Large Language Models through the lens of Logic, Language and Information: An ESSLLI Journey

Raffaella Bernardi

Free University of Bozen-Bolzano

ESSLLI 2026, Prague

unibz

- ESSLLI 2026, 37th European Summer School in Logic, Language and Information, 1 – 14 August, Prague, Czech Republic.
- ESSLLI 2025, 36th European Summer School in Logic, Language and Information, 28 July – 8 August, Ruhr University Bochum, Germany.
- ESSLLI 2024, 35th European Summer School in Logic, Language and Information, 29 July – 9 August 2024, Leuven, Belgium.
- ESSLLI 2023, 34th European Summer School in Logic, Language and Information, 31 July – 11 August 2023, Ljubljana, Slovenia.
- ESSLLI 2022, 33rd European Summer School in Logic, Language and Information, 8 August – 19 August 2022, Galway, Ireland.
- ESSLLI 2021, 32nd European Summer School in Logic, Language and Information, 26 July – 13 August 2021, Held virtually.
- ESSLLI 2020 was due to to take place 3 to 14 August, 2020, Utrecht, The Netherlands. Postponed due to COVID-19.
- ESSLLI 2019, 31st European Summer School in Logic, Language and Information, 5 to 16 August, 2019, Riga, Latvia
- ESSLLI 2018, 30th European Summer School in Logic, Language and Information, 6 to 17 August, 2018, Sofia, Bulgaria
- ESSLLI 2017, 29th European Summer School in Logic, Language and Information, 17 to 28 July, 2017, Toulouse, France
- ESSLLI 2016, 28th European Summer School in Logic, Language and Information, 15 to 26 August, 2016, Bolzano-Bozen, Italy
- ESSLLI 2015, 27th European Summer School in Logic, Language and Information, 3 to 14 August, 2015, Barcelona, Spain
- ESSLLI 2014, 26th European Summer School in Logic, Language and Information, 11 to 22 August, 2014 Tübingen, Germany
- ESSLLI 2013, 25th European Summer School in Logic, Language and Information, 5 to 16 August, 2013 in Düsseldorf, Germany
- ESSLLI 2012, 24rd European Summer School in Logic, Language and Information, 6 to 17 August, 2012 in Opole, Poland
- ESSLLI 2011, 23rd European Summer School in Logic, Language and Information, 1 to 12 August, 2011 in Ljubljana, Slovenia
- ESSLLI 2010, 22st European Summer School in Logic, Language and Information, 9 to 20 August 2010. Copenhagen, Denmark
- ESSLLI 2009, 21st European Summer School in Logic, Language and Information, 20 to 31 July 2009. Bordeaux, France
- ESSLLI 2008, 20th European Summer School in Logic, Language and Information, 4 to 15 August 2008. Hamburg, Germany
- ESSLLI 2007, 19th European Summer School in Logic, Language and Information, 6 to 17 August 2007. Dublin, Ireland
- ESSLLI 2006, 18th European Summer School in Logic, Language and Information, 31 July to 11 August 2006. Málaga, Spain
- ESSLLI 2005, 17th European Summer School in Logic, Language and Information, 8 to 19 August 2005. Edinburgh, Scotland
- ESSLLI 2004, 16th European Summer School in Logic, Language and Information, 9 to 20 August 2004. Nancy, France
- ESSLLI 2003, 15th European Summer School in Logic, Language and Information, 18 to 29 August 2003. Vienna, Austria
- ESSLLI 2002, 14th European Summer School in Logic, Language and Information, 5 to 16 August 2002. Trento, Italy
- ESSLLI 2001, 13th European Summer School in Logic, Language and Information, 13 to 24 August 2001. Helsinki, Finland
- ESSLLI 2000, 12th European Summer School in Logic, Language and Information, 6 to 18 August 2000. Birmingham, United Kingdom
- ESSLLI 1999, 11th European Summer School in Logic, Language and Information, 9 to 20 August 1999. Utrecht, Netherlands
- ESSLLI 1998, 10th European Summer School in Logic, Language and Information, 17 to 28 August 1998. Saarbrücken, Germany
- ESSLLI 1997, 9th European Summer School in Logic, Language and Information, 11 to 22 August 1997. Aix en Provence, France
- ESSLLI 1996, 8th European Summer School in Logic, Language and Information, 12 to 23 August 1996. Prague, The Czech Republic
- ESSLLI 1995, 7th European Summer School in Logic, Language and Information, 14 to 25 August 1995. Barcelona, Spain
- ESSLLI 1994, 6th European Summer School in Logic, Language and Information, 8 to 19 August 1994. Copenhagen, Denmark
- ESSLLI 1993, 5th European Summer School in Logic, Language and Information, 16 to 27 August 1993. Lisbon, Portugal
- ESSLLI 1992, 4th European Summer School in Logic, Language and Information, 17 to 28 August 1992. Essex, United Kingdom
- ESSLLI 1991, 3rd European Summer School in Logic, Language and Information, 12 to 23 August 1991. Saarbrücken, Germany
- ESSLLI 1990, 2nd European Summer School in Logic, Language and Information, 13 to 24 August 1990. Leuven, Belgium
- ESSLLI 1989, 1st European Summer School in Logic, Language and Information, 14 to 25 August 1989. Groningen, The Netherlands.

My first ESSLLI

Philosophy of Language

Claudia Casadio

- Frege (1891, 1892): Verbs are **unsaturated functions**, they require to be completed by an argument.
- Frege, G. (1892). *Über Sinn und Bedeutung* ("On Sense and Reference").

The Morning Star & *The Evening Star* different senses **(Sinn)**

Venus the same reference **(Bedeutung)**

Propositional **Logic**

P	Q	R	$P \to Q$	$Q \to R$	$(P \to Q) \wedge (Q \to R)$
T	T	T	T	T	T
T	T	F	T	F	F
T	F	T	F	T	F
T	F	F	F	T	F
F	T	T	T	T	T
F	T	F	T	F	F
F	F	T	T	T	T
F	F	F	T	T	T

Aristotle' syllogism

P1 All men are mortal
P2 Socrates is men
C: Socrates is mortal

Language & Reasoning interplay ❓

Logic & Language

Word	Category	Semantics
all	$(S/(S\backslash NP))/N$	$\lambda P.\lambda Q.\forall x.(P(x) \rightarrow Q(x))$
no	$(S/(S\backslash NP))/N$	$\lambda P.\lambda Q.\forall x.(P(x) \rightarrow \neg Q(x))$

Michael Moortgat

ESSLLI lectures ⊆ interesting lectures

All good AI researchers (attended ESSLLI lectures)$^{+}$

All good AI researchers attended interesting lectures

No bad AI researchers (attended interesting lectures)$^{-}$

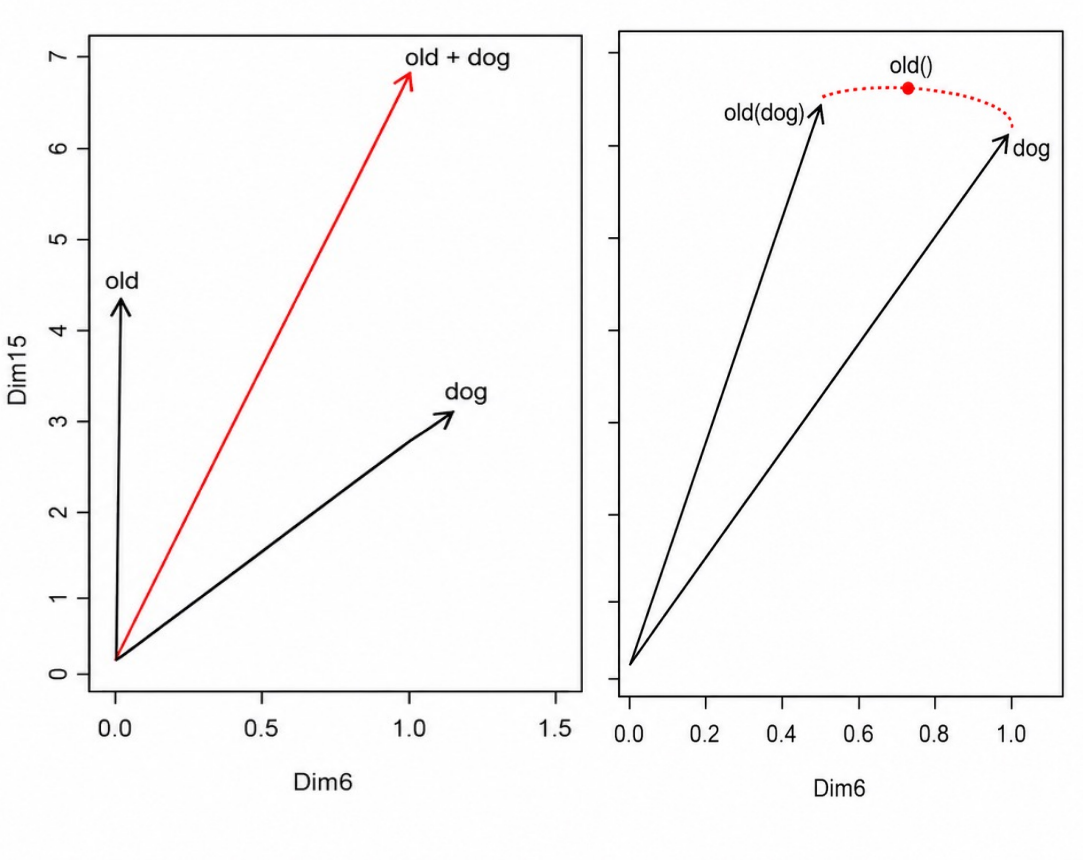
No bad AI researchers attended ESSLLI lectures

PhD Thesis, Utrecht University, 2002

Logic & Language

→ Logic Representations (NLU)
→ Built compositionally: logic rules
→ Reasoning

Core: **Logical Words** and **Syntax-Semantics interface**



Frege in Space, Baroni, Bernardi, Zamparelli 2014

M. Baroni R. Zamparelli

Distributional Semantics

→ Vector Representations (NLU)
→ Beyond Lexicon: algebraic operations
→ Reasoning (SICK dataset)

Core: **Content words**

GuessWhat?!

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Our model

1. is it a donut?	yes
2. is it on the left?	no
3. is it on the right?	yes
4. is it the whole donut?	yes
5. does it have pink icing?	no
6. is it touching the donut with the sprinkles?	yes
7. does it have chocolate icing?	no
8. is it the third donut?	yes

RL model

1. is it food?	yes
2. is it a donut?	yes
3. is it in left?	no
4. is it in top?	yes
5. is it in top?	yes
6. is it in top?	yes
7. is it in top?	yes
8. is it top?	yes

Task Success ➔ **Quality of the dialogues**

Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat, NAACL 2019
Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, Raquel Fernández

B. Plank R. Fernández

Visually grounded conversation

→ Vector representations
→ Sequence of polar Q-A turns (NLG and NLU)
→ Visually Grounded Reasoning

Core: **Model architecture** and learning paradigm

From Formal Semantics to Conversational Vision Language Models

Logic & Language (NLU)

→ Logic Representation
→ Built compositionally
→ Logic Reasoning

Core: **Logical words**

Distributional Semantics (NLU)

→ Vector Representation
→ Beyond Lexicon
→ Reasoning

Core: **Content words**

Visually grounded conversations (NLU & NLG)

→ Vector representation
→ Sequence of polar Q-A turns
→ Visually Grounded Reasoning

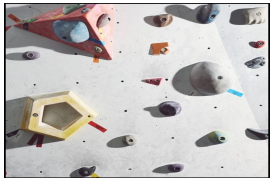
Core: **Model architecture** and learning paradigm

First lesson I learned

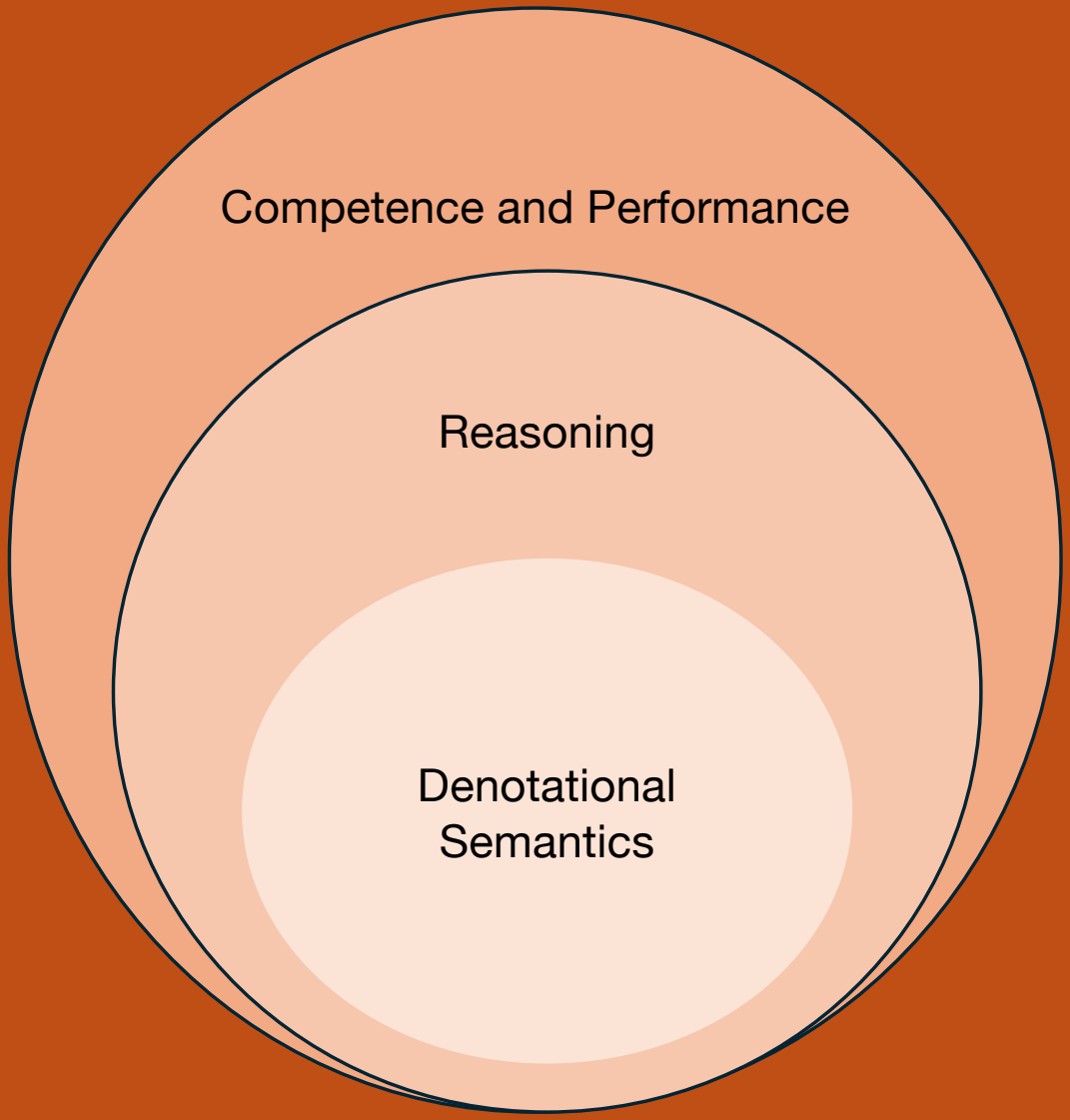
There is no pre-defined path

Research is driven by curiosity

While doing research in a team, you build for ever lasting relations



LARGE LANGUAGE MODELS



Denotational Semantics

Key points of Denotational Semantics

The meaning of an expression is the **object it denotes**.

Expression	**Denotation**
John	an entity
runs	a set of entities
every	a generalized quantifier (a set of sets of entities)
John runs	a truth value

- Pretty good in representing Logical Words: e.g. negation and quantifiers
- Pretty good in accounting for Scope Ambiguity

(1) a. Every farmer owns a donkey.

b. Surface Scope: $\forall y[farmer(y) \rightarrow \exists x[donkey(x) \wedge owns(y,x)]]$

c. Inverse Scope: $\exists x[donkey(x) \wedge \forall y[farmer(y) \rightarrow owns(y,x)]]$

➔ LLMs and Denotational Semantics key points

Denotational Semantics

Visually Grounded Verification Tasks: **True vs False**

Ravi Shekhar

a) People riding bicycles down the road approaching a dog

b) People riding bicycles down the road approaching a bird

True	False
	X
X	

ACL 2017: Models were at *chance level* on FOIL-it

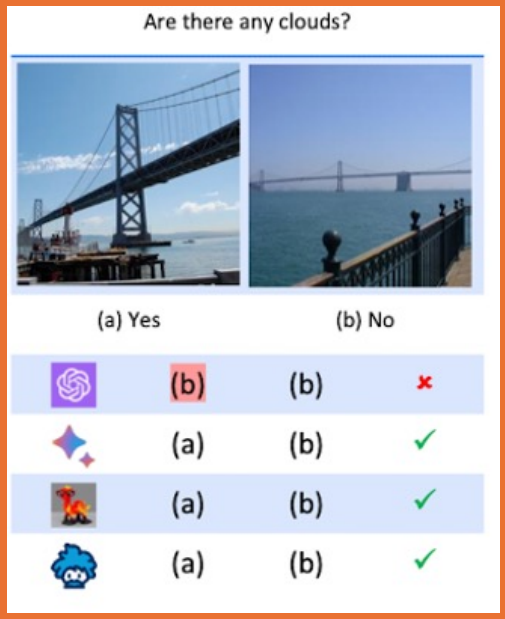
CLIP: 89%

➔SOTA good in grounding **nouns**

➔SOTA still struggles with **verbs**

Parcalabescu et al (Deep Mind) ACL 2022

Bugliarello et al 2023: on verbs SOTA 90% but they fail with **spatial relations**



VLMs struggle with carefully designed verification tasks

CVPR 2024: Tong, Liu, Zhai, Ma, LeCun, Xie

Denotational Semantics

Logical Words: Visually grounded **negation**


Claudio Greco


Alberto Testoni

Questioner **Oracle**

1. Is it an object? No
2. Is it a person? Yes

3. Does he have his right arm on the other's shoulder? No

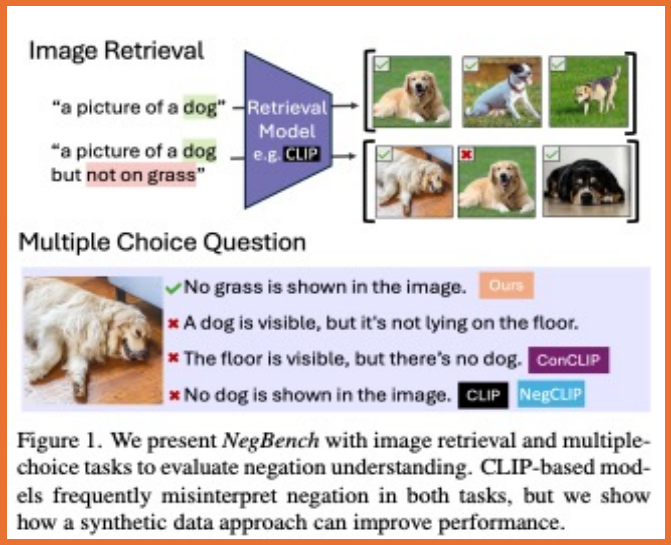


Figure 1. We present *NegBench* with image retrieval and multiple-choice tasks to evaluate negation understanding. CLIP-based models frequently misinterpret negation in both tasks, but we show how a synthetic data approach can improve performance.

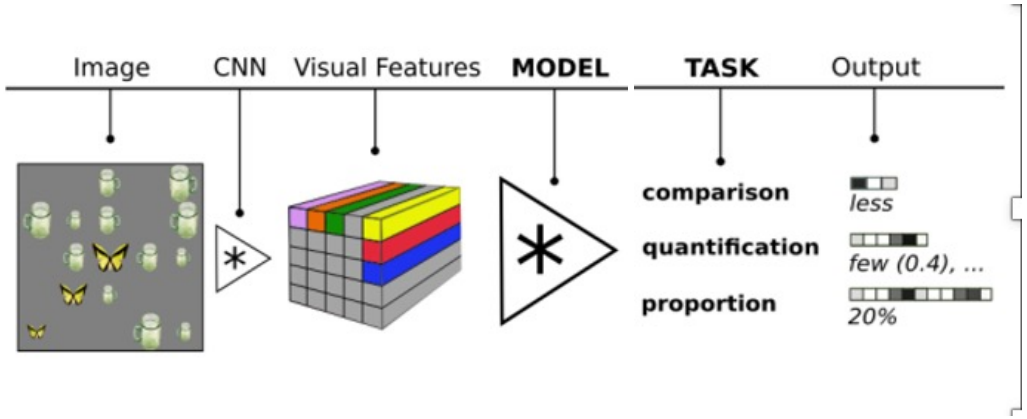
Vision-Language Models Do Not Understand Negation, Alhamoud et al (MIT, OpenAI, Un. Oxford) **CVPR 2025**

Testoni, Greco, Bernardi
Artificial Intelligence Models Do Not Ground Negation, Humans Do. GuessWhat?! Dialogues as a Case Study
Front. Big Data, 2022

Denotational Semantics

Logical Words: Visually grounded **quantifiers**

Sandro Pezzelle



Pezzelle et al Cognition 2018
Pezzelle et al NAACL 2018

VLMs, **like humans, are influenced by object count** in vague quantifier use. However, we find significant inconsistencies across models in different evaluation settings, suggesting **that judging and producing vague quantifiers rely on two different processes.**

Wong, Nouwen, Gatt, **ACL 2025**
VAQUUM: Are Vague Quantifiers Grounded in Visual Data?

Denotational Semantics

Logical Words: Generalized quantifiers
Scope Ambiguity

(1) a. Every farmer owns a donkey.

b. Surface Scope: $\forall y[farmer(y) \rightarrow \exists x[donkey(x) \wedge owns(y,x)]]$

c. Inverse Scope: $\exists x[donkey(x) \wedge \forall y[farmer(y) \rightarrow owns(y,x)]]$

Using these datasets, we find evidence that several models

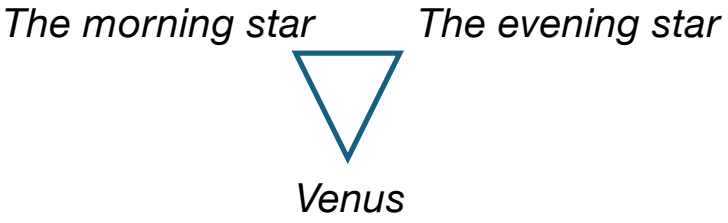
(i) are **sensitive to the meaning ambiguity** in these sentences, in a way that **patterns well with human judgments**, and

(ii) can successfully identify **human-preferred readings** at a high level of accuracy (over 90% in some cases).

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, Siva Reddy Scope Ambiguities in Large Language Models. **TACL 2024**

Denotational Semantics

Conclusion on Denotation

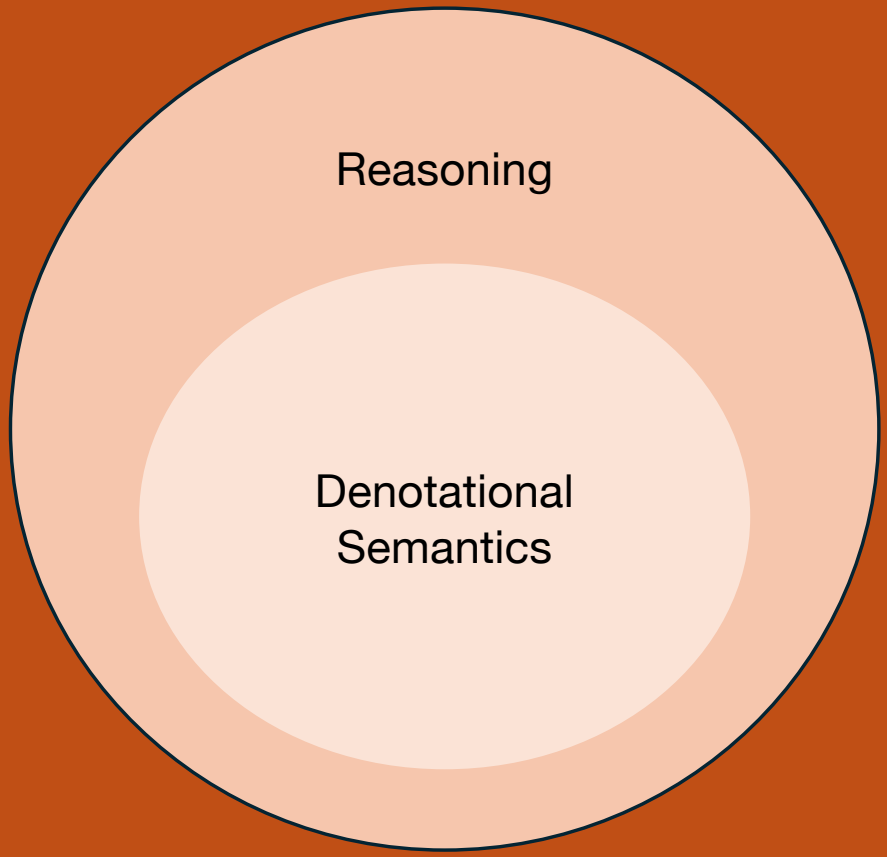


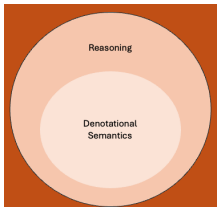
There is no semantics without denotation.

A. It is either an object or a person.
B. Is it an object? No.
C. Therefore it's a person.

LLMs may have learned to align with humans' preferred interpretations of logical words, **but we cannot yet claim that they truly master them.**

LARGE LANGUAGE MODELS





Trends in
Cognitive Sciences

CellPress

Feature Review

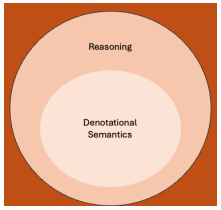
Dissociating language and thought in large language models

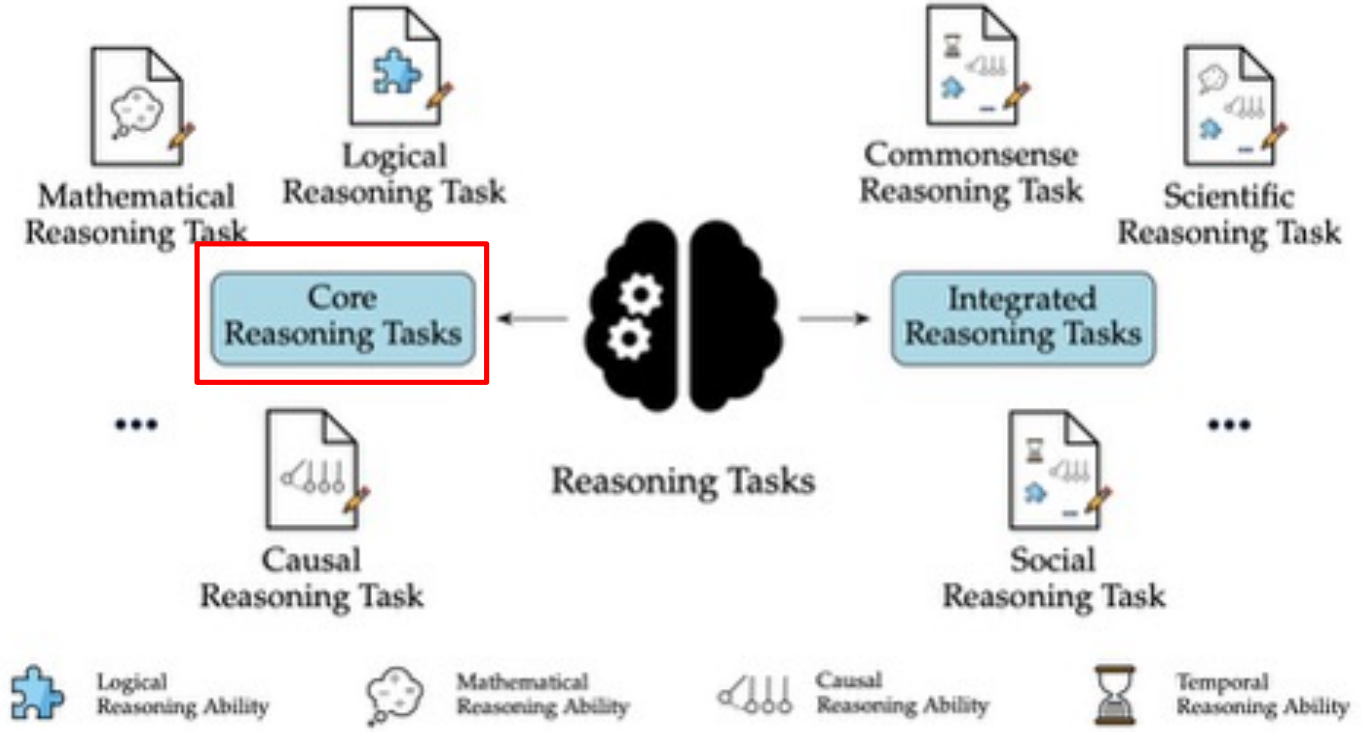
Kyle Mahowald,[1,5,*] Anna A. Ivanova,[2,5,*] Idan A. Blank,[3,*] Nancy Kanwisher,[4,*] Joshua B. Tenenbaum,[4,*] and Evelina Fedorenko[4,*]

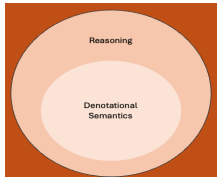
➔ LLMs should be taken seriously as models of **formal linguistic skills**

➔ 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.**

Back to my interest:
Language and Reasoning interplay.


Reasoning
Denotational Semantics


Mathematical Reasoning Task
Logical Reasoning Task
Commonsense Reasoning Task
Scientific Reasoning Task
Core Reasoning Tasks
Integrated Reasoning Tasks
Reasoning Tasks
Causal Reasoning Task
Social Reasoning Task
Logical Reasoning Ability
Mathematical Reasoning Ability
Causal Reasoning Ability
Temporal Reasoning Ability



Large Scale Benchmarks to evaluate LLMs reasoning ability

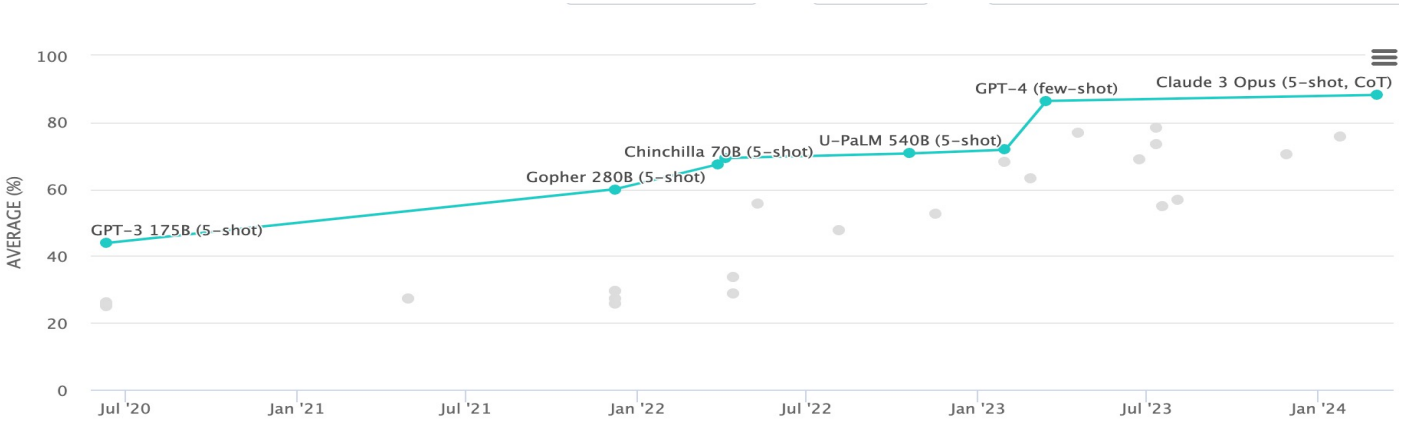
GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it.Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

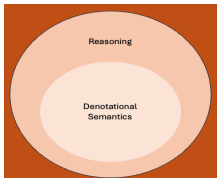
Let T be the number of bouncy balls in the tube.
After buying the tube of balls, Sophie has 31+8+9+ T = 48 + T =62 toys for her nephew.
Thus, T =62-48 = <<62-48=14>>14 bouncy balls came in the tube.

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



Cobbe et al, 2021



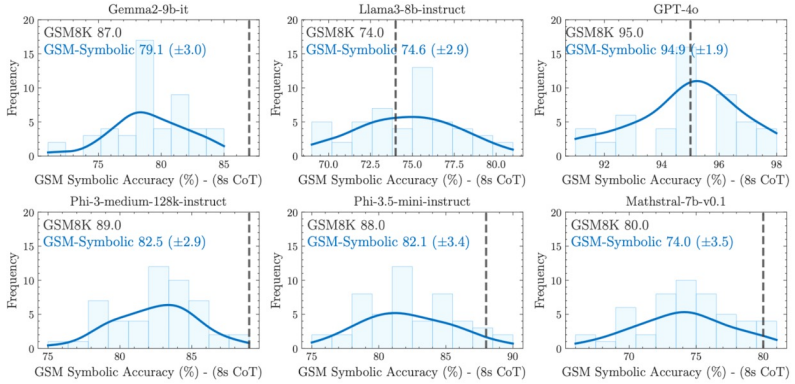
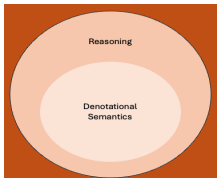


Figure 2: The distribution of 8-shot Chain-of-Thought (CoT) performance across 50 sets generated from `GSM-Symbolic` templates shows significant variability in accuracy among all state-of-the-art models. Furthermore, for most models, the average performance on `GSM-Symbolic` is lower than on `GSM8K`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

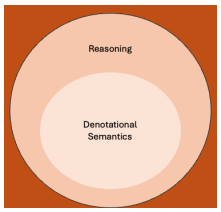


Massive Benckmarks

Open LLM Leaderboard Archived

Comparing Large Language Models in an **open** and **reproducible** way

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU-...	CO_2 Cost
	1	🔶	MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %	66.01 kg
	2	💬	MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %	64.44 kg
	3	💬	dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %	25.99 kg
	4	💬	MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %	25.95 kg
	5	🔶	huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %	50.41 %	76.77 kg
	6	💬	Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %	47.65 kg
	7	💬	MaziyarPanahi/calme-2.1-qwen2.5-72b	47.86 %	86.62 %	61.66 %	59.14 %	15.10 %	13.30 %	51.32 %	29.50 kg
	8	🔶	newsbang/Homer-v1.0-Qwen2.5-72B	47.46 %	76.28 %	62.27 %	49.02 %	22.15 %	17.90 %	57.17 %	29.55 kg
	9	💬	ehristoforu/qwen2.5-test-32b-it	47.37 %	78.89 %	58.28 %	59.74 %	15.21 %	19.13 %	52.95 %	29.54 kg
	10	🔶	Saxo/Linkbricks-Horizon-AI-Avengers-V1-32B	47.34 %	79.72 %	57.63 %	60.27 %	14.99 %	18.16 %	53.25 %	7.95 kg



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

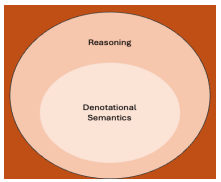
P1: All siameses are **cats**
P2: Some felines are not **cats**
C: Some felines are not siameses

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

Syllogisms are 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

Bertolazzi et. al. 2024. *A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences*. EMNLP 2024.
Eisape et al 2024 A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models, NAACL 2024
Lampinen et. al. 2024. *Language models, like humans, show content effects on reasoning tasks*. PNAS Nexus.



LLMs do not treat syllogisms formally

Syllogism AO3

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

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.

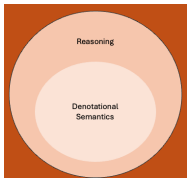
Syllogism EO1

P1: No dogs are felines.
P2: Some felines are not cats.
———————————————
C: Nothing follows

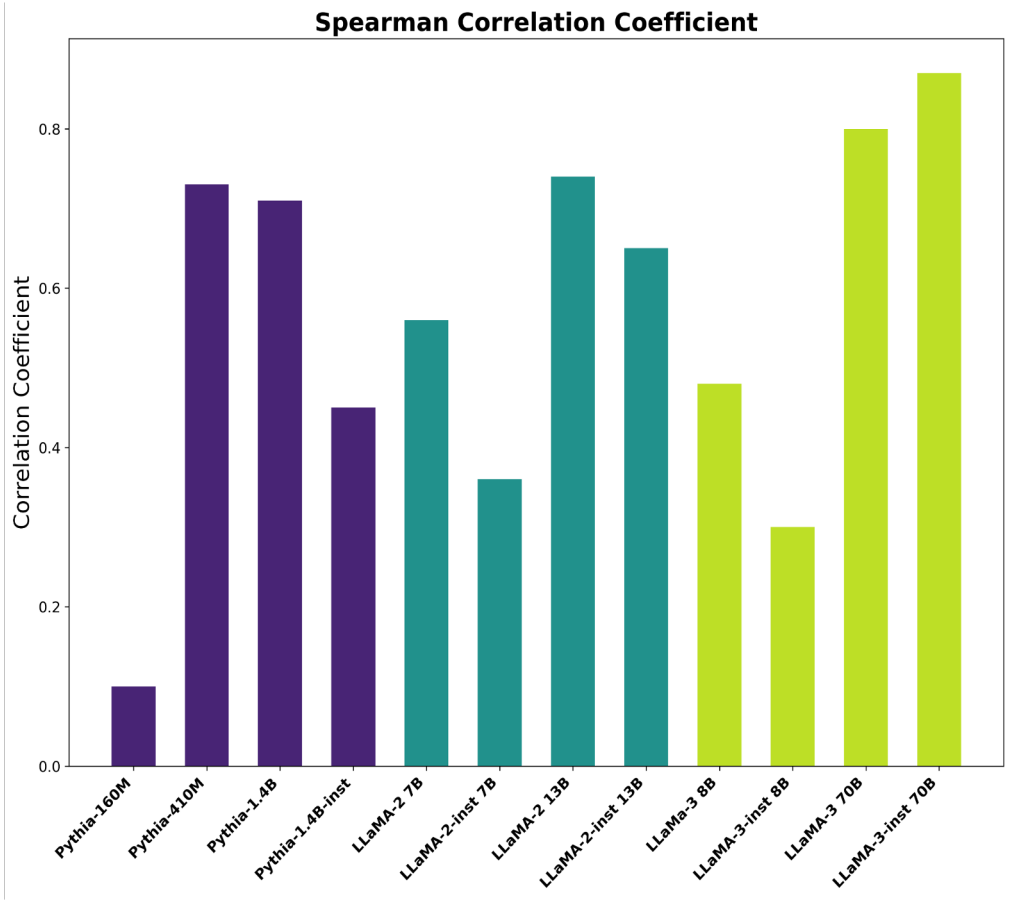
Leaderboard approach

Model	Believable			Unbelievable	Content Effect
	Acc	Invalid	Valid	Valid	Difference
Pythia-160M	0.16	0.00	0.37	0.00	-100.00%*
Pythia-410M	9.53	0.00	22.59	19.63	-13.10%
Pythia-1.4B	14.22	1.94	31.11	33.33	+7.14%
Pythia-1.4B-inst	17.03	15.28	19.26	15.93	-17.29%*
LLaMA-2 7B	24.38	3.06	53.70	47.78	-11.02%
LLaMA-2-inst 7B	19.69	10.83	32.22	19.26	-40.22%
LLaMA-2 13B	22.81	0.28	53.70	40.37	-24.82%
LLaMA-2-inst 13B	22.34	8.06	42.22	35.93	-14.19%*
LLaMa-3 8B	30.31	0.83	70.74	52.59	-10.44%
LLaMA-3-inst 8B	50.00	30.28	78.15	64.81	-17.07%
LLaMA-3 70B	30.63	10.28	58.89	54.07	-8.18%
LLaMA-3-inst 70B	41.25	32.50	54.44	57.41	+5.46%

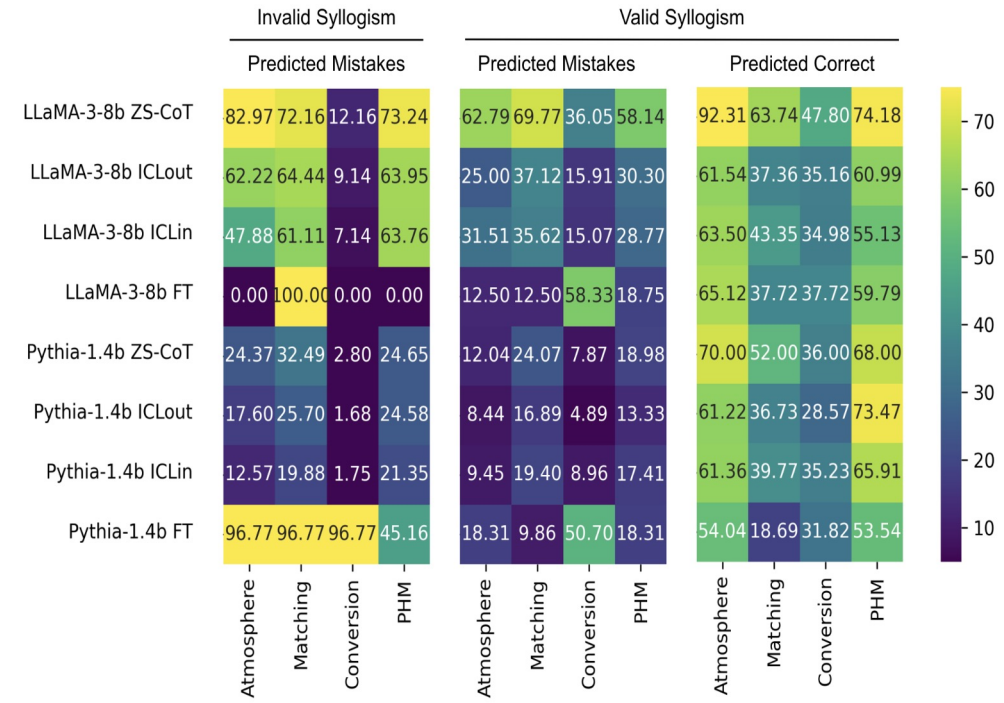
Bertolazzi et. al. 2024. *A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences*. EMNLP 2024



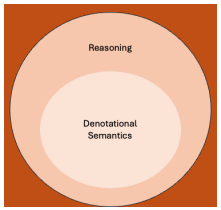
LLMs and Humans' reasoning



Human data from: Khemlani and Johnson-Laird 2012



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. We hypothesize **that they rely on a shallow pattern-matching strategy, using quantifiers as cues.**



Post-training: in-context learning (ICL) or supervised finetuning (SFT)?

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

Premises + Options

LLM

Some felines are not siameses. Nothing follows.

SFT

Premises + Options

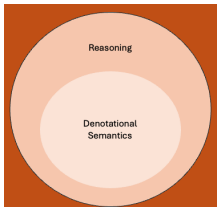
LLM

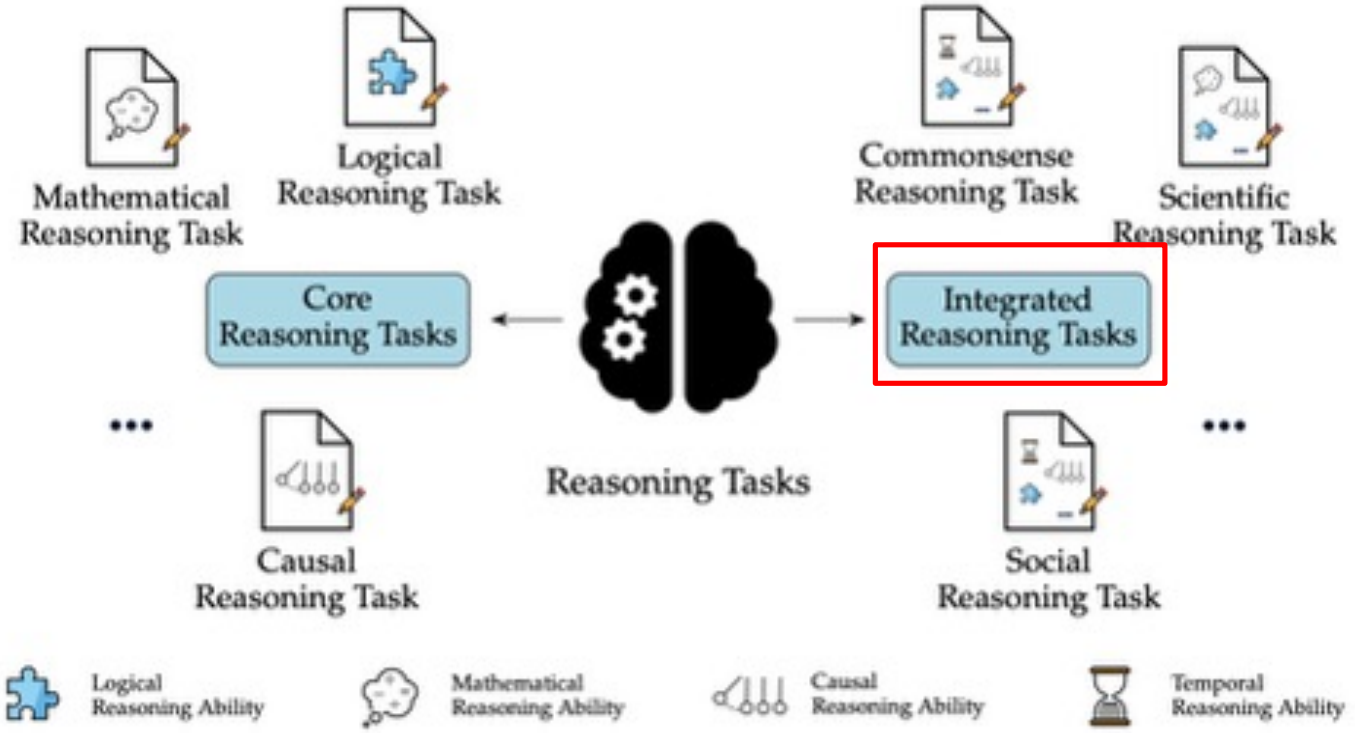
Some felines are not siameses. Some felines are siameses

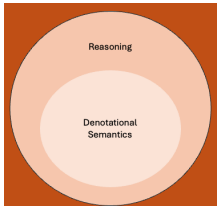
Our results suggest that

- ✓ ICL does not mitigate reasoning biases
- ✓ Through SFT models better learn to dissociate *form* from *meaning*, but not completely yet.

ICL examples/SFT training: pseudowords


Reasoning
Denotational Semantics


Mathematical Reasoning Task
Logical Reasoning Task
Commonsense Reasoning Task
Scientific Reasoning Task
Core Reasoning Tasks
Integrated Reasoning Tasks
Reasoning Tasks
Causal Reasoning Task
Social Reasoning Task
Logical Reasoning Ability
Mathematical Reasoning Ability
Causal Reasoning Ability
Temporal Reasoning Ability

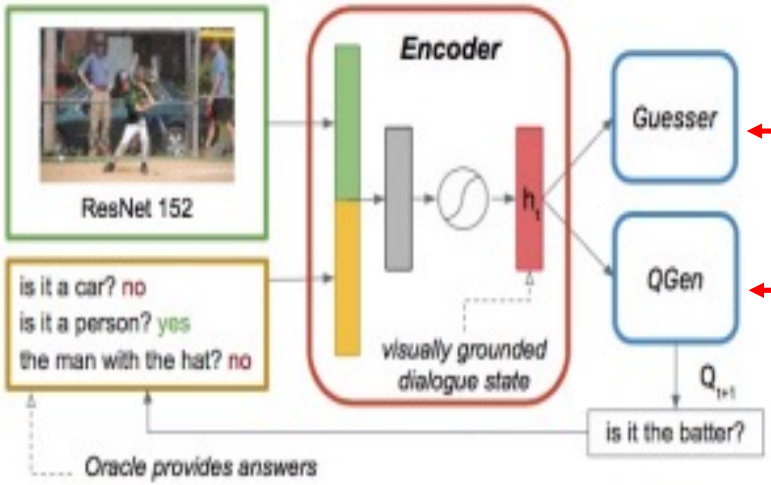


Reasoning and NLG

GuessWhat?!

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Beyond task Success:
Quality of the dialogues

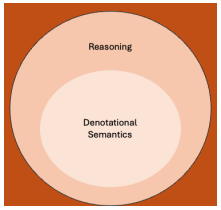


REASONING

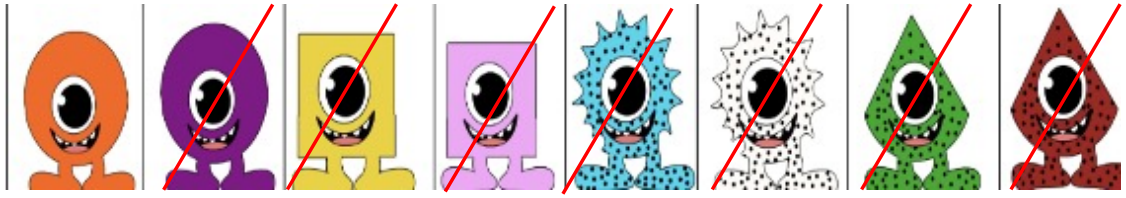
LANGUAGE GENERATION

interplay

Shekhar et al NAACL 2019



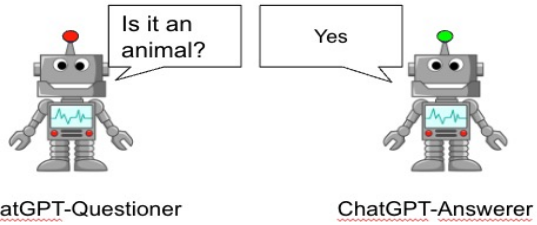
Language and Reasoning Interplay



Guess the target

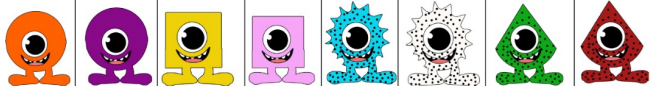
Optimal Questioner **Answerer**

1. Is it dotted? No
2. Is it round? Yes
3. Is it purple? No
4. It's the orange one. GUESSED



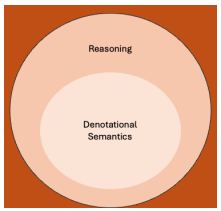
4Y child: Is it orange?

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

ChatGPT's Information Seeking Strategy: Insights from the 20-Questions Game
Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, Raffaella Bernardi, INLG 2023



Expected Information Gain (EIG)

	ChatGPT-Q	ChatGPT-A	HypSpace
1.	Is the item edible?	Yes	[squash, watermelon, mulberry, gourd]
2.	Is the item a fruit?	Yes	[watermelon, mulberry, gourd]
3.	Is the fruit typically round in shape?	Yes	[watermelon, mulberry, gourd]
4.	Is the fruit green on the outside and pink on the inside?	No	[mulberry, gourd]
5.	Is the fruit typically green on the outside, with a hard exterior and soft interior?	No	[mulberry]
6.	Is the fruit commonly used in savory dishes?	No	[mulberry]
7.	Is the fruit commonly consumed raw?	Yes	[mulberry]
8.	Does the name of the fruit contain a type of vehicle?	No	[mulberry]
9.	Is the fruit part of the melon family?	No	[mulberry]
10.	Based on your answers, the item assigned to you is most likely a **mulberry**	Yes! That's correct.	

Q1: EIG = 1.0

Q2: EIG = 0.81

Q3: EIG = 0.0

Q4: EIG = 0.92

Q5: EIG = 1.0

Q6: EIG = 0.0 – unnecessary

Q7: EIG = 0.0 – unnecessary

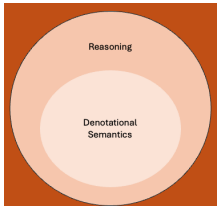
Q8: EIG = 0.0 – unnecessary

Q9: EIG = 0.0 – unnecessary

Q10: Guess

	Av. #Q
Optimal	3.5
4Y child	4.5
Chat-GPT	7.24

Again, Cognitive Science studies turned out to provide useful evaluation settings (and data)



Target = blouse

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

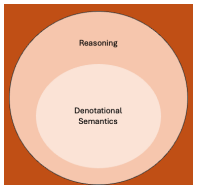
Questioner Is the item you have edible?
Answerer No, it is not edible.
Questioner Is it an article of clothing?
Answerer Yes, it is an article of clothing.
Questioner Is it a type of footwear?
Answerer No, it is not a type of footwear.
Questioner Does the item have long sleeves?

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!**

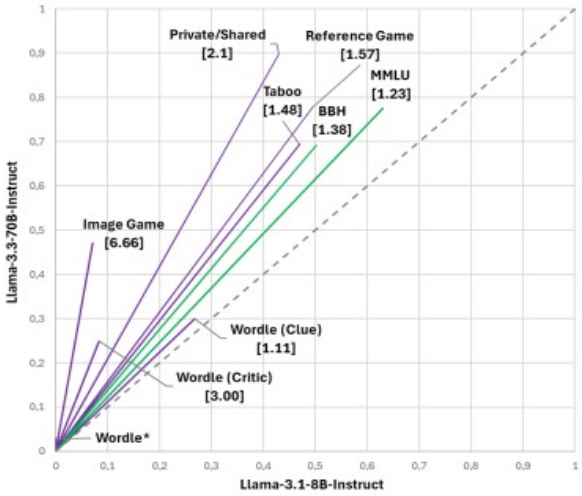
It misses the *Language Use*

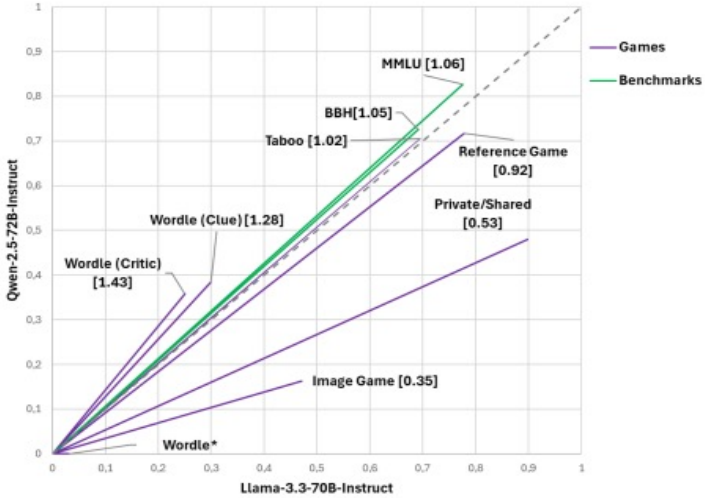
Miss the interlocutor communicative goal!
12% ChatGPT-Q and 7% ChatGPT-Q-stepwise

GPT in 2023



The power of interactive setting





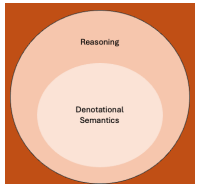
Triangulating LLM Progress through Benchmarks, Games, and Cognitive Tests
Filippo Momentè, Alessandro Suglia, Mario Giulianelli , Ambra Ferrari, Alexander Koller, Oliver Lemon, David Schlangen, Raquel Fernández , Raffaella Bernardi, ACL 2025

Dialogue Games:
✓ useful for evaluation
Useful for *efficient training* to learn *language use*?

Workshop at EMNLP 2026
Improving Language Models through **Learning from Dialogue Interaction**
https://lm-playschool.github.io/

LLADIGA 3-Y project DFG and Provincia Autonoma di Bolzano

Autonome Provinz Bozen
Provincia autonoma di Bolzano
Provinzia autonoma de Bulsan
SÜDTIROL · ALTO ADIGE



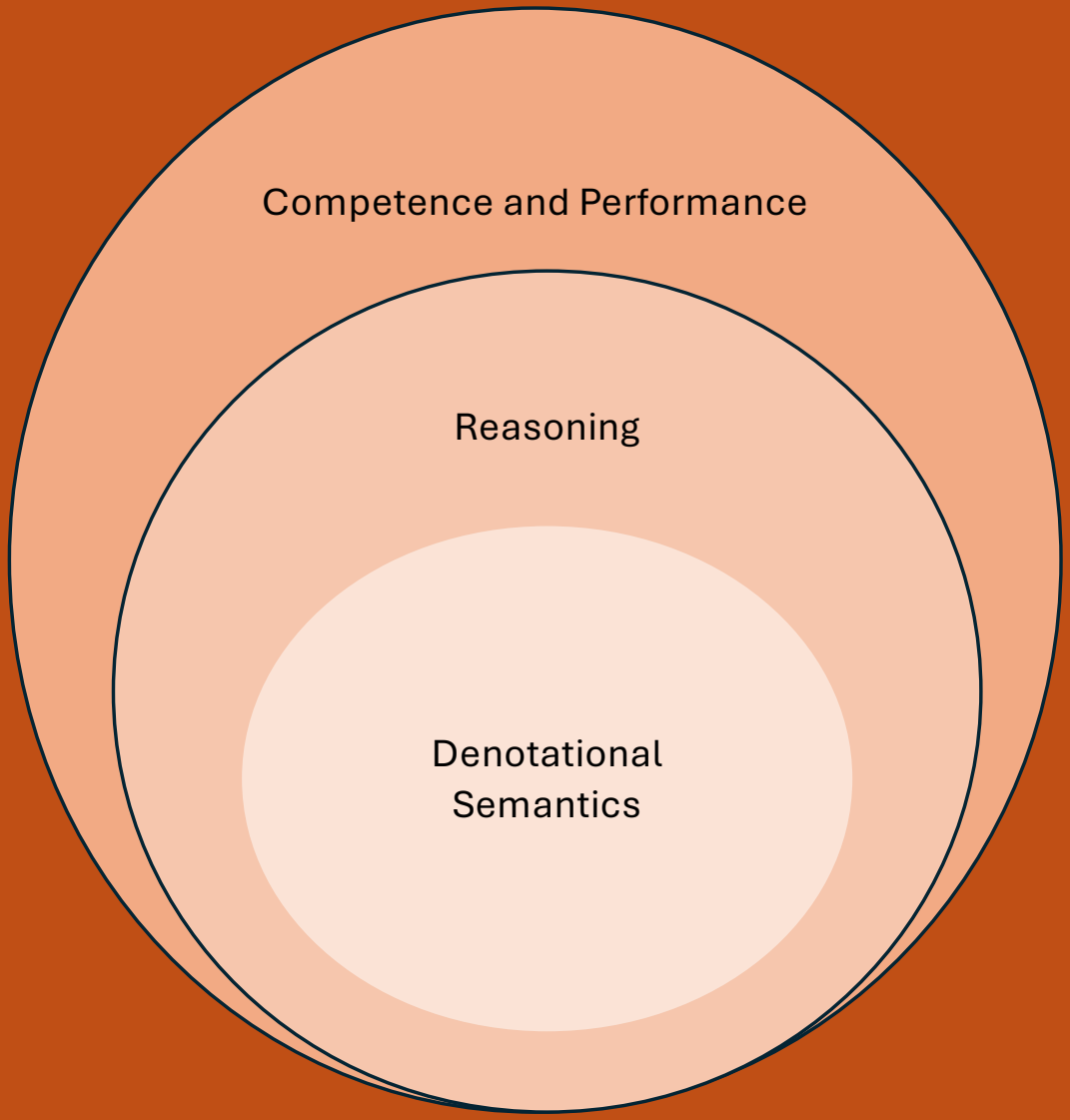
Conclusion on Reasoning

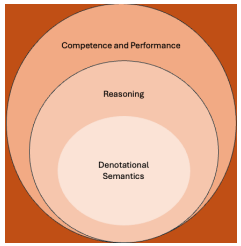
There is no intelligence, without reasoning;

reasoning does not happen in isolation:
it happens within a linguistic or visual context,
it unfolds through interaction.

Language and Reasoning interplay
Should reasoning benchmarks be based on task requiring grounding information against an image or a linguisticaly descreibed world, or in a dialogue?

LARGE LANGUAGE MODELS





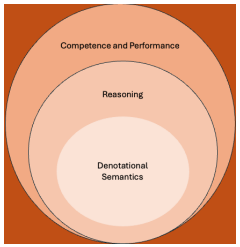
Competence and Performance

"**Linguistic theory** is concerned primarily with an **ideal speaker-listener**, in a completely homogeneous speech-community..." (Chomsky, 1965)

A record of **natural speech** will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on. The problem for the linguist, as well as for the child learning the language, is to **determine** from the data of performance **the underlying system** he puts to use in actual performance. (Chomsky 1965:3f)

A "grammar of a language" is "a description of the ideal speaker-hearer's intrinsic competence"

competence = knowledge of language
performance = actual use of language in concrete situations.



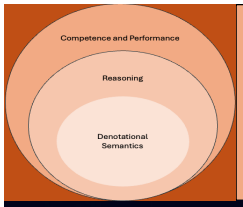
Large Language Models as Neurolinguistic Subjects: Discrepancy between Performance and Competence

Linyang He[1,2] **Ercong Nie**[3,4] **Helmut Schmid**[4]
Hinrich Schütze[3,4] **Nima Mesgarani**[1] **Jonathan Brennan**[2]
[1]Columbia University [2]University of Michigan
[3]Munich Center for Machine Learning, Germany [4]LMU Munich, Germany
linyang.he@columbia.edu {nie,schmid}@cis.lmu.de
hinrich@hotmail.com nima@ee.columbia.edu jobrenn@umich.edu

ACL 2025

When **treating LLMs as psycholinguistic** subjects, their responses may leverage their grasp of form, relying on statistical correlations, to create an illusion of understanding meaning. [..] psycholinguistic evaluations tend to reflect performance rather than competence**, as they assess external outputs that may not fully capture the underlying linguistic knowledge encoded within the model**. This mismatch suggests that psycholinguistic evaluation results might not accurately represent the true linguistic competence of LLMs.

In contrast, examining **LLMs as neurolinguistic** subjects focuses on **internal representations**, **providing a more direct assessment of competence by moving beyond surface-level biases. [..]** This approach allows for **a finer distinction between performance and competence**, revealing insights that psycholinguistic methods might overlook



Competence: Internal Representation Analyses

The cat sat on the mat.

The cat rested comfortably.

Probing Classifiers

Train simple classifiers on representations to detect what information is present.

Linear Classifier

Label

0.12 ✗

0.88 ✓

Is subject = animal?

Steering Vectors

Use direction vectors in activation space to steer model behavior.

x + v (steered)

v

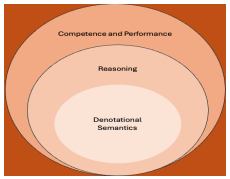
v (steering vector)

x

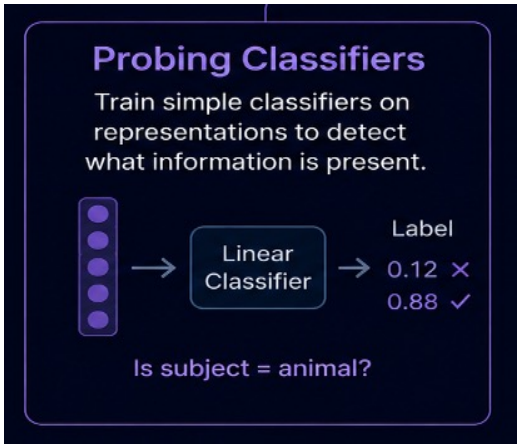
Circuit Analysis

Identify components and pathways that implement specific behaviors.

Credits: ChatGPT



Probing Classifiers



What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

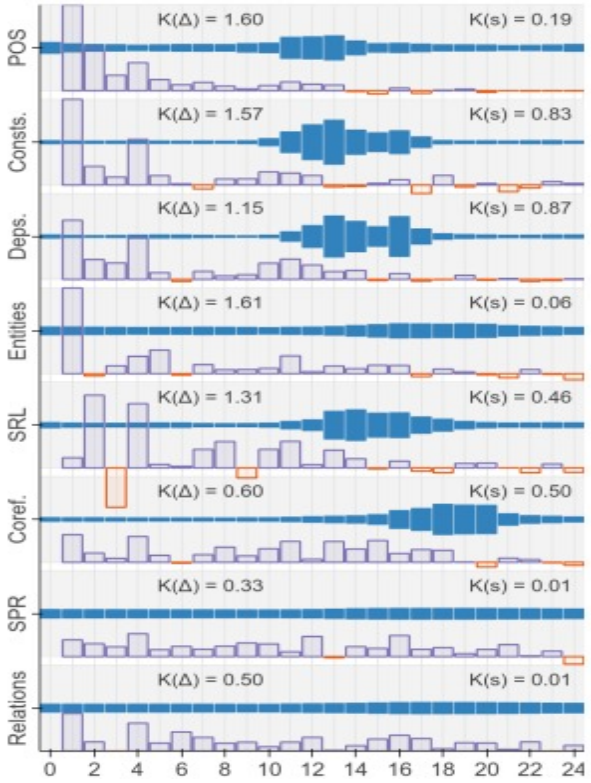
German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

Marco Baroni
Facebook AI Research
mbaroni@fb.com

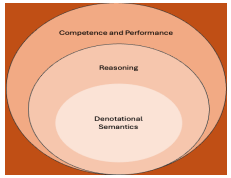
ACL 2018



BERT Rediscovers the Classical NLP Pipeline

Ian Tenney[1] **Dipanjan Das**[1] **Ellie Pavlick**[1,2]
[1]Google Research [2]Brown University
{iftenney,dipanjand,epavlick}@google.com

ACL 2019



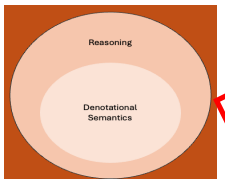
Probing Classifiers: Syntax-Semantics Interface

Using our **probe**, we show that such transformations exist for both **ELMo and BERT** but not in baselines, **providing evidence that entire syntax trees are embedded implicitly in deep models' vector geometry.**

John Hewitt and Christopher D. Manning,
A Structural Probe for Finding Syntax in Word Representation,
NAACK-HLT 2019

We study how syntactic and semantic information is encoded in inner layer representations [..] of **DeepSeek-V3**. [..] suggesting that syntax and semantics are, at least partially, linearly encoded. We also find that the **cross-layer encoding profiles of syntax and semantics are different**, and that the two signals can to some extent be decoupled.

Santiago Acevedo, Alessandro Laio, Marco Baroni,
Differential syntactic and semantic enconding in LLMs ICML 2026



RECALL

LLMs display content bias

Syllogism AO3

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

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.

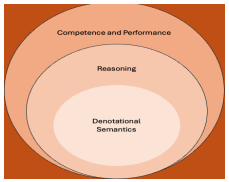
Syllogism EO1

P1: No dogs are felines.
P2: Some felines are not cats.
———————————————
C: Nothing follows

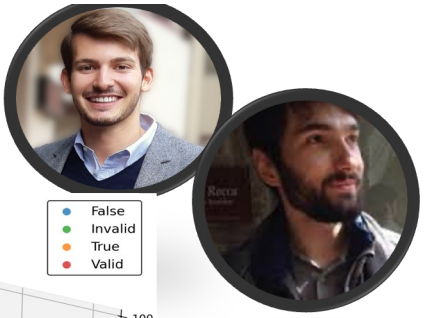
Leaderboard approach

Model	Believable			Unbelievable	Content Effect
	Acc	Invalid	Valid	Valid	Difference
Pythia-160M	0.16	0.00	0.37	0.00	-100.00%*
Pythia-410M	9.53	0.00	22.59	19.63	-13.10%
Pythia-1.4B	14.22	1.94	31.11	33.33	+7.14%
Pythia-1.4B-inst	17.03	15.28	19.26	15.93	-17.29%*
LLaMA-2 7B	24.38	3.06	53.70	47.78	-11.02%
LLaMA-2-inst 7B	19.69	10.83	32.22	19.26	-40.22%
LLaMA-2 13B	22.81	0.28	53.70	40.37	-24.82%
LLaMA-2-inst 13B	22.34	8.06	42.22	35.93	-14.19%*
LLaMa-3 8B	30.31	0.83	70.74	52.59	-10.44%
LLaMA-3-inst 8B	50.00	30.28	78.15	64.81	-17.07%
LLaMA-3 70B	30.63	10.28	58.89	54.07	-8.18%
LLaMA-3-inst 70B	41.25	32.50	54.44	57.41	+5.46%

Bertolazzi et. al. 2024. *A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences*. EMNLP 2024



Why LLMs display content bias



Validity classification

P1: All dogs are labradors
P2: Some labradors are not canines
C: No dogs are canines

LLM

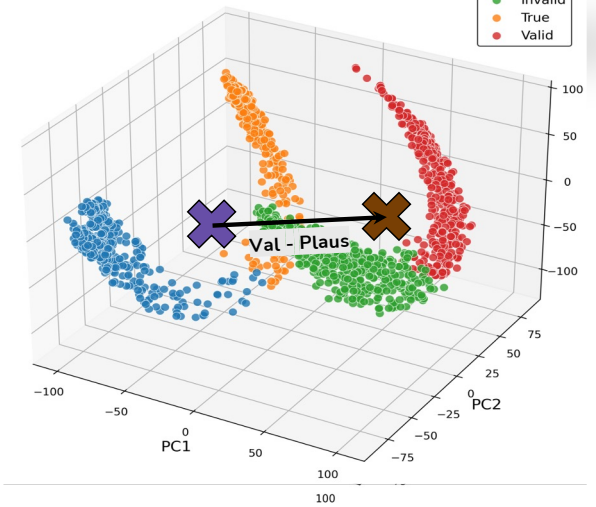
Invalid

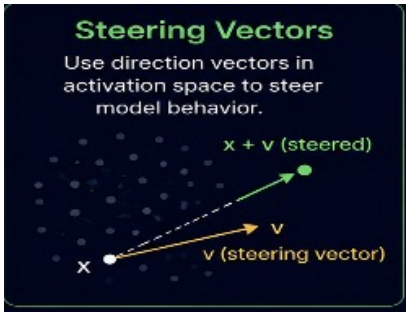
Plausibility classification

Statement:
No dogs are canines

LLM

False





Validity classification

P1: All dogs are labradors
P2: Some labradors are not canines
C: No dogs are canines

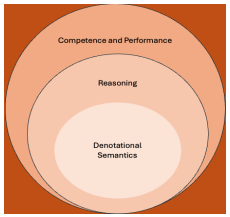
LLM <- V

Valid

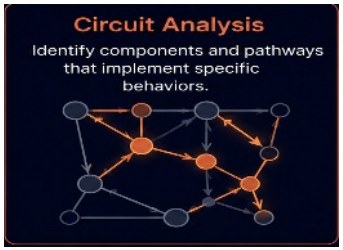
The parallel geometric structure between true/false and valid/invalid clusters suggests **shared representational directions for plausibility and validity**

V= validity - plausibility

Bertolazzi, Pezzelle, Bernardi
How Language Models **Conflate Logical Validity with Plausibility**:
A Representational Analysis of Content Effects, ACL 2026



Circuit Analysis:
Why LLMs fail to detect math errors?



Frank initially has 5 apples. After buying 8 more apples, how many apples does he have?

To solve this, we add 5 + 8 = 16.
Thus, Frank has a total of 16 apples.

The above reasoning is: **INVALID**

If we compute 5 + 8 in our mind (13) and compare it with 16, we notice a mistake!

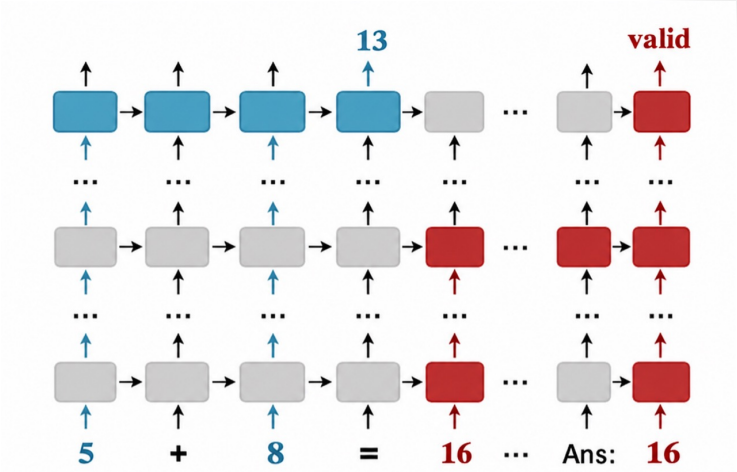


Figure 1: A schematic overview of the structurally dissociated circuits responsible for arithmetic computation and validation. While the models' internal arithmetic computation primarily occurs in higher layers, validation takes place in mid-to-lower layers. before the final arithmetic results are fully encoded (see Section 4.3).

Bertolazzi, Mondorf, Plank, Bernardi
The Validation Gap: A Mechanistic Analysis of How Language Models Compute Arithmetic but Fail to Validate It, EMNLP 2025

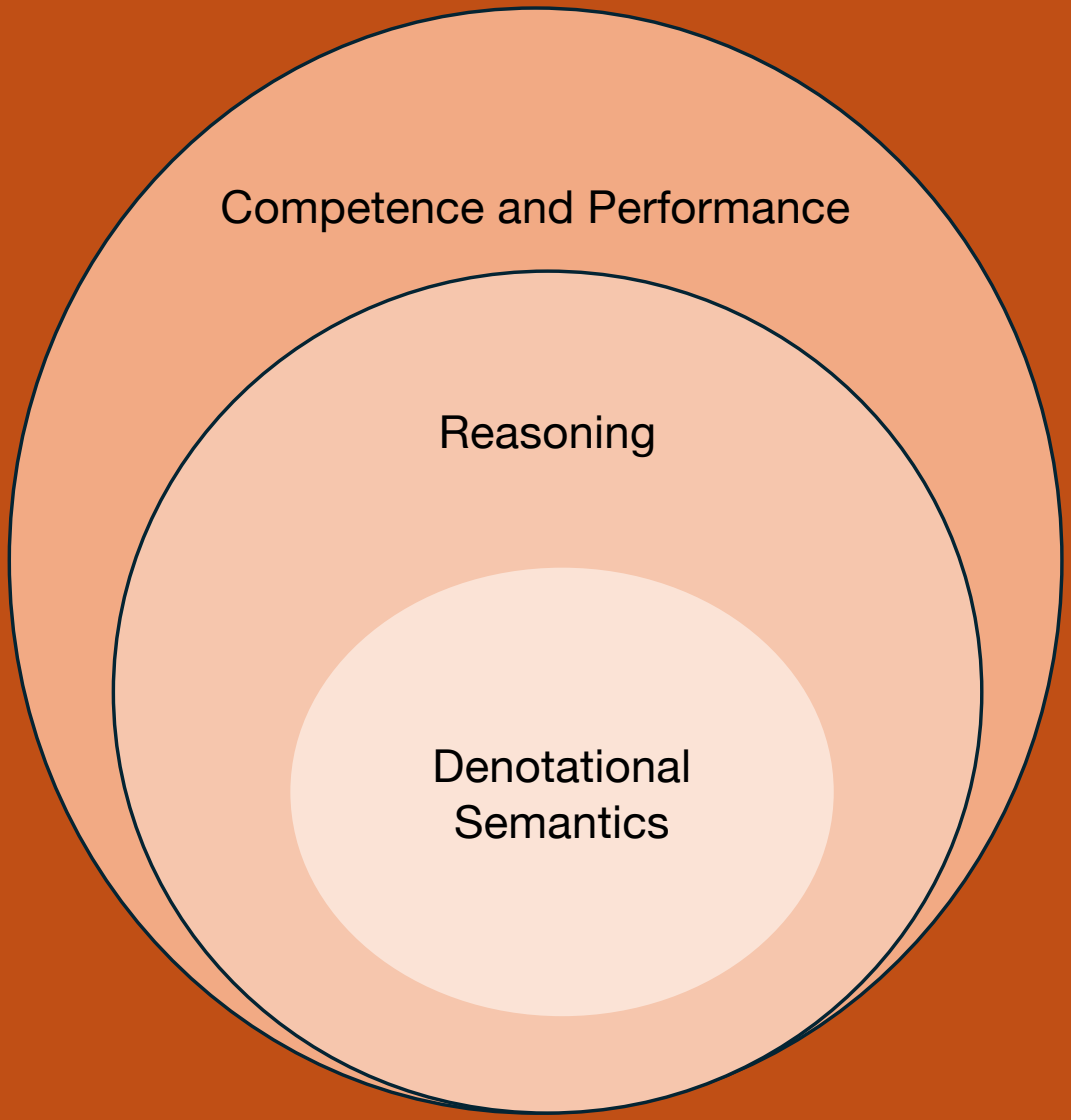
Conclusion on Competence & Performance

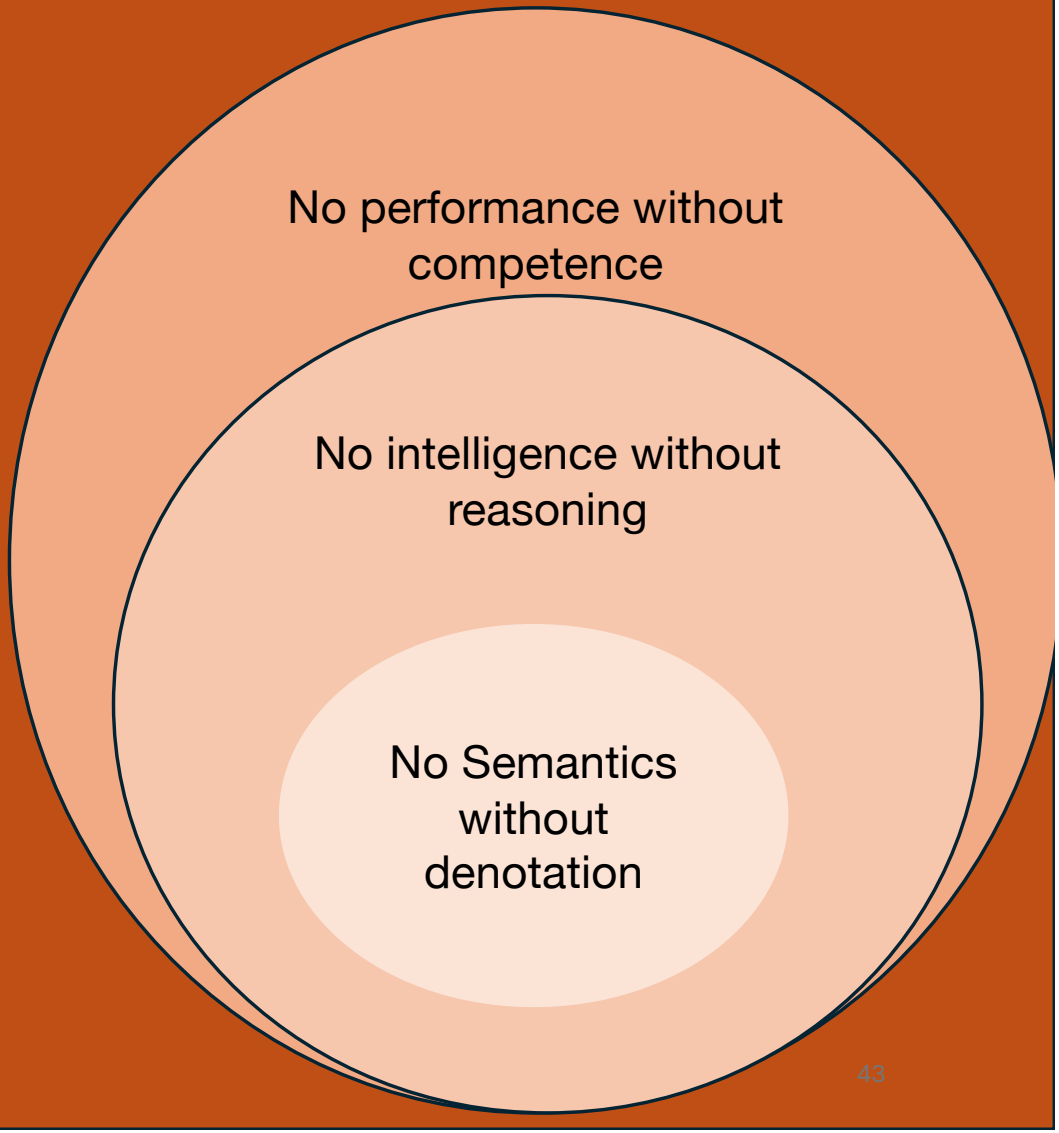
No performance without competence

Interpretability methods provide a **new bridge** between neural language models and formal linguistics, allowing us to investigate and **compare formal theories against the internal representations of LLMs**.

Can interpretability methods help inspecting whether and how LLMs process logical words?

LARGE LANGUAGE MODELS





Through my ESSLLI life I have appreciated the elegance of **Formal Linguistics Theories**. They shed light on how language works, and on the logic properties of linguistic expressions.

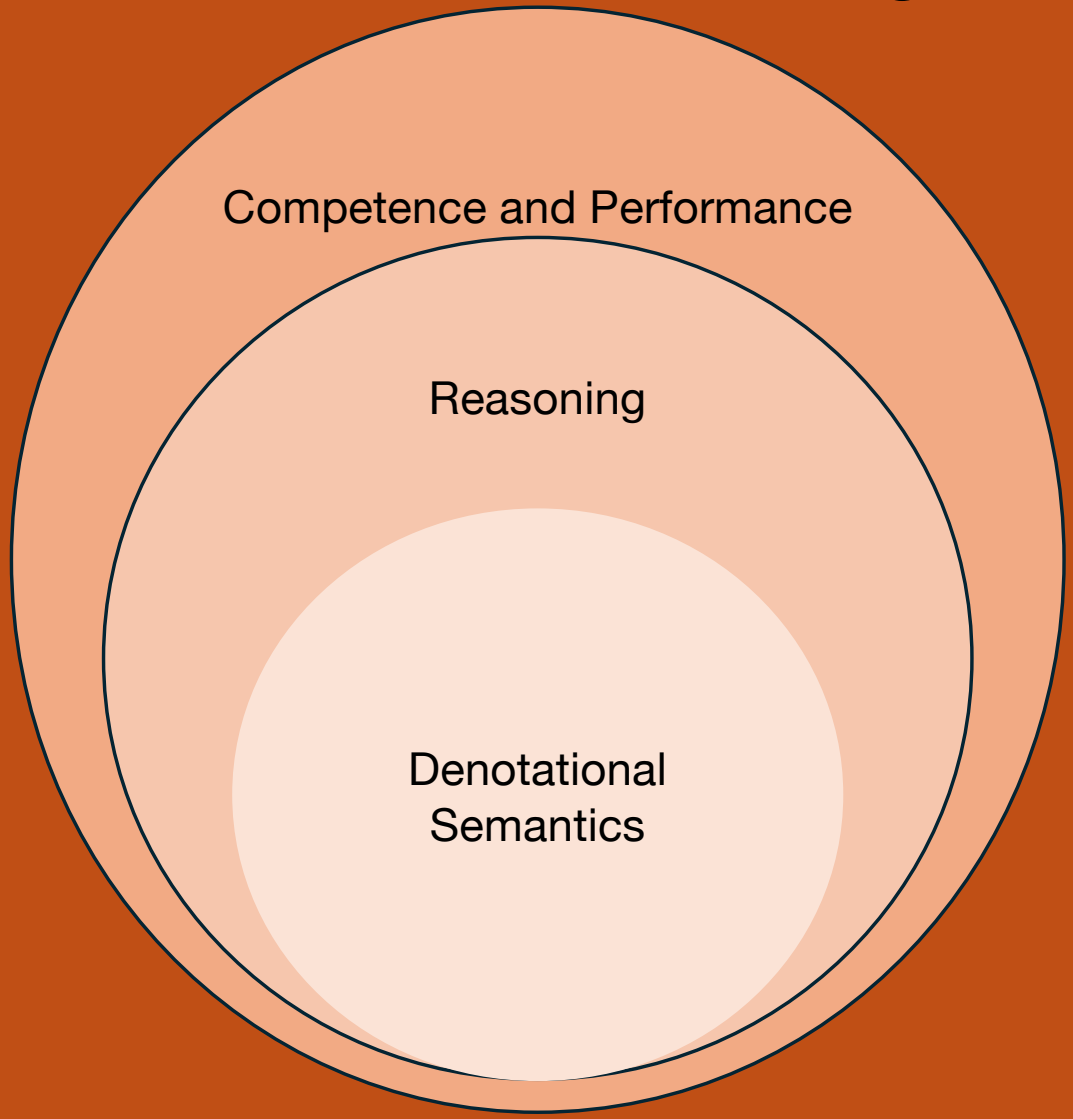
The knowledge of Formal Linguistic Theories helps **creating diagnostic evaluation** datasets to inspect LLMs's behaviour as well as their internal representations.

I am amazed by how LLMs have learned language, I am **still agnostic** on whether they can arrive to properly capture the concept of

entities, truth values, and, hence, of logical words.

I believe we are in an interesting phase in which *Logic, Language and Information* lens can help advance our understanding on human language, of human reasoning and of computational models of them.

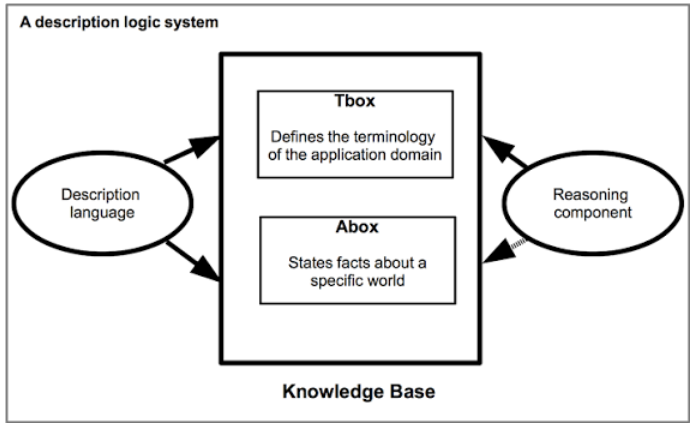
Looking at the future



ChatGPT's generated image

Journey through the ESSLLI courses: future challenges for LLMs

Description Logic



TBox
(intensional knowledge)

Planet ⊑ CelestialBody

Star ⊑ CelestialBody

Planet ⊓ Star ⊑ ⊥

ABox
(extensional knowledge)

Planet(Venus)

Planet(Mars)

visibleAt(Venus, Morning)

hasColor(Mars, Red)

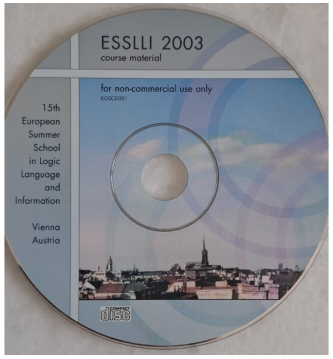
Enrico Franconi and Diego Calvanese (ESSLLI 2003)

LLMs may have learned a lot of intensional knoweldge;

they have also learned "extensional" knowledge of entities about which information is available in their training data.

But do the truly capture the distinction
extensional vs. intentional?

Discourse Representation Theory



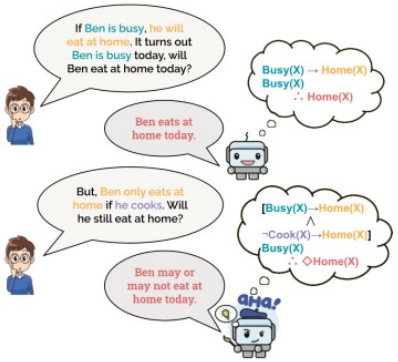
Gemma Boleda's ERC aimed at bringing DRT into Distributional Semantics'view

+ Gemma Boleda

+ Julia Hockenmaier

Renata Wasserman (ESSLLI 1997)

Belief Revision



Open to collaboration and .. new friendship ☺

Belief Revision: The Adaptability of Large Language Models Reasoning Wilie et al EMNLP 2024

Last lesson I learned

Believe in your ideas

Do a proper literature review tracing back the history and development of ideas

Enjoy your work

Enjoy working with your co-authors

Filippo Merlo

Davide Mazzaccara

Leonardo Bertolazzi

Filippo Momentè

Philipp Mondorf

Alberto Testoni

Alessandro Suglia

Claudio Greco

Sandro Pezzelle

Ravi Shekhar

B. Plank

R. Fernández

D. Schlangen

A. Gatt

Thanks to ESSLLI