Logic, Plausibility, and Generalization: Making LLMs More Systematic

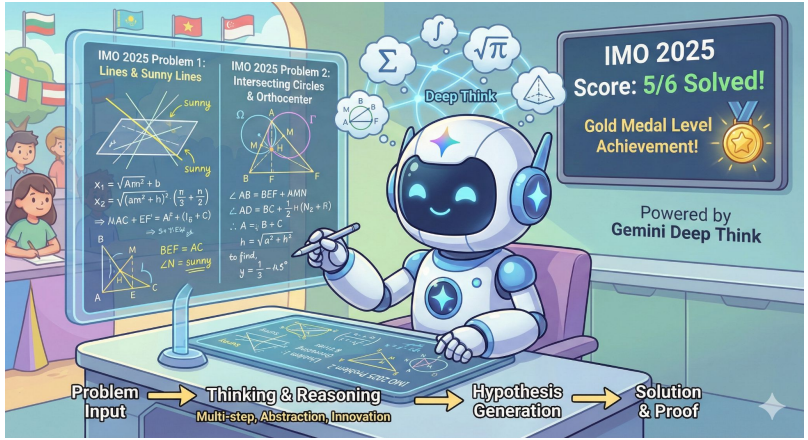
Leonardo Bertolazzi

Ph.D. Candidate at University of Trento

An interesting puzzle about current AI systems

The following two statements are both true at the same time:

1. LLMs have achieved **superhuman** (or better-than-most-human) **abilities** in very challenging tasks

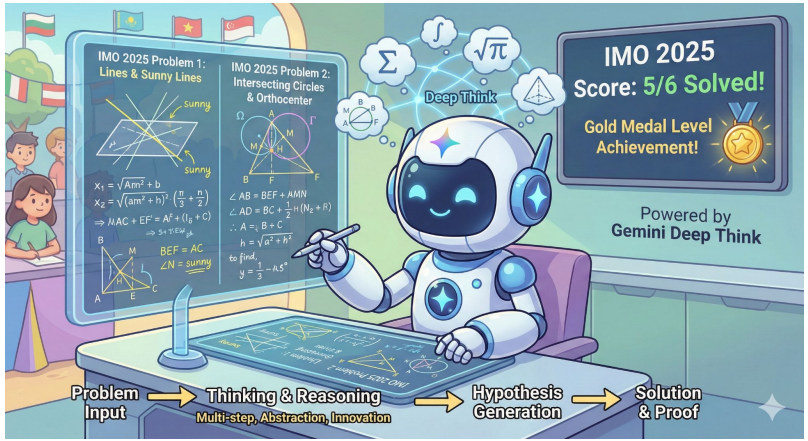


2. LLMs can **fail in unpredictable ways** compared to how humans would fail

An interesting puzzle about current AI systems

The following two statements are both true at the same time:

1. LLMs have achieved **superhuman** (or better-than-most-human) **abilities** in very challenging tasks



2. LLMs can **fail in unpredictable ways** compared to how humans would fail

What are the reasons behind this inconsistency?

A classic critique from cognitive science

Over thirty years ago, Fodor and Pylyshyn[2] (1988) argued that neural networks differ fundamentally from human minds because they lack **systematicity** + other linked properties.

Systematicity: The understanding of certain mental representations is structurally related to the understanding of associated ones. If you can think "John loves Mary," you can necessarily think "Mary loves John."

Productivity: The capacity to generate and understand an indefinite number of novel representations from a finite base.

Compositionality: complex representations are built from simpler constituents in rule-governed ways, and the meaning of the whole depends systematically on the meanings of the parts.

[2] Fodor, J. A., & Plyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition.

Which AI systems are systematic?

Fodor and Pylyshyn's idea of **systematicity** is easily instantiated in **classic rule-based models** that perform **symbol manipulation**.

- they have **explicit symbols** for LOVES, John, and Mary and can combine them freely

What about SOTA LLMs?

Anecdotally:

IF a AI system can solve IMO problems but fails at understanding the "greater-than" relation, THEN it is **not systematic** in its understanding of the mathematical domain.

More in general, for widely used models, **not systematic** because:

- Do **not** encode **explicit rules**
- Even if a **neural configuration** that can represent a systematic rule (like logical inference) **exists**, it doesn't mean a standard network will actually **learn** that rule from data using Gradient Descent.

How systematic is the human mind?

Fodor and Pylyshyn argued that the human cognitive architecture must be **symbolic** to account for the systematicity of thought.

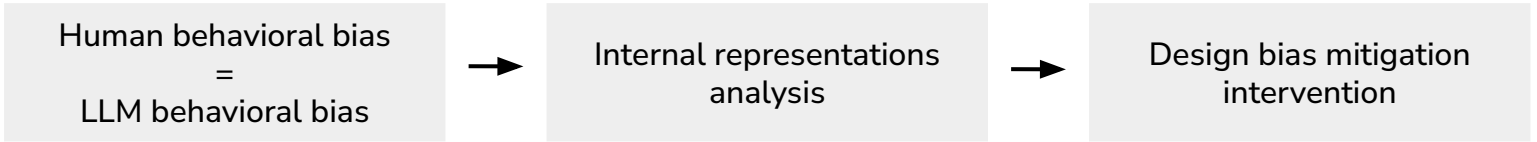
A large body of observed phenomena in humans, even in compositional and systematic domains like **logical reasoning** and **language use**, are **not** systematic and might be better explained by alternative hypothesis:

- **Domain-Specific Reasoning** → Performance varies dramatically across familiar vs. unfamiliar domains
- **Framing Effects** → Responses driven by presentation format rather than identical structure

Implication: while a symbolic architecture may capture some aspects of cognition, it may fail to account for many interesting and pervasive phenomena in human reasoning behavior.

Using what we know about humans to make LLMs more systematic

1. **Debiasing LLMs towards systematicity using representation engineering**



2. **Using meta-learning to induce systematicity in learning logical inferences**

- A body of work has shown empirically that meta-learning can induce **human-like systematic generalization** in Transformer NNs in linguistic tasks
- Can we use meta-learning to teach **logical inferences in a systematic way** to LLMs?

1. Debiasing towards systematicity

How Language Models Conflate Logical Validity with Plausibility: A Representational Analysis of Content Effects

Leonardo Bertolazzi[1], Sandro Pezzelle[2], Raffaella Bernardi[3],

[1]University of Trento, [2]University of Amsterdam, [3]Free University of Bozen-Bolzano
Correspondence: leonardo.bertolazzi@unitn.it

Abstract

Both humans and large language models (LLMs) exhibit *content effects*: biases in which the plausibility of the semantic content of a reasoning problem influences judgments regarding its logical validity. While this phenomenon in humans is best explained by the dual-process theory of reasoning, the mechanisms behind content effects in LLMs remain unclear. In this work, we address this issue by investigating how LLMs encode the concepts of validity and plausibility within their internal representations. We show that both concepts are linearly represented and strongly aligned in representational geometry, leading models to conflate plausibility with validity. Using steering vectors, we demonstrate that plausibility vectors can causally bias validity judgements, and vice versa, and that the degree of alignment between these two concepts predicts the magnitude of behavioral content effects across models. Finally, we construct debiasing vectors that disentangle these concepts, reducing content effects and improving reasoning accuracy. Our findings advance understanding of how abstract logical concepts are represented in LLMs and highlight representational interventions as a path toward more logical systems.

influential, positing two distinct modes of thought: a fast, intuitive, heuristic-driven system (System 1), and a slower, deliberative system responsible for analytical reasoning (System 2; Evans, 2008; Kahneman, 2011). Neuroscientific studies have provided empirical support for this framework, highlighting different neural substrates associated with these reasoning processes (Goel et al., 2000; Luo et al., 2014). Recent work has found that LLMs exhibit similar content effects in reasoning tasks (Lampinen et al., 2024); however, the underlying mechanisms driving these effects remain unknown.

In this work, we provide a representational account of why content effects may emerge in LLMs, investigating how the abstract concepts of validity and plausibility are encoded in their hidden representation space. Specifically, we build upon the linear representation hypothesis (Park et al., 2024), which proposes that many high-level concepts are encoded linearly within the latent space of LLMs. This hypothesis has been supported by empirical findings showing that concepts can often be captured by linear probes or manipulated with steering vectors (Liu et al., 2024; Rimsky et al., 2024; Marks and Tegmark, 2024). We hypothesize that content effects in LLMs may arise from the way

Content effects in reasoning

Human reasoning on logical problem is often **non systematic**. It is heavily influenced by the **semantic content** of the problem at hand[4].

All humans are mortal.
Italians are humans.
Therefore, Italians are mortal.

Human: **VALID**

All humans are plants.
Italians are humans.
Therefore, Italians are plants.

Human: **INVALID**

Definition: we call the first syllogism **plausible** since its conclusion is true in the actual world, **while the** second is **implausible** since its conclusion is false in the actual world.

[4]Evans et. al. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition.*

Content effects in Humans and LLMs

Humans

Memory & Cognition
1983, 11 (3), 295-306

On the conflict between logic and belief in syllogistic reasoning

J. St. B. T. EVANS, JULIE L. BARSTON, and PAUL POLLARD
Plymouth Polytechnic, Plymouth PL4 8AA, England

Three experiments are reported that investigate the weighting attached to logic and belief in syllogistic reasoning. Substantial belief biases were observed despite controls for possible conversions of the premises. Equally substantial effects of logic were observed despite controls for two possible response biases. A consistent interaction between belief and logic was also recorded; belief bias was more marked on invalid than on valid syllogisms. In all experiments, verbal protocols were recorded and analyzed. These protocols are interpreted in some cases as providing rationalizations for prejudiced decisions and, in other cases, as reflecting a genuine process of premise to conclusion reasoning. In the latter cases, belief bias was minimal but still present. Similarly, even subjects who focus primarily on the conclusion are influenced to an extent by the logic. Thus a conflict between logic and belief is observed throughout, but at several levels of extent.

An important debate in cognitive psychology surrounds the notion of rationality with respect to human inference (see Cohen, 1981, and associated commentaries). Recent reviews by Evans (1982) and Nisbett and Ross (1980) have stressed the role of apparently irra-

In this paper, we will focus on the alleged "belief-bias" effect in reasoning. The claim is that when presented with deductive arguments to evaluate, subjects will make judgments upon a priori beliefs rather than on the basis of logical argument. Specifically, they will

The Quarterly Journal of Experimental Psychology (1985) **37A**, 553–569

The Effects of Belief on the Spontaneous Production of Syllogistic Conclusions

J. V. Oakhill
MRC Perceptual and Cognitive Performance Unit, Laboratory of Experimental Psychology, University of Sussex, Brighton, U.K.

P. N. Johnson-Laird
MRC Applied Psychology Unit, Cambridge, U.K.

Two experiments examined the effects of subjects' beliefs on syllogistic inference. The first experiment showed that beliefs biased the spontaneous conclusions that subjects drew for themselves. These effects were more marked for indeterminate premises (which yield no non-trivial valid conclusions) than for determinate premises (which yield valid conclusions). There was also an effect of the nature of the beliefs: conclusions that were false by definition had a bigger effect on deductions than those that were false as a matter of fact. The second experiment replicated the finding for determinate syllogisms, using problems in moods in which the status of the valid conclusion could not be altered by conversion of the premises. Beliefs accordingly appear to affect the process of reasoning rather than the interpretation of premises.

Memory & Cognition
1989, 17 (1), 11-17

The belief-bias effect in the production and evaluation of logical conclusions

HENRY MARKOVITS and GUILAINE NANTEL
Université du Québec à Montréal, Canada

In this study, we examined whether adult subjects' beliefs regarding the empirical truth of a conclusion affected their production as well as their evaluation of a logical conclusion in a reasoning task. In addition, the relation between the ability to resolve an abstract reasoning problem correctly and the effect of belief-bias was examined. The subjects were given one of four paper-and-pencil reasoning tasks, two of them using an evaluation paradigm, and two of them using a production paradigm. Each paradigm comprised either neutral problems or belief problems. The neutral problems were constructed to be as similar as possible to the belief problems, in order to control for extraneous factors. All four tasks also included an abstract reasoning problem. The results indicate a significant belief-bias effect for both the evaluation and the production tasks. Qualitative analysis indicated that the belief-bias effect was more pervasive in the production condition. In addition, the belief-bias effect was found to exist independently of the subjects' abstract reasoning ability. The results are discussed with reference to a two-stage model, in which belief is used to resolve uncertainties in inferentially produced conclusions.

One of the more interesting phenomena in the research on reasoning concerns the *belief-bias* effect. Several researchers have claimed that subjects tend to evaluate the logical validity of deductive arguments on the basis of their personal beliefs regarding the empirical status of the conclusion. Specifically, subjects will tend to rate an argument as valid if they think that the conclusion is empirically true, and vice versa, irrespective of the textbook validity of the argument. The reality of the belief-bias effect has been questioned (Revlin & Leirer, 1978; Revlin, Leirer, Yopp, & Yopp, 1980), mainly on the grounds that some of the effects observed may be attributable to conversion effects due to subjects' idiosyncratic encoding of premises. However, Evans, Barston, and Pollard (1983) have demonstrated a strong belief-bias effect in experiments designed to control for both conversion and atmosphere effects.

The belief-bias effect has generally been found in

et al.'s (1983) finding that subjects who accepted valid conclusions despite the influence of their beliefs referred more often to the premises than subjects who decided according to belief. One way of minimizing such a possibility and ensuring that subjects attend to the premises is to have subjects produce a conclusion. In other words, it may be argued that if the observed effects in evaluation tasks are due mainly to some process whereby subjects (mistakenly) examine only the conclusion of an argument, then forcing them to generate a conclusion from given premises should enable them to use their logical abilities more fully. Now, the effect of belief on the production of conclusions has been examined by Oakhill and Johnson-Laird (1985). They found that beliefs do tend to influence subjects' conclusions in a production task. However, they did not attempt to compare this effect with that found in an evaluation task. In addition, they compared subjects' performance with respect to believable as

LLMs

If Pigs Could Fly... Can LLMs Logically Reason Through Counterfactuals?

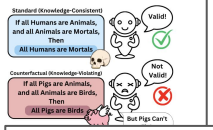
Ishwar B Balappanawar[1,*] Vamshi Krishna Bonagiri[1,*] Anish R Joishy[1,*]

Manas Gaur[2] Krishnaprasad Thirunarayan[3] Ponnurangam Kumaraguru [1]

[1]IIIT Hyderabad [2]University of Maryland, Baltimore County [3]Wright State University
ishwar.balappanawar@students.iiit.ac.in,{vamshi.b,anish.joishy}@research.iiit.ac.in
manas@umbc.edu, t.k.prasad@wright.edu, pk.guru@iiit.ac.in

Abstract

Large Language Models (LLMs) demonstrate impressive reasoning capabilities in familiar contexts, but struggle when the context conflicts with their parametric knowledge. To investigate this phenomenon, we introduce **CounterLogic**, a dataset containing 1,800 examples across 9 logical schemas, explicitly designed to evaluate logical reasoning through counterfactual (hypothetical knowledge-conflicting) scenarios. Our systematic evaluation of 11 LLMs across 6 different datasets reveals a con-



PNAS nexus
PNAS Nexus, 2024, 3, pgae233
https://doi.org/10.1093/pnasnexus/pgae233
Advance access publication: 16 July 2024
Research Report

Language models, like humans, show content effects on reasoning tasks

Andrew K. Lampinen[a,1], Ishita Dasgupta[a,†,1], Stephanie C. Y. Chan[a], Hannah R. Sheahan[b], Antonia Creswell[a], Dharshan Kumaran[a,c], James L. McClelland[a,d,e] and Felix Hill[a]

[a]Google DeepMind, Mountain View, CA, 94043 USA
[b]Google DeepMind, London N1C 4DN, UK
[c]Stanford University, Stanford, CA 94306, USA
[d]To whom correspondence should be addressed. Email: lampinen@google.com; id@google.com
[e]A.K.L. and I.D. contributed equally to this work.
Edited by Derek Abbott

Abstract

Abstract reasoning is a key ability for an intelligent system. Large language models (LMs) achieve above-chance performance on abstract reasoning tasks but exhibit many imperfections. However, human abstract reasoning is also imperfect. Human reasoning is affected by our real-world knowledge and beliefs, and shows notable "content effects"; humans reason more reliably when the semantic content of a problem supports the correct logical inferences. These content-entangled reasoning patterns are central to debates about the fundamental nature of human intelligence. Here, we investigate whether language models—whose prior expectations capture some aspects of human knowledge—similarly mix content into their answers to logic problems. We explored this question across three logical reasoning tasks: natural language inference, judging the logical validity of syllogisms, and the Wason selection task. We evaluate state of the art LMs, as well as humans, and find that the LMs reflect many of the same qualitative human patterns on these tasks—like humans, models answer more accurately when the semantic content of a task supports the logical inferences. These parallels are reflected in accuracy patterns, and in some lower-level features like the relationship between LM confidence over possible answers and human response times. However, in some cases the humans and models behave differently—particularly on the Wason task, where humans perform much worse than large models, and exhibit a distinct error pattern. Our findings have implications for understanding possible contributors to these human cognitive effects, as well as the factors that influence language model performance.

Keywords: language models, content effects, reasoning, logic, cognitive science

Significance Statement

Language models and humans both mix semantic content into their performance on logical reasoning problems, which generally results in greater success in familiar situations, but more errors in unusual ones. These results may inform the search for the origins of these human behaviors and may help improve applications of language models.

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

Abstract

The reasoning abilities of Large Language Models (LLMs) are becoming a central focus of study in NLP. In this paper, we consider the case of syllogistic reasoning, an area of deductive reasoning studied extensively in logic and cognitive psychology. Previous research has shown that pre-trained LLMs exhibit reasoning biases, such as *content effects*, avoid answering that *no conclusion follows*, display human-like difficulties, and struggle with multi-step reasoning. We contribute to this research line by systematically investigating the effects of chain-of-thought reasoning, in-context learning (ICL), and supervised fine-tuning (SFT) on syllogistic reasoning, considering syllogisms with conclusions that support or violate world knowledge, as well as ones with multiple premises. Crucially, we go beyond the standard focus on accuracy, with an in-depth analysis of the conclusions generated by the models. Our results suggest that the behavior of pre-trained LLMs

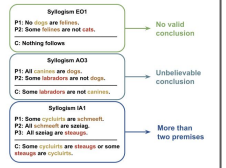


Figure 1: LLMs have difficulty with invalid inferences (Top); suffer from content effects (Middle); and struggle with longer chains of premises (Bottom). What is behind such weaknesses? Can LLMs learn to use only the form to draw deductively valid conclusions?

Controlled dataset

Abstract problem

- Categorical syllogisms
- 2 premises and 1 conclusion
- 64 possible combinations of premises
- 512 combinations of premises + conclusions

↓

Assigned meaning

- 10 triples of hierarchically organized terms
- e.g. “labradors”, “dogs”, “canines”
- 2 versions: plausible vs. implausible.

Plausible

Input

Premises:
All labradors are dogs.
All dogs are canines.

Conclusion:
All labradors are canines

Output

Valid

Implausible

Input

Premises:
All dogs are cats.
All cats are felines.

Conclusion:
All dogs are felines

Output

Valid

Behavioral performance

Model	Prompt	CE	Acc
Qwen2.5-32B-Instruct	Zero-shot	0.348	81.62
	CoT	0.095	94.61
Qwen3-14B	Zero-shot	0.213	86.54
	CoT	0.017	98.47
Qwen2.5-7B-Instruct	Zero-shot	0.418	75.68
	CoT	0.147	89.66
Qwen2.5-14B-Instruct	Zero-shot	0.361	77.47
	CoT	0.072	94.75
Qwen3-4B	Zero-shot	0.194	80.61
	CoT	0.003	97.90
Qwen3-8B	Zero-shot	0.218	85.97
	CoT	0.014	96.30
Qwen3-32B	Zero-shot	0.063	90.91
	CoT	0.064	95.64
Gemma3-4B-it	Zero-shot	0.213	81.02
	CoT	0.104	89.29
Gemma3-12B-it	Zero-shot	0.129	86.71
	CoT	-0.006	94.69
Gemma3-27B-it	Zero-shot	0.182	87.29
	CoT	0.021	97.47

$$\text{CE} = \frac{(\text{valid plausible} - \text{valid implausible}) + (\text{invalid implausible} - \text{invalid plausible})}{2}$$

- Content effects are still observed in all models in the zero-shot setting (except Qwen3-32B)
- With CoT, as model gets bigger, they become more accurate and less biased, with some models completing the task almost perfectly

Looking at the model internals

We looked at the internal representation of LLMs on two classification tasks:

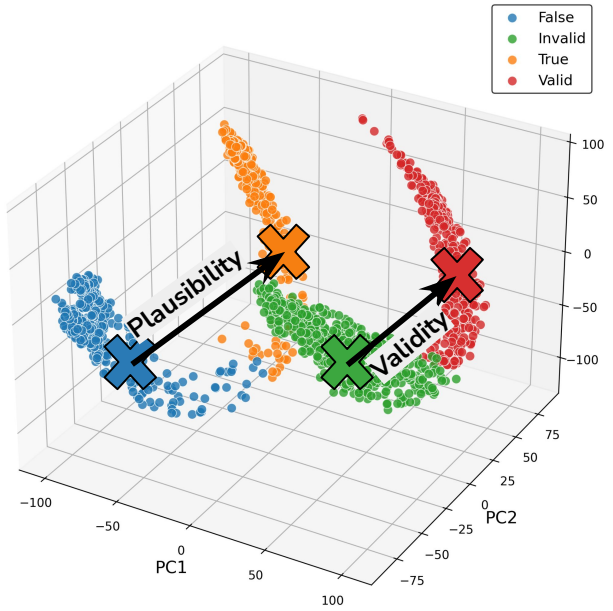
- Classify syllogisms as valid or invalid

```
Premise:
All labradors are dogs. All dogs are canines.
Conclusion:
All labradors are canines.

Is the syllogism valid or invalid? [VALID | INVALID]
```

- Classify sentences as true or false

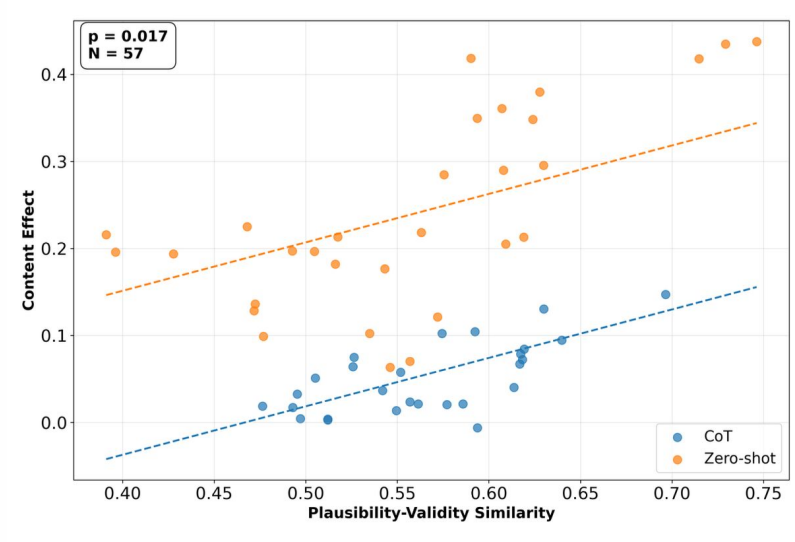
```
Sentence:
All labradors are canines.

Is the sentence true or false? [TRUE | FALSE]
```

We take a single representation of **plausibility** and **validity** as the **difference-in-means** vector between the positive and negative classes:



What can we learn about behavioral content effects?

- **Validity** and **plausibility** vectors are **highly similar** and their degree of similarity is **predictive** of observed behavioral content effects



- We can **control predictions** about **validity** using plausibility vectors, and we can control prediction about **plausibility** using validity vectors

Premises: All dogs are cats. All cats are felines.
Conclusion: All dogs are felines.

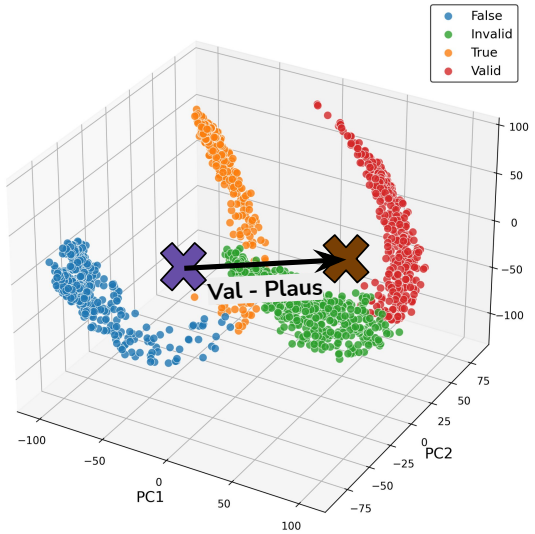
All dogs are felines.

Debiasing models towards systematicity

We compute a task difference vector that isolate those **dimensions of validity that are disentangled from plausibility**:

$$\mu^l_{V-P} = \mu^l_V - \mu^l_P$$

Adding μ^l_{V-P} to hidden states during validity classification should **push the representation away from plausibility-sensitive directions**, thereby reducing the influence of content effects.



Model	Metric	Orig.	Interv.
Qwen2.5-32B	Acc	81.62	82.21
	CE	0.348	0.072
Qwen3-14B	Acc	86.54	96.70
	CE	0.213	0.043

We improve accuracy and reduce CE at the same time!

Summing up

- We used **controlled data** disentangling content vs. logical form to demonstrate that content influences deductive reasoning in LLMs
- By investigating the internal representations we found that **validity and plausibility judgements are latently similar** and that they can **causally influence one another**
- This analysis led us to design an **intervention** that makes models reason more **systematically**

2. Learning logic through meta-learning

Teaching Small Language Models to Learn Logic through Meta-Learning

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

Correspondence: leonardo.bertolazzi@unitn.it

Abstract

Large language models (LLMs) are increasingly evaluated on reasoning tasks, yet their logical abilities remain contested. To address this, we study LLMs' reasoning in a well-defined fragment of logic: syllogistic reasoning. We cast the problem as premise selection and construct controlled datasets to isolate logical competence. Beyond evaluation, an open challenge is enabling LLMs to acquire abstract inference patterns that generalize to novel structures. We propose to apply few-shot meta-learning to this domain, thereby encouraging models to extract rules across tasks rather than memorize patterns within tasks. Although meta-learning has been little explored in the context of logic learnability, our experiments show that it is effective: small models (1.5B–7B) fine-tuned with meta-learning demonstrate strong gains in generalization, with especially pronounced benefits in low-data regimes. These meta-learned models outperform GPT-4o and o3-mini on our syllogistic reasoning task.

1 Introduction

With the advent of increasingly capable large language models (LLMs), logical reasoning has become a central domain for evaluating and comparing these systems (Huang and Chang, 2023; Mondorf and Plank, 2024; Liu et al., 2025). However,

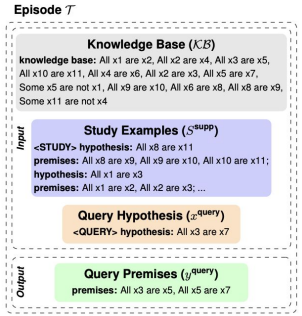


Figure 1: **Overview of a ML episode**. Given a set of premises (the knowledge base, $\mathcal{KB}$), a set of task demonstrations (or Study Examples), and a Query Hypothesis x^{query} that is entailed from $\mathcal{KB}$, models must generate the *minimal* subset of premises, the Query Premises y^{query}, from which x^{query} can be derived. During each ML episode, by being trained on the Study Examples, models learn to extract the abstract logical patterns. The examples show how we frame syllogistic inferences as a premise selection task. The dataset is built with pseudo-words, where here we have variables for space reasons.

2. Learning logic through meta-learning

Teaching Small Language Models to Learn Logic through Meta-Learning

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

Correspondence: leonardo.bertolazzi@unitn.it

Abstract

Large language models (LLMs) are increasingly evaluated on reasoning tasks, yet their logical abilities remain contested. To address this, we study LLMs' reasoning in a well-defined fragment of logic: syllogistic reasoning. We cast the problem as premise selection and construct controlled datasets to isolate logical competence. Beyond evaluation, an open challenge is enabling LLMs to acquire abstract inference patterns that generalize to novel structures. We propose to apply few-shot meta-learning to this domain, thereby encouraging models to extract rules across tasks rather than memorize patterns within tasks. Although meta-learning has been little explored in the context of logic learnability, our experiments show that it is effective: small models (1.5B–7B) fine-tuned with meta-learning demonstrate strong gains in generalization, with especially pronounced benefits in low-data regimes. These meta-learned models outperform GPT-4o and o3-mini on our syllogistic reasoning task.

1 Introduction

With the advent of increasingly capable large language models (LLMs), logical reasoning has become a central domain for evaluating and comparing these systems (Huang and Chang, 2023; Mondorf and Plank, 2024; Liu et al., 2025). However,

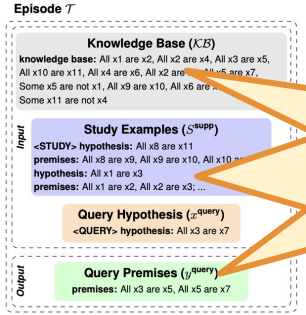


Figure 1: **Overview of a ML episode**. Given a set of premises (the knowledge base, $\mathcal{KB}$), a set of task demonstrations (or Study Examples), and a Query Hypothesis x^{query} that is entailed from $\mathcal{KB}$, models must generate the *minimal* subset of premises, the Query Premises y^{query}, from which x^{query} can be derived. During each ML episode, by being trained on the Study Examples, models learn to extract the abstract logical patterns. The examples show how we frame syllogistic inferences as a premise selection task. The dataset is built with pseudo-words, where here we have variables for space reasons.

Just accepted to EACL2026!

Meta-learning or “learning to learn”

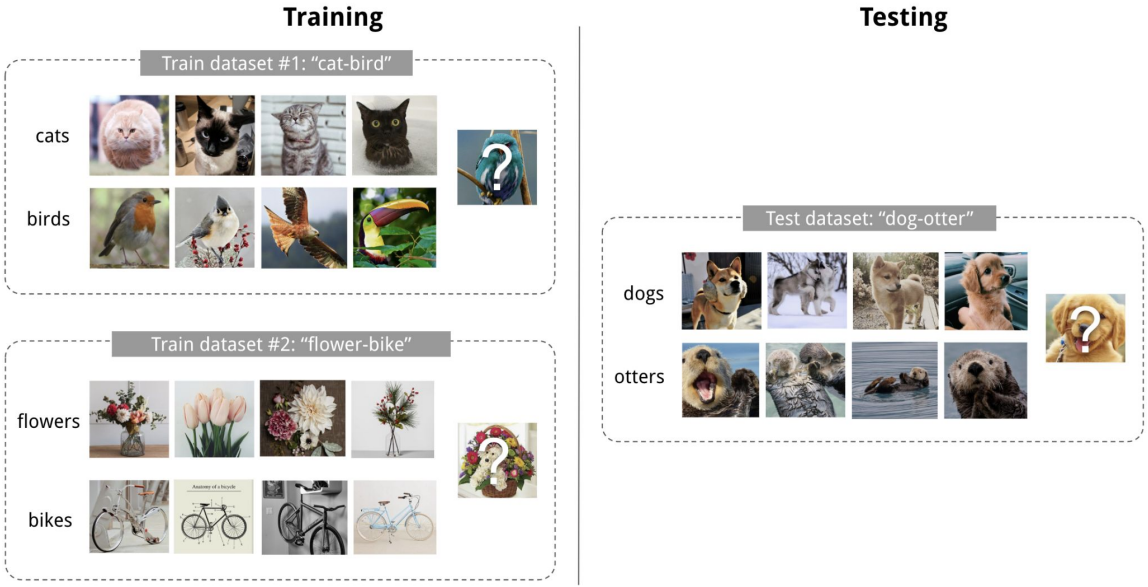
Meta-learning[5] methods aim to equip models with the capability of **adapting** or **generalizing** to new tasks and new environments that have never been encountered at training time

Supervised learning

- train dataset drawn from a **single data distribution**

Meta-learning

- models are trained over a **distribution of datasets** (tasks)

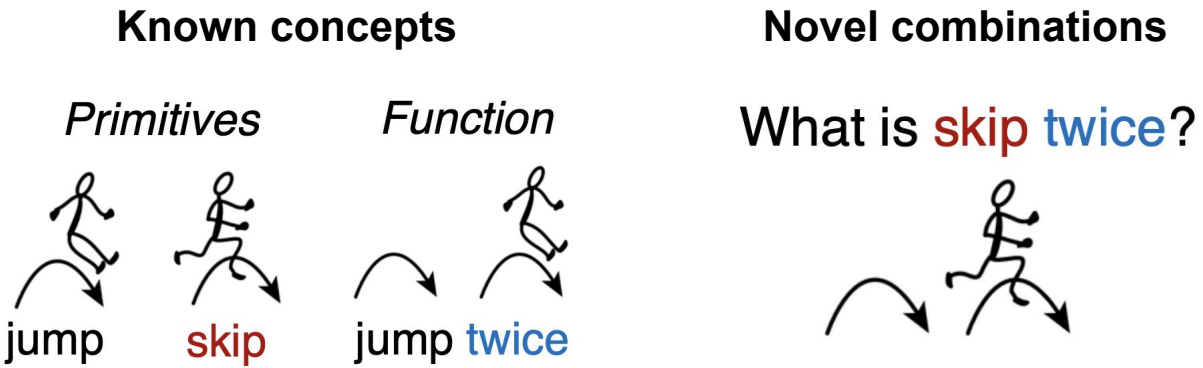


[5]T. Hospedales, A. Antoniou, P. Micaelli and A. Storkey, "Meta-Learning in Neural Networks: A Survey" in IEEE Transactions on Pattern Analysis & Machine Intelligence

Meta-learning for systematic generalization

Lake and Baroni[6] (2023) propose to use **few-shot meta-learning** as a way to induce **human-like systematic generalization** in neural networks inspired by how human can combine known concepts.

E.g. child learns how to 'skip' + knows the meaning of doing something *n* times
→ understand how to 'skip twice' or 'skip thrice' due to their **compositional skills**.



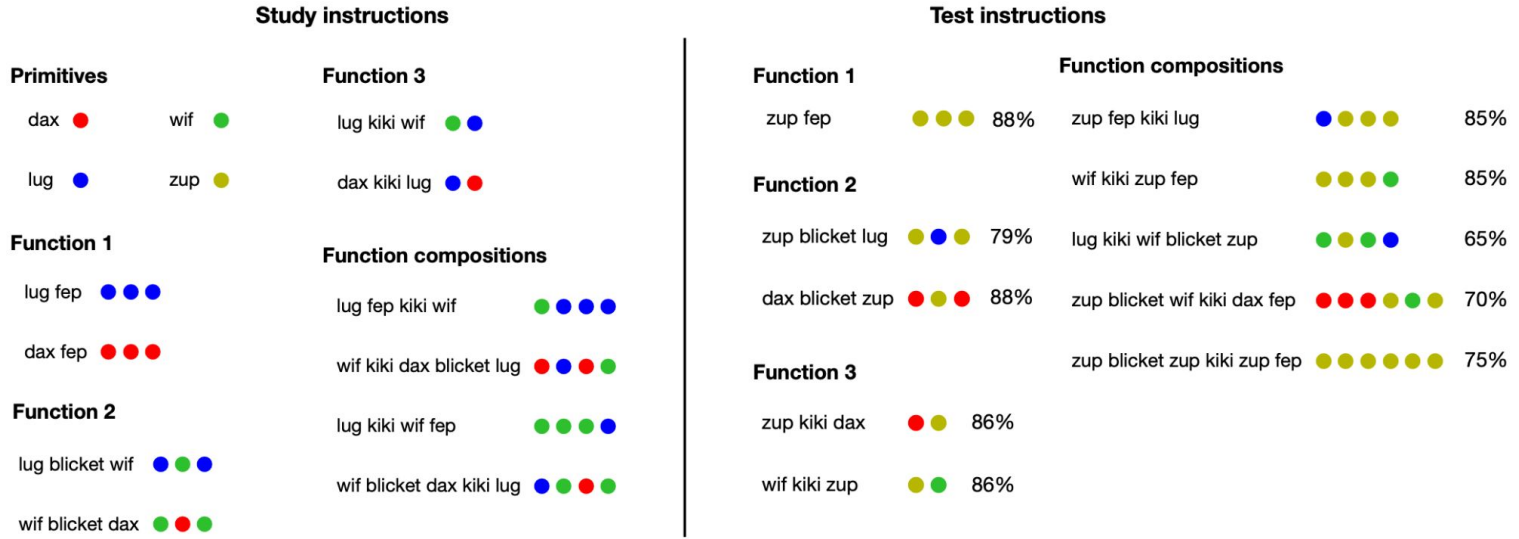
[6]Lake, B.M., Baroni, M. Human-like systematic generalization through a meta-learning neural network. Nature 623, 115–121 (2023)

* Images partially taken from: *Irie, K., Lake, B.M. Overcoming classic challenges for artificial neural networks by providing incentives and practice. Nat Mach Intell 7, 1602–1611 (2025).*

Meta-learning for compositionality (MLC)

Artificial setting: **input strings in a psedolanguage**, e.g. “dax” → **abstract symbols output**, e.g. <RED>.

Models can learn how primitives, functions, and function compositions are mapped to symbols in the **study instructions**, then they are given a new unseen set of **test instructions** and have to **infer their outputs.**



Meta-learning and logic

The systematicity criterion from Fodor and Pylyshyn was directly inspired by properties of **logic** and **formal languages**.

When we ask *"Can LLMs learn to reason logically?"* we are asking if by learning certain logical patterns they will **systematically generalize** to structurally related ones

→

Logic operates on formal structures, and superficially different expressions may share identical underlying structure

- e.g. *plausible syllogism == implausible syllogism*

Meta-learning has demonstrated human-like systematic generalization

***Open question*: Can it transfer to the systematic generalization required for logical systems?**

The premise selection task

In our experiments, we focus on the **syllogistic fragment of first-order logic**:

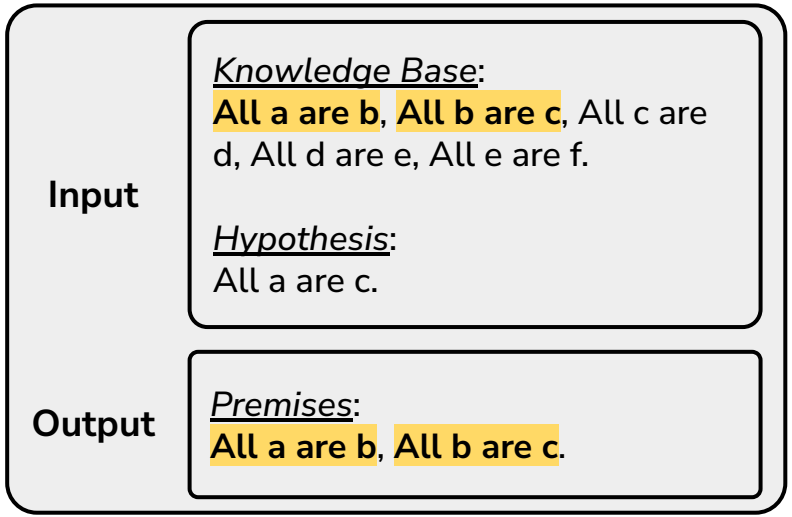
- Generalizes syllogisms to **more than two premises**
- Includes **seven** types of **minimal** syllogistic inferences

Task:

- Knowledge base of **atomic logical statements**.
- Models have to identify the **minimal subset of premises** that logically entail a given **test hypothesis**.

Core aspect of deductive reasoning: determining which known facts are **necessary** and **sufficient** to justify a conclusion.

Task example using a transitive inference – the simplest in the syllogistic logic



*We generate data with pseudowords in place of letters

Meta-learning setup

Learning methods (fine-tuning):

- **Meta-learning:** $p(y^{\text{query}} \mid x^{\text{query}}, S^{\text{supp}}, KB)$
- **Baseline:** $p(y^{\text{query}} \mid x^{\text{query}}, KB)$

Meta-learning episode

Knowledge Base:
All a are b, **All b are c**, All c are d, **All d are e**, **All e are f**, All f are g, **All g are h**, **All h are i**, **All i are j**, All j are k.

Study Examples:

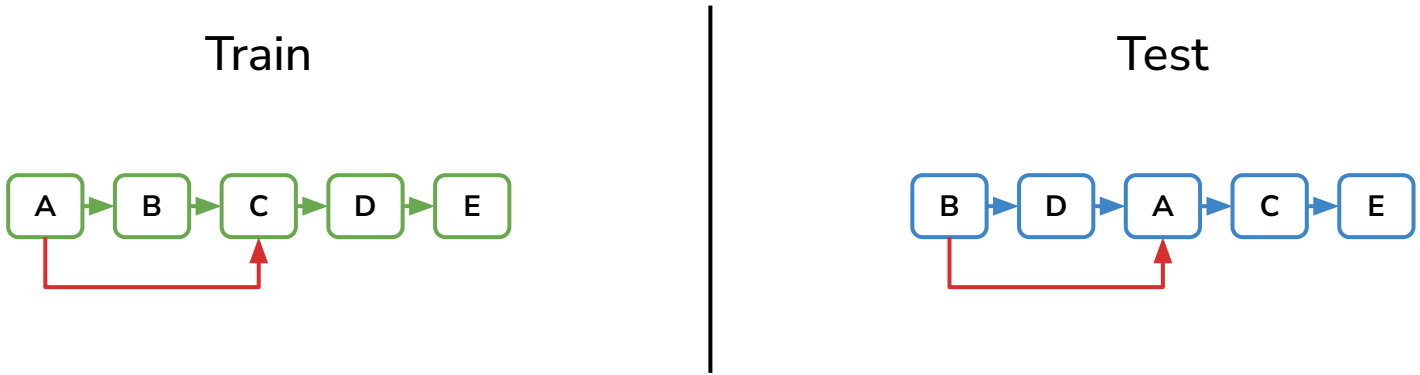
Hypothesis: All d are f. *Premises*: **All d are e**, **All e are f**.
Hypothesis: All g are j. *Premises*: **All g are h**, **All h are i**, **All i are j**.

Hypothesis: All a are c *Premises*: **All a are b**, **All b are c**.

The metal learning model can learn to **abstract the logical pattern** from the study examples and apply it to the test hypothesis

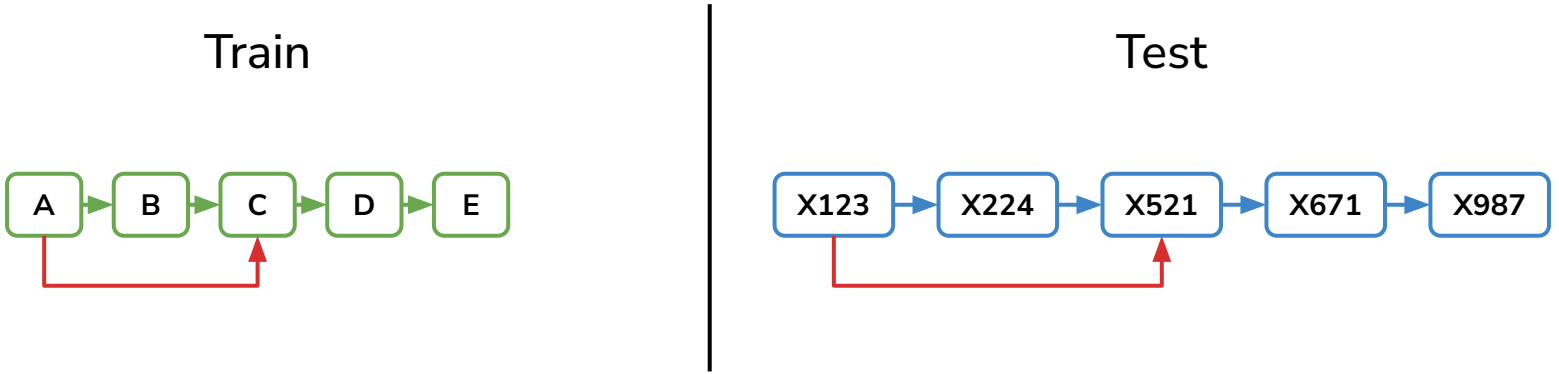
Types of systematic generalization in logic (I)

- **Core** → applying known inference types to novel unseen sets of premises using the same vocabulary as during training



Types of systematic generalization in logic (II)

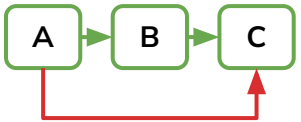
- **Lexical** → applying known inference types to novel unseen sets of premises with an unseen OOD vocabulary



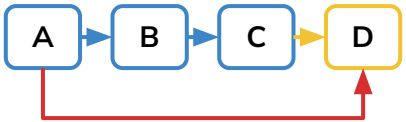
Types of systematic generalization in logic (III)

- **Recursive** → applying known inference types to more complex (longer) sets of premises than seen during training

Train



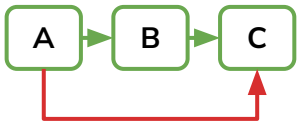
Test



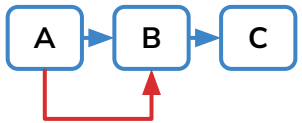
Types of systematic generalization in logic (IV)

- **Compositional** → applying known inference types to less complex (shorter) sets of premises than seen during training

Train



Test



Results

- **Core** → applying known inference types to novel unseen sets of premises with known vocabulary

	Model	Method	All
Fine-tuning	Qwen-2.5 1.5B	ML	93.11 ± 0.61
		Baseline_S	85.56 ± 1.24
	Qwen-2.5 3B	ML	96.16 ± 0.44
		Baseline_S	93.03 ± 1.15
	Qwen-2.5 7B	ML	98.13 ± 0.98
		Baseline_S	95.76 ± 1.10
Prompting	GPT-4o	Few-shot	39.76
		Zero-shot	15.90
	o3-mini	Few-shot	88.45
		Zero-shot	67.98

- ML models are always better than baseline but not by a large margin
- The advantage is bigger for smaller models
- We compare with bigger prompted models for a better understanding of the task difficulty

Results

- **Lexical** → applying known inference types to novel sets of premises with an unseen OOD vocabulary

Model	Type	Core	Unseen Pseudowords	Unseen Constants
Qwen-2.5 1.5B	ML	93.11 ± 0.61	93.15 ± 0.11	74.24 ± 1.07
	$\text{Baseline}_{\_S}$	85.56 ± 1.24	83.34 ± 1.90	38.49 ± 1.06
Qwen-2.5 3B	ML	96.16 ± 0.44	96.09 ± 0.30	83.21 ± 1.19
	$\text{Baseline}_{\_S}$	93.03 ± 1.15	91.49 ± 0.68	53.12 ± 2.03
Qwen-2.5 7B	ML	98.13 ± 0.98	98.03 ± 1.19	86.87 ± 0.31
	$\text{Baseline}_{\_S}$	95.76 ± 1.10	94.89 ± 1.55	57.81 ± 2.17

Results

- **Lexical** → applying known inference types to novel sets of premises with an unseen OOD vocabulary

Model	Type	Core	Unseen Pseudowords	Unseen Constants
Qwen-2.5 1.5B	ML	93.11 ± 0.61	93.15 ± 0.11	74.24 ± 1.07
	Baseline$_S$	85.56 ± 1.24	83.34 ± 1.90	38.49 ± 1.06
Qwen-2.5 3B	ML	96.16 ± 0.44	96.09 ± 0.30	83.21 ± 1.19
	Baseline$_S$	93.03 ± 1.15	91.49 ± 0.68	53.12 ± 2.03
Qwen-2.5 7B	ML	98.13 ± 0.98	98.03 ± 1.19	86.87 ± 0.31
	Baseline$_S$	95.76 ± 1.10	94.89 ± 1.55	57.81 ± 2.17

  - Unseen constants is the most OOD case, and ML is significantly more robust!

Results

- **Recursive** → applying known inference types to more complex (longer) sets of premises than seen during training
- **Compositional** → applying known inference types to less complex (shorter) sets of premises than seen during training

Model	Method	Recursive		Compositional	
		Disaligned	Aligned	Disaligned	Aligned
Qwen-2.5 1.5B	ML	76.42 ± 2.95	91.75 ± 1.10	70.94 ± 2.27	71.13 ± 1.83
	Baseline_{-S}	63.53 ± 1.16	63.53 ± 1.16	56.67 ± 1.22	56.67 ± 1.22
Qwen-2.5 3B	ML	87.61 ± 1.97	95.86 ± 0.70	77.19 ± 3.53	78.53 ± 1.71
	Baseline_{-S}	76.78 ± 1.63	76.78 ± 1.63	71.88 ± 1.49	71.88 ± 1.49
Qwen-2.5 7B	ML	90.03 ± 1.09	96.84 ± 0.15	76.23 ± 2.91	83.41 ± 1.63
	Baseline_{-S}	80.76 ± 2.65	80.76 ± 2.65	71.08 ± 1.55	71.08 ± 1.55

- ML models are always better than baseline

Results

- **Recursive** → applying known inference types to more complex (longer) sets of premises than seen during training
- **Compositional** → applying known inference types to less complex (shorter) sets of premises than seen during training

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

- ML models are always better than baseline
- Recursive case is easier than the compositional one

Results

- **Recursive** → applying known inference types to more complex (longer) sets of premises than seen during training
- **Compositional** → applying known inference types to less complex (shorter) sets of premises than seen during training

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

- In the aligned case study examples have same answer length as the query
- ML models can learn from simpler or more complex inference in-context

Summing up

- This work is **foundational** in the sense that it asks about the learnability of logic in a systematic way in a neural system (no neuro-symbolic approach!)
- In **core generalization**, meaning that we test how learned inferences are applied to an unseen set of premises using known vocabulary, baseline models almost approach meta-learning models
- **Meta-learning** is most effective when there is a **large distributional shift** (abstract lexicon, recursive or compositional generalization) – this is what matters most for **systematicity** in logical reasoning!
- Future work should investigate how to generalize this approach to more **naturalistic** settings

Conclusions

Take-home message: I believe that thinking about LLMs' reasoning capabilities inspired by what we know about human reasoning capabilities and limitations can guide approaches to make LLMs more robust reasoners.

Approach 1: LLMs have similar reasoning biases to humans → we can engineer ways of debiasing models to make them more systematic.

Approach 2: Human experts can think about logical problems in a systematic way → we can investigate new learning methods, such as meta-learning, to mimic this capability in LLMs.

Presented papers

1. Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. 2025. **How Language Models Conflate Logical Validity with Plausibility: A Representational Analysis of Content Effects.** Under review.

2. Leonardo Bertolazzi , Manuel Vargas Guzmán , Raffaella Bernardi , Maciej Malicki , Jakub Szymanik. 2026. **Teaching Small Language Models to Learn Logic through Meta-Learning.** In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics.