected Information Gain (EIG): computes questions' informativeness as the reduction in entropy based on the expected answers to the question.
The optimal question in terms of EIG half split candidates in the set.

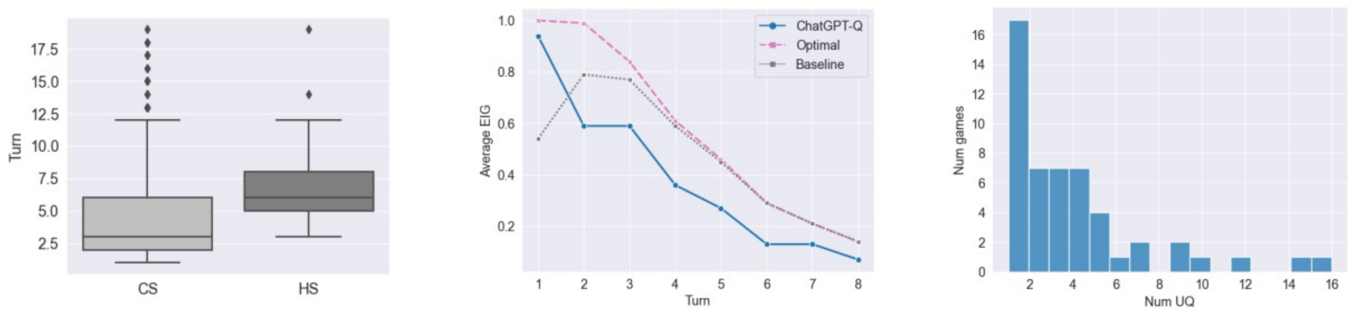
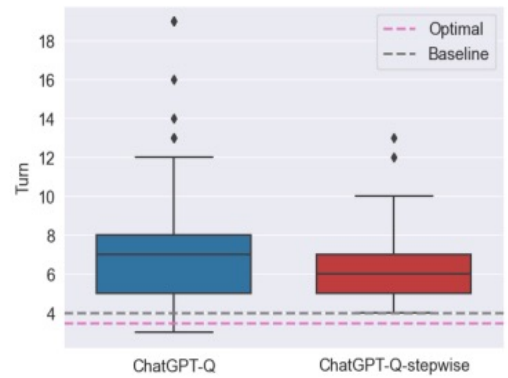


Figure 2: **Left**: ChatGPT-Q asks CS questions mostly in the earlier turns and HS in the later ones. Mann-Whitney U test shows the difference is statistically significant $p < 0.001$; **Middle**: ChatGPT's EIG is almost as high as the optimal model's EIG at the first turn, but it is lower than of the baseline's EIG at later turns. **Right**: Distribution of unnecessary questions.



Bertolazzi et al INLG 2023: ChatGPT's information seeking strategies

Small Open Access LLMs: 20-Questions Game

➢ **Problem**: smaller open-source LLMs generate appropriate CS questions but characterized by low informativeness. **Llama 2-chat (7B)**:

Target: seal
Candidates set: goat, ~~tomahawk~~, ~~sword~~, seal, walrus, ~~missile~~, cougar, ~~revolver~~.

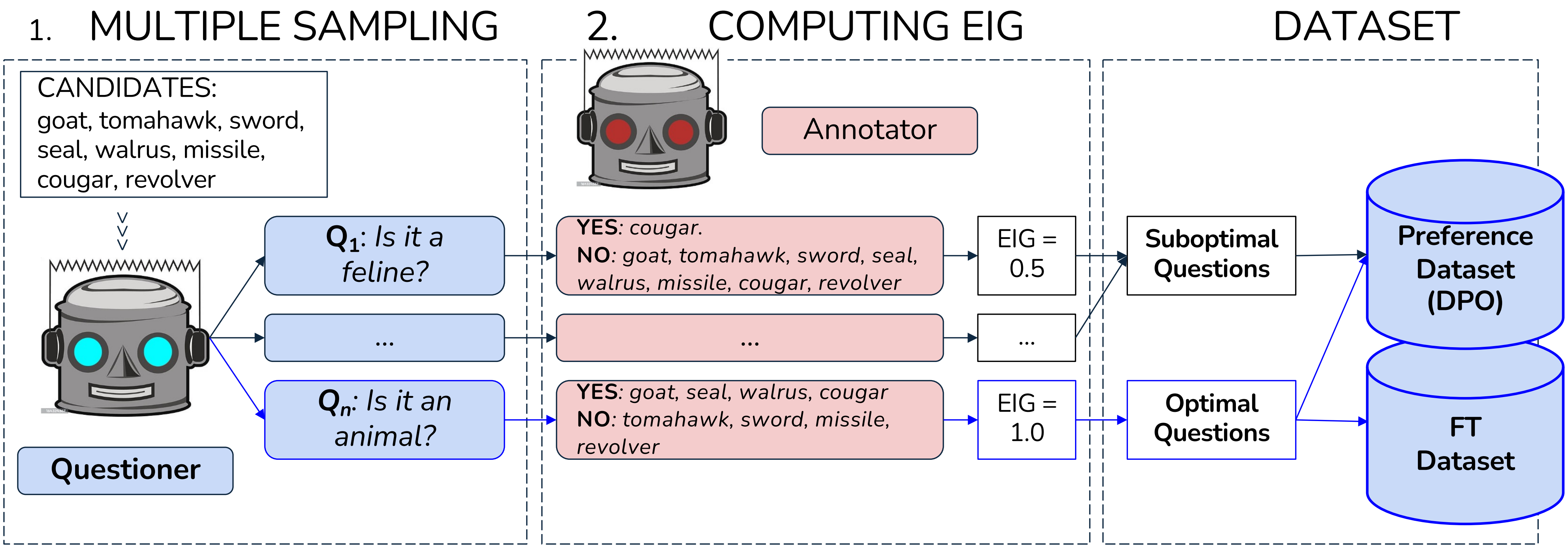
Zero-shot Dialogue

Q: Is the target a type of weapon? **A:** No.
Q: Is the target an animal? *EIG = 0* **A:** Yes.
Q: Does the target have a long, pointed snout? *EIG = 0* **A:** No.
Q: Does the target have a thick, blubbery layer of fat? **A:** Yes.
Q: I'm going to take a wild guess and say that the target is... a walrus! Am I right? **A:** No.
Q: Does the target have four legs? *EIG = 0* **A:** Yes.
Q: I'm going to take another wild guess and say that the target is... a seal! **A:** Yes.

➢ **Proposed Solution**: we propose a method consisting of three steps performed by the same LLM (Llama 2-chat (7B)):

1. multiple sampling questions,
2. evaluating questions in terms of (self-annotated) EIG,
3. DPO training.

Proposed Method



3. TRAINING

>>> **DPO** dataset: 55k **pairs** of optimal question vs suboptimal question

>>> **FT** dataset: 4k **dialogues** of only 1-EIG questions

Results: Different Domains

Candidate Sets (|Ω|=8):

- INLG: 90 cds of unseen candidates from known categories
- Things: 90 cds of unseen but common-life candidates
- Celebrities: 90 cds of unseen candidates from unknown categories (celebrities)

Results:

- For **INLG** and **Things**, DPO improves over the base-line on **effectiveness (S@1)**, **dialogue efficiency (AQ)** and **questions' informativeness (EIG)**,
- For **Celebrities**, DPO improves in terms of dialogue efficiency (**AQ**) and questions' informativeness (**EIG**), not in effectiveness (S@1).

Set	Method	S@1 ↑	AQ ↓	EIG ↑
INLG	zero-shot	56.7%	7.1	0.34
	FT	46.6%	**4.6**	0.41
	DPO	**68.9%**	5.2	**0.45**
Things	zero-shot	51.1%	7.5	0.29
	FT	42.2%	5.4	0.31
	DPO	**61.1%**	**5.2**	**0.40**
Celebrities	zero-shot	71.1%	7.6	0.35
	FT	46.7%	5.5	0.39
	DPO	72.2%	**5.1**	**0.47**

Results: Different Size

Larger candidate sets:

- INLG 16 (|Ω|=16): cds of unseen candidates of seen categories
- BigBench (|Ω|=29): cds of unseen candidates from unseen categories

Results:

Set	Method	S@1 ↑	AQ ↓	EIG ↑
INLG 16	zero-shot	44.4%	9.5	0.31
	DPO	**51.1%**	**6.3**	**0.38**
BigBench	zero-shot	17.2%	8.8	0.31
	DPO	**31.0%**	8.1	0.28

▾ items [] 16 items
0 "parakeet"
1 "seal"
2 "goose"
3 "stork"
4 "deer"
5 "finch"
6 "seagull"
7 "dog"
8 "dolphin"
9 "cow"
10 "budgie"
11 "otter"
12 "rooster"
13 "canary"
14 "walrus"
15 "sheep"
target "cow"

▾ items [] 29 items
THINGS
0 "apple"
1 "television"
2 "dinosaur"
3 "airplane"
4 "house"
5 "tree"
6 "coat"
7 "shoes"
8 "car"
9 "train"
10 "shower"
11 "frisbee"
12 "cow"
13 "giganotosaurus"
14 "siberian husky"
15 "glass micropipette"
ABSTRACTS
16 "anger"
17 "love"
18 "hate"
19 "contentment"
20 "jealousy"
21 "surprise"
22 "disgust"
23 "hopefulness"
24 "global poverty"
25 "phase transition"
26 "positive sum game"
27 "beauty"
28 "representative democracy"
target "apple"

Analysis: Types of Questions

More **effective information-seeking strategy**: CS questions → HS questions

Set	Method	HS		CS	
		%	EIG	%	EIG
INLG	zero-shot	47.45	0.25	52.55	0.42
	DPO	60.05	0.33	39.95	**0.62**
Things	zero-shot	44.13	0.22	55.87	0.35
	DPO	67.59	0.33	32.41	**0.56**
Celebrities	zero-shot	31.84	0.15	68.16	0.44
	DPO	53.92	0.36	46.08	**0.59**
INLG 16	zero-shot	38.66	0.19	61.34	0.39
	DPO	51.47	0.27	48.53	**0.51**
BigBench	zero-shot	64.2	0.22	35.71	0.48
	DPO	61.64	0.21	38.36	0.41

Target: anger
Candidates set: car, glass micropipette, global poverty, house, anger, train, airplane, television, representative democracy, love, giganotosaurus, hate, tree, surprise, disgust, siberian husky, phase transition, shoes, frisbee, jealousy, positive sum game, beauty, coat, apple, cow, contentment, hopefulness, dinosaur, shower

DPO Dialogue

Q: Is the target an animal? **A:** No
Q: Is the target a type of vehicle (car, train, airplane, etc.)? **A:** No.
Q: Is the target a type of building or structure (house, train station, airport, etc.)? **A:** No.
Q: Is the target a form of technology (computer, phone, micropipette, etc.)? **A:** No.
Q: Is the target a type of clothing or accessory (shirt, pants, hat, etc.)? **A:** No.
Q: Is the target a type of emotion or feeling (love, hate, joy, etc.)? **A:** Yes.
Q: Great! Is the target "**love**"? **A:** No.
Q: Is the target "**hate**"? **A:** No.
⋮
Q: Is the target "**anger**"? **A:** Yes.

Analysis: Good Results in INLG

Target: seal
Candidates set: $\boxed{\text{goat}}$, tomahawk, sword, $\boxed{\text{seal}}$, $\boxed{\text{walrus}}$, missile, $\boxed{\text{cougar}}$, revolver.

DPO Dialogue

Q: Is the target a type of weapon? *EIG = 1*	**A:** No.
Q: Is the target a marine animal? *EIG = 1*	**A:** Yes.
Q: Is the target the walrus?	**A:** No.
Q: Is the target the seal?	**A:** Yes.

Conclusion

Training with DPO and EIG, Llama-2 learns to:

1. ask **more informative questions** (EIG)
2. follow a **more efficient strategy** (AQ and types of questions)
3. be **more effective** in the 20Qs game (S@1)

More broadly:

❖ LLMs could **learn to ask more informative questions** and **more effective information-seeking behavior**

❖ LLMs informative behavior **generalize to different domains**

Back to the interplay between Language and Reasoning **in Communication**

Target = blouse

This is the list of candidates: zucchini, trousers, onions, potato, blouse, pants, slippers, rhubarb.

Questioner Is the item you have edible?

Answerer No, it is not edible.

Questioner Is it an article of clothing?

Answerer Yes, it is an article of clothing.

Questioner Is it a type of footwear?

Answerer No, it is not a type of footwear.

Questioner Does the item have long sleeves?

Language USE

Answerer It depends on the specific blouse, but some blouses do have long sleeves.

Questioner **Is the item made specifically for women?**

Answerer **Yes, it is typically made for women.**

Questioner **Is the item you have a blouse?**

Answerer **Yes! That's correct. Congratulations!**

Miss the interlocutor communicative goal!
12% ChatGPT-Q and 7% ChatGPT-Q-stepwise

Bertolazzi et al 2023

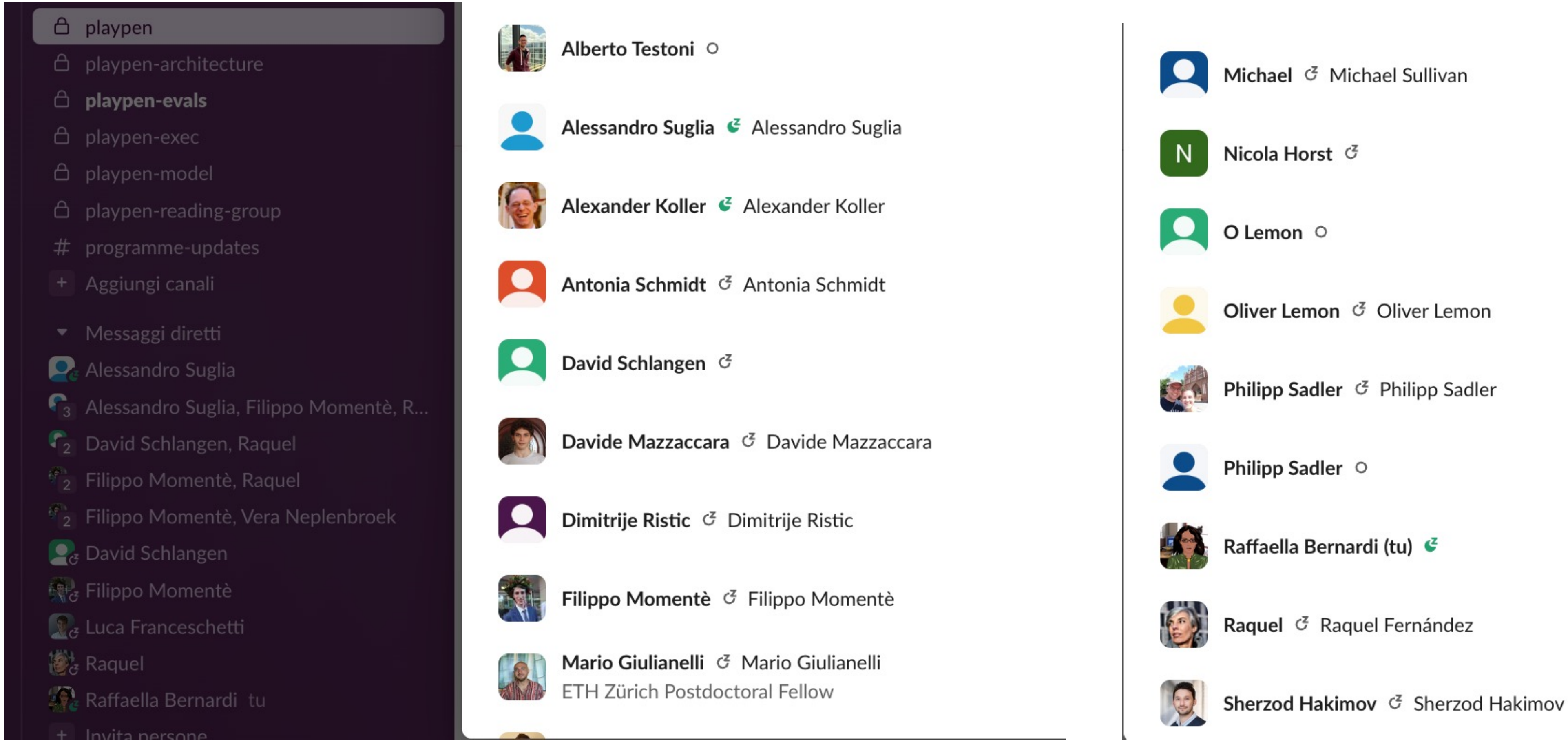
What do these two papers have in common?

Both studies profit from Cognitive Science literature to investigate Language & Reasoning interplay:

- **Syllogisms:** We studied LLMs formal reasoning ability through a **Heuristic Theory proposed to study human reasoning bias.**

- **20Q:** We valuate LLMs reasoning-driven generation with an **evaluation method used to evaluate children problem-solving skills.**

Overall message:
strength the collaboration between Cognitive Neuroscience and NLP to develop **carefully controlled tests** that should be paired with the ”*Language in Action*” tasks.



PLAYPEN TEAM

(Aim to) organize a shared task. ➔ Stay Tuned


Filippo Merlo


Davide Mazzaccara


Leonardo Bertolazzi


Filippo Momentè


Alberto Testoni


Philipp Mondorf


M. Vargas Guzmán


Sandro Pezzelle


Ravi Shekhar


B. Plank


J. Szymanik


R. Fernández


M. Malicki


A. Gatt