Dissociating language and thought in large language models

Anna (Anya) Ivanova
Georgia Tech

Feb 6, 2025

The rise of AI psychology

P2-EE-382: *Evaluating human-like similarity biases at every scale in Large Language Models: Evidence from remote and basic-level triads.*
Presenting Author: **Simon De Deyne**, *University of Melbourne*

VP-CC-52: *Incremental Comprehension of Garden-Path Sentences by Large Language Models: Semantic Interpretation, Syntactic Re-Analysis, and Attention*
Presenting Author: **Andres Li**, *Georgia Institute of Technology*

P2-E-249: *Stick to your Role! Stability of Personal Values Expressed in Large Language Models*
Presenting Author: **Grgur Kovač**, *INRIA*

P2-C-203: *Assessing Common Ground through Language-based Cultural Consensus in Humans and Large Language Models*
Presenting Author: **Sophie Domanski**, *University of Maryland*

P3-E-592: *Implicit Bias in Language Models – A Narrative Literature Review with Systematic Elements*
Presenting Author: **Kevin D. Kiy**, *Maynooth University*

T.22.02: *Does reading words help you to read minds? A comparison of humans and LLMs at a recursive mindreading task*
Presenting Author: **Cameron R. Jones**

P3-E-591: *Testing Causal Models of Word Meaning in LLMs*
Presenting Author: **Sam Musker**, *Brown University*

P3-U-705: *The Role of Episodic Memory in Storytelling: Comparing Large Language Models with Humans*
Presenting Author: **Charlotte Cornell**, *Rutgers University--New Brunswick*

P3-LL-915: *Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading*
Presenting Author: **Weiyan Liao**, *University of Hong Kong*

P3-EE-760: *Do Large language Models know who did what to whom?*
Presenting Author: **Joseph Denning**, *UCLA*

P2-E-250: *Different Trajectories through Option Space in Humans and LLMs*
Presenting Author: **Alina Dracheva**, *Dartmouth College*

P3-M-671: *Self-Hint Prompting Improves Zero-shot Reasoning in Large Language Models via Reflective Cycle*
Presenting Author: **Jindou Chen**, *Shanghai Jiao Tong University*

P1-EE-93: *Experimental Pragmatics with Machines: Testing LLM Predictions for the Inferences of Plain and Embedded Disjunctions*
Presenting Author: **Polina Tsvilodub**, *University of Tübingen*

P3-E-576: *Compositionality is underutilized in language: The case of English adjective-noun phrases in humans and large language models*
Presenting Author: **Aalok Sathe**, *Massachusetts Institute of Technology*

P3-E-579: *Estimating human color-concept associations from multimodal language models*
Presenting Author: **Kushin Mukherjee**, *University of Wisconsin-Madison*

P2-E-248: *Creative goal generation in humans and language models*
Presenting Author: **Junyi Chu**, *Harvard University*

VP-E-14: *LLMs Don't "Do Things with Words" but Their Lack of Illocution Can Inform the Study of Human Discourse*
Presenting Author: **Zachary Rosen**, *University of California, Los Angeles*

P2-E-252: *The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents*
Presenting Author: **Timothy T Rogers**, *University of Wisconsin - Madison*

P3-EE-774: *Investigating Iconicity in Vision-and-Language Models: A Case Study of the Bouba/Kiki Effect in Japanese Models*
Presenting Author: **Hinano Iida**, *Nagoya University*

P2-E-221: *Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models*
Presenting Author: **Martha Lewis**, *University of Bristol*

P3-EE-766: *Large Language Models and Human Discourse Processing*
Presenting Author: **Eyal Sagi**, *University of St. Francis*

VP-CC-53: *Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making*
Presenting Author: **Yinuo Du**, *Software and Societal Systems Department*

P2-E-256: *Large Language Models Show Human-Like Abstract Thinking Patterns: A Construal-Level Perspective*
Presenting Author: **Seung Joo Yoo**, *Seoul National University*

P3-E-596: *Interpretation of Novel Literary Metaphors by Humans and GPT-4*
Presenting Author: **Nicholas Ichien**, *University of Pennsylvania*

P2-E-553: *Evaluating language model alignment with free associations*
Presenting Author: **Dirk Wulff**, *Max Planck Institute for Human Development*

P3-M-676: *Simulating Opinion Dynamics with Networks of LLM-based Agents*
Presenting Author: **Yun-Shiuan Chuang**, *University of Wisconsin - Madison*

The rise of AI psychology

P2-EE-382: *Evaluating human-like similarity biases at every scale in Large Language Models: Evidence from remote and basic-level triads.*
Presenting Author: **Simon De Deyne**, *University of Melbourne*

VP-CC-52: *Inc*
Path Sentenc
Interpretatio
Presenting Auth

P2-E-249: *Stick to your Role! Stability of Personal Values Expressed in Large Language Models*
Presenting Author: **Grgur Kovač**, *INRIA*

P2-C-203: *Assessing Common Ground through Language-based Cultural Consensus in Humans and Models*
ophie Domanski, *University of Maryland*

in Language Models – A
view with Systematic Elements
iy, *Maynooth University*

T.22.02: *Does reading wor*
of humans and LLMs at
Presenting Author: **Cameron R. Jon**

LLMs
Presenting Author: **Sam Musker**, *Brown University*

P3-LL-915: *Do large language models resolve*
se
Th
re
Pre

P2-E-248: *Creative goal generation in humans and language models*
Presenting Author: **Junyi Chu**, *Harvard Univers*

ies through Option Space
, Dartmouth College

P3-M-671: *Self-Hint Prompting Improves Zero-shot Reasoning in Large Language Models via Reflective*

P3-E-579: *Estimating human color-concept associations from multimodal language models*
Presenting Author: **Kushin Mukherjee**, *University of Wisconsin-Madison*

P3-U-705: *The Role of Episodic Memory in Storytelling: Comparing Large Language Models with Humans*
Presenting Author: **Charlotte Cornell**, *Rutgers University--New wick*

P3-EE-774: *Investigating Iconicity in Vision-and-*

P2-E-221: *Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models*
Presenting Author: **Martha Lewis**, *University of Bristol*

VP-CC-53: *Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making*
Presenting Author: **Yinuo Du**, *Software and Societal Systems Department*

P2-E-256: *Large Language Models Show Human-Like Abstract Thinking Patterns: A Construal-Level Perspective*
Presenting Author: **Seung Joo Yoo**, *Seoul National University*

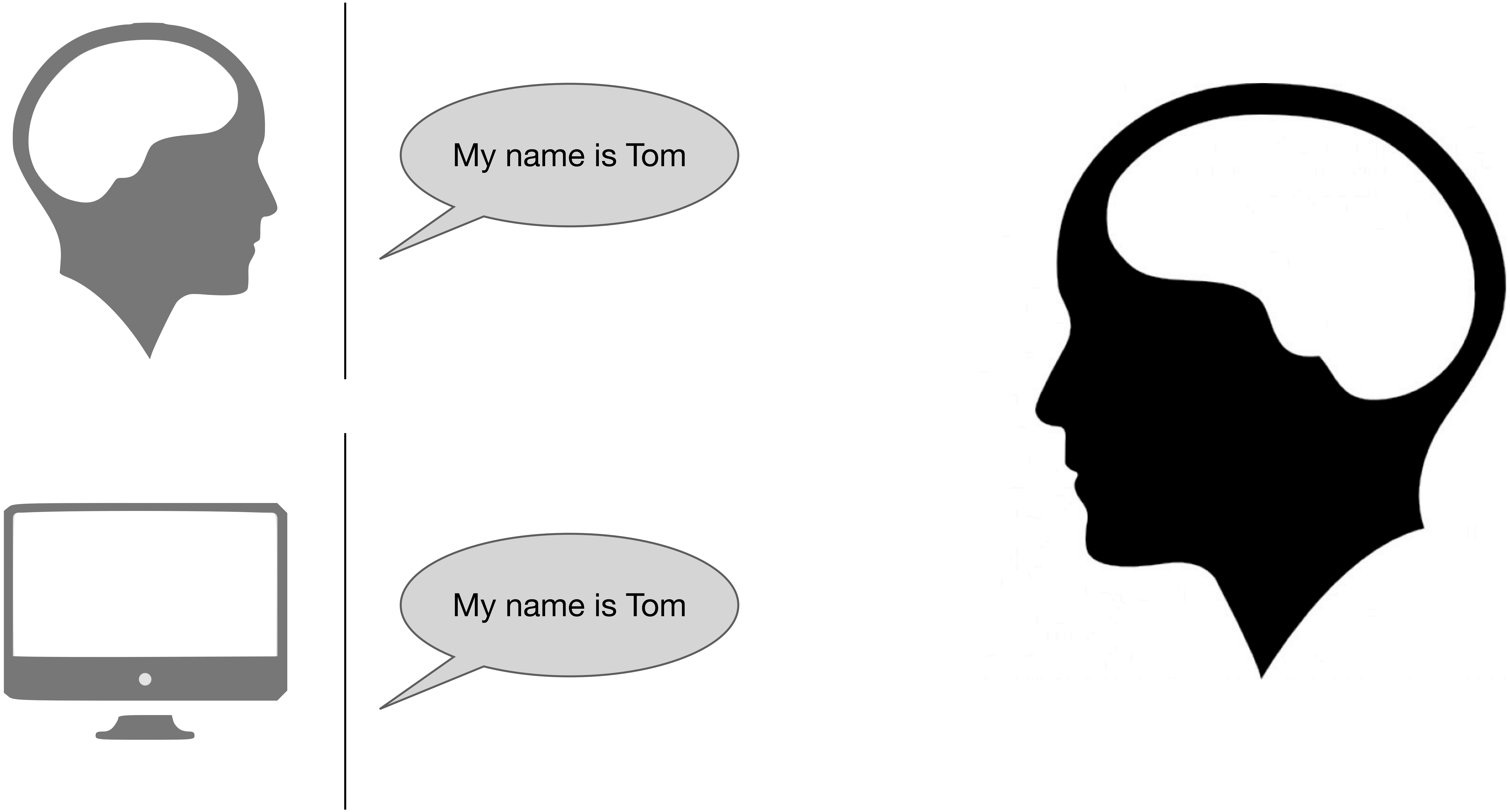
with

by Humans and GPT-4
Presenting Author: **Nicholas Ichien**, *University of Pennsylvania*

free associations
Presenting Author: **Dirk Wulff**, *Max Planck Institute for Human Development*

P3-M-676: *Simulating Opinion Dynamics with Networks of LLM-based Agents*
Presenting Author: **Yun-Shiuan Chuang**, *University of Wisconsin - Madison*

The Turing test



COMMON ERROR:

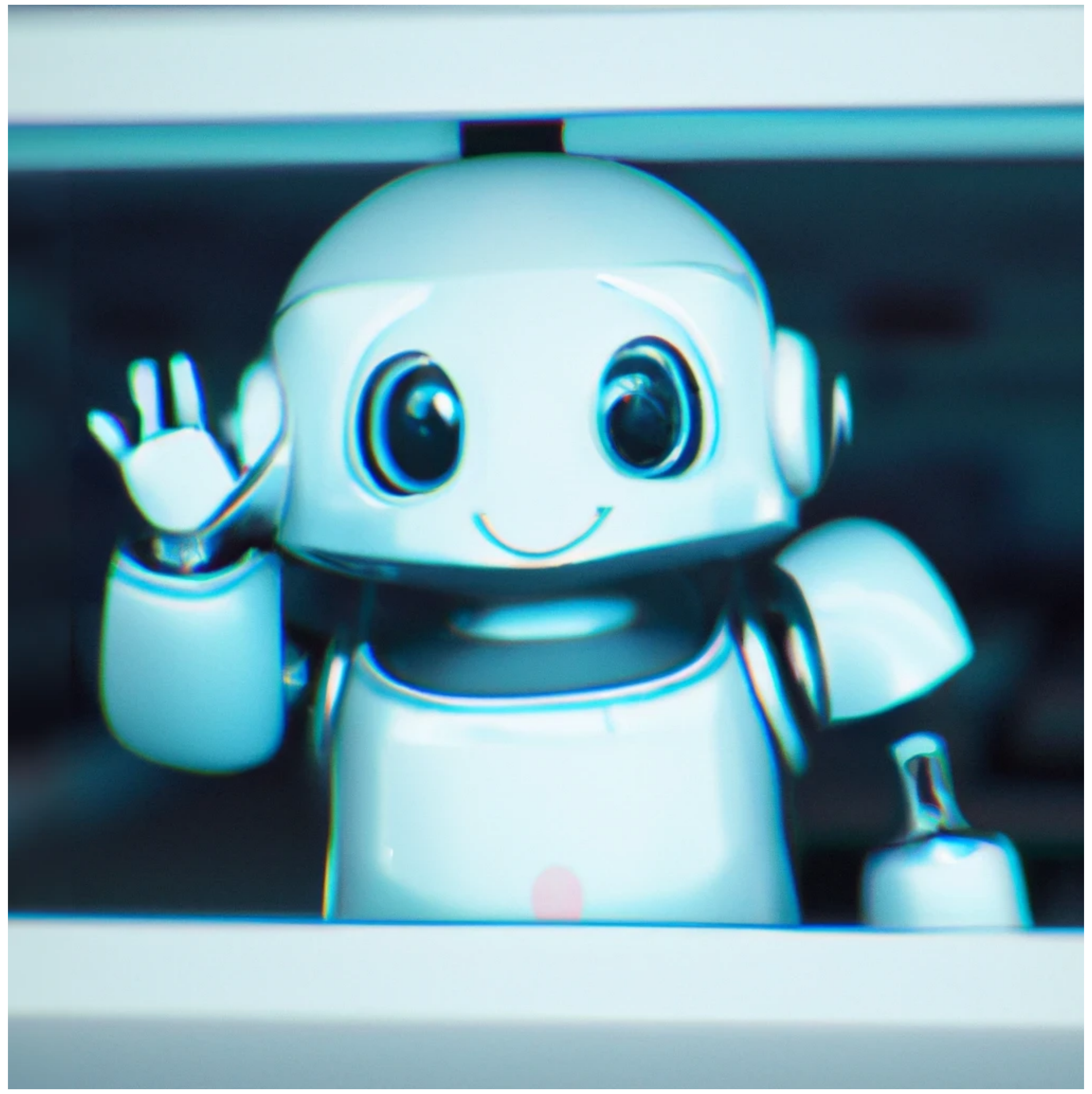
MISTAKING FLUENT LANGUAGE FOR FLUENT THOUGHT

Fallacy #1

good at language
↓
good at thought

Fallacy #2

bad at language
↑
bad at thought



When evaluating LLM capabilities, we should dissociate language and cognition/intelligence/thought.

Roadmap

Formal vs functional
linguistic competence

It gets complicated:
generalized world knowledge

Moving forward

Roadmap

Formal vs functional
linguistic competence

It gets complicated:
generalized world knowledge

Moving forward

Kyle Mahowald

Ev Fedorenko

Trends in
Cognitive Sciences

CellPress

Feature Review

Dissociating language and thought in large language models

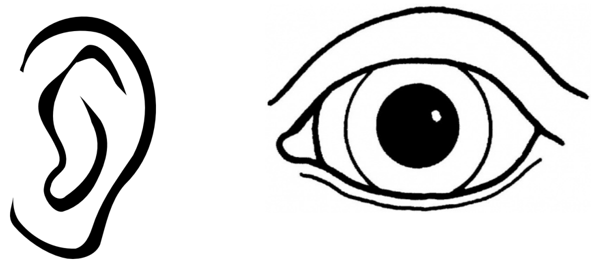
Kyle Mahowald,[1,5,*] Anna A. Ivanova,[2,5,*] Idan A. Blank,[3,*] Nancy Kanwisher,[4,*] Joshua B. Tenenbaum,[4,*] and Evelina Fedorenko[4,*]

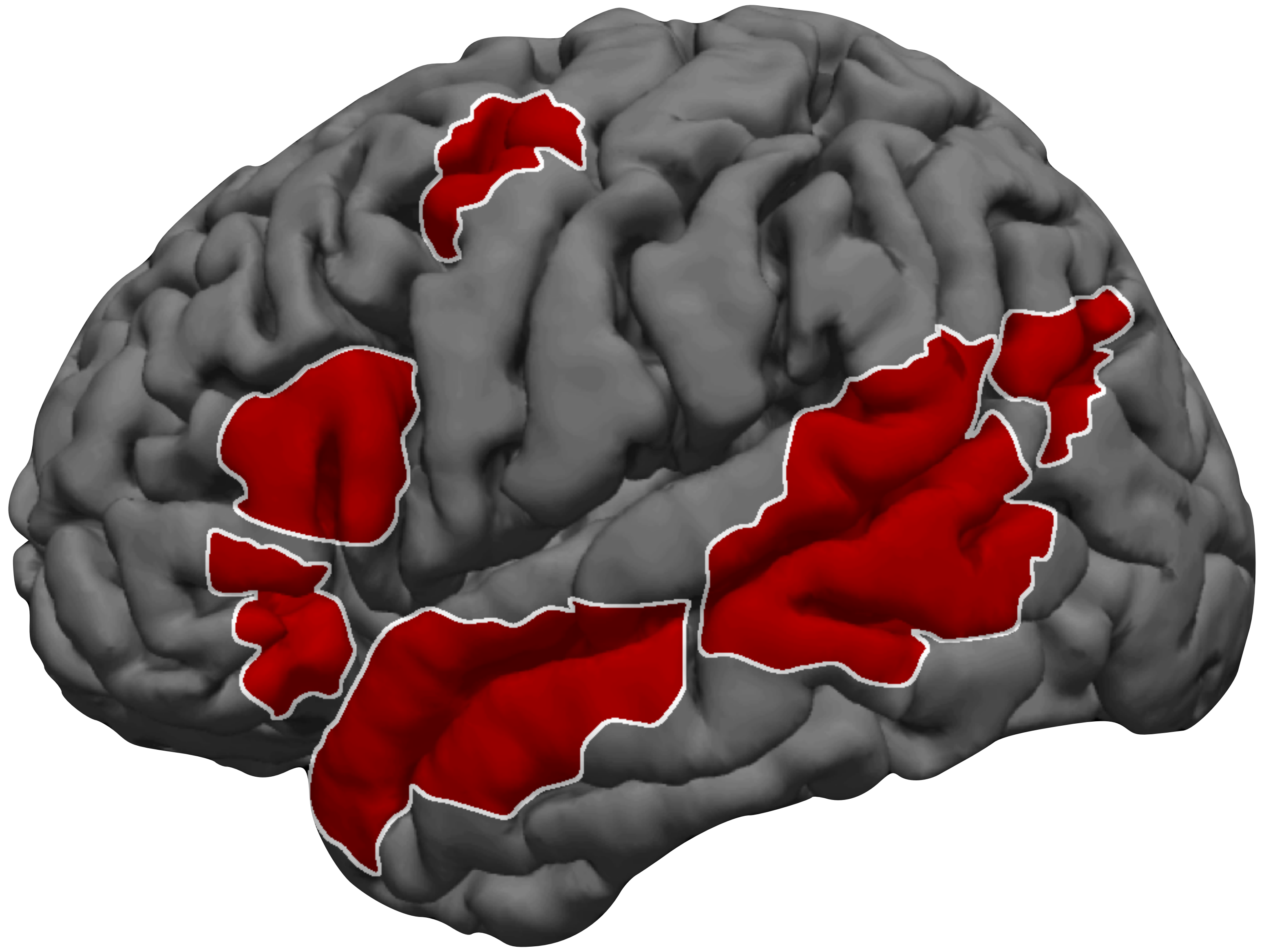
Language and the brain

Language processing in the brain takes place within a separate network.

Words, phrases, sentences

Listening and reading



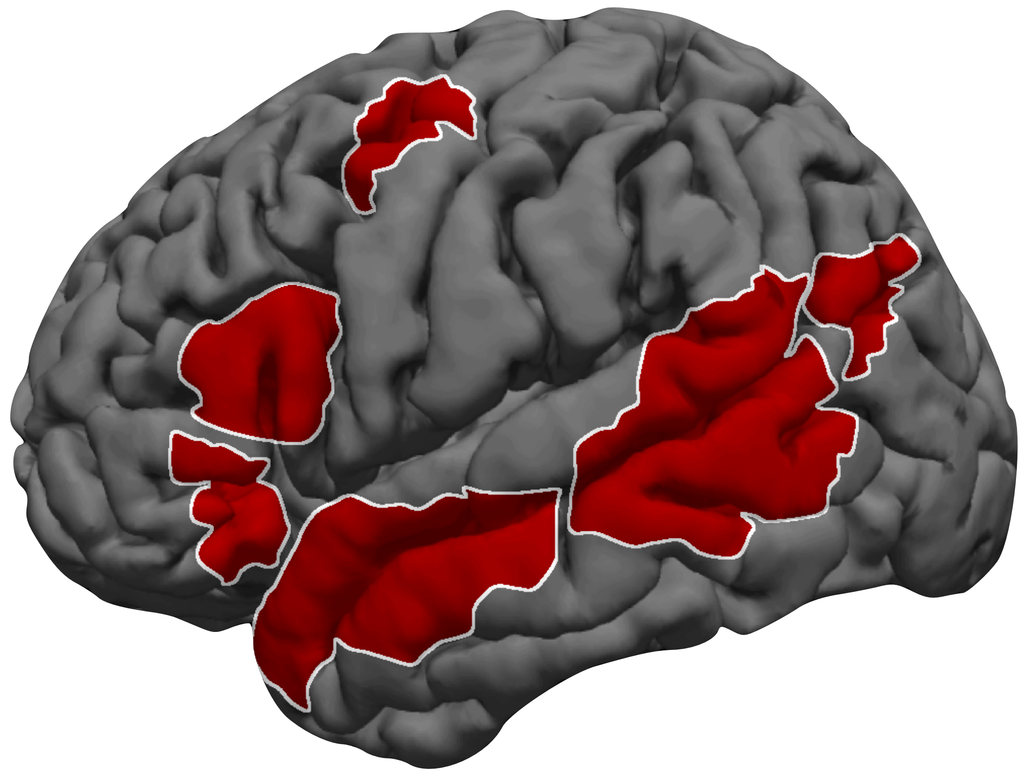


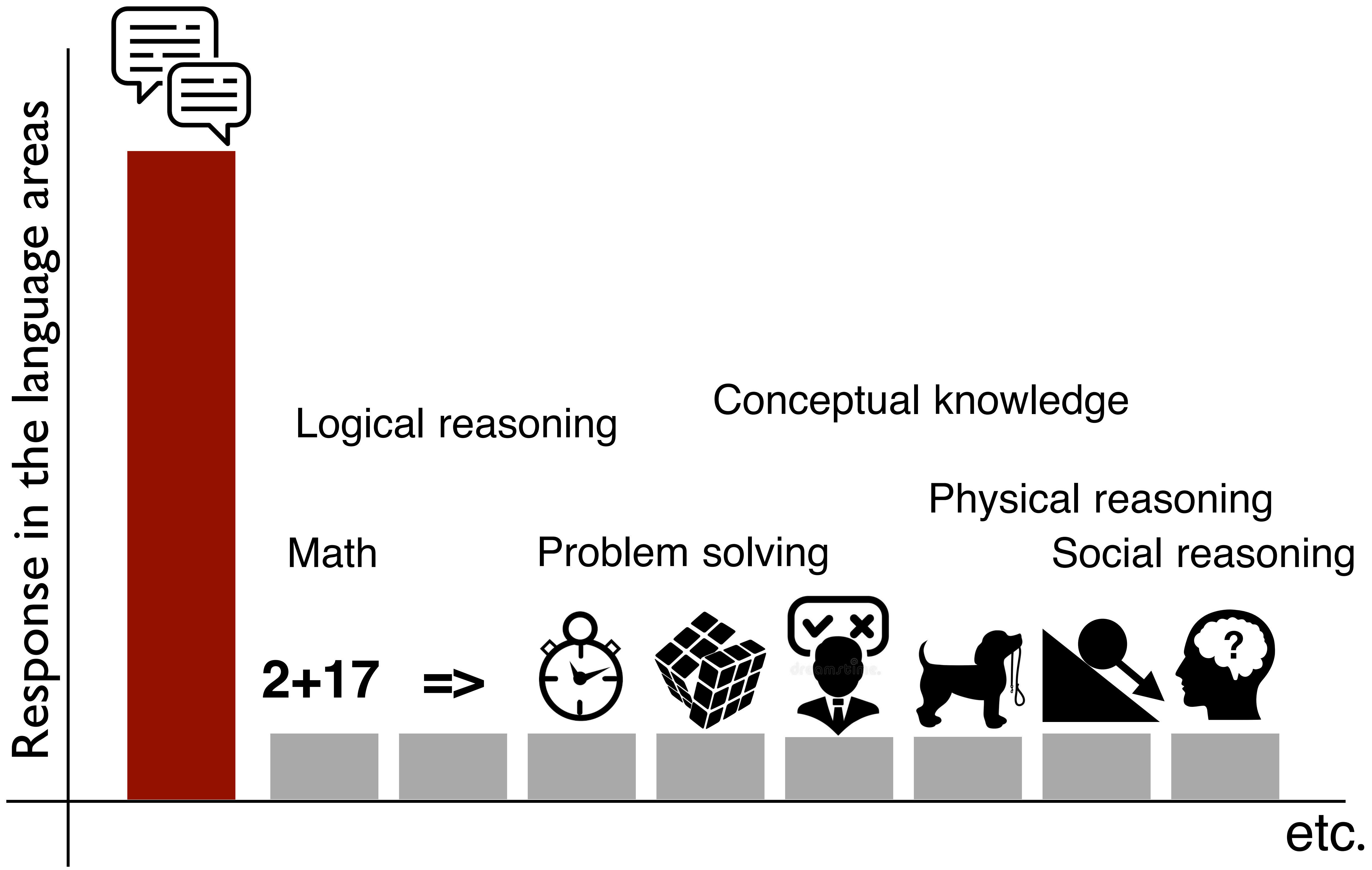
Speaking and writing

Fedorenko et al, 2010, 2011; Scott et al, 2017; Hu, Small et al, 2022; etc, etc

Language and the brain

Language areas **show little/no response** when we engage in diverse thought-related activities.



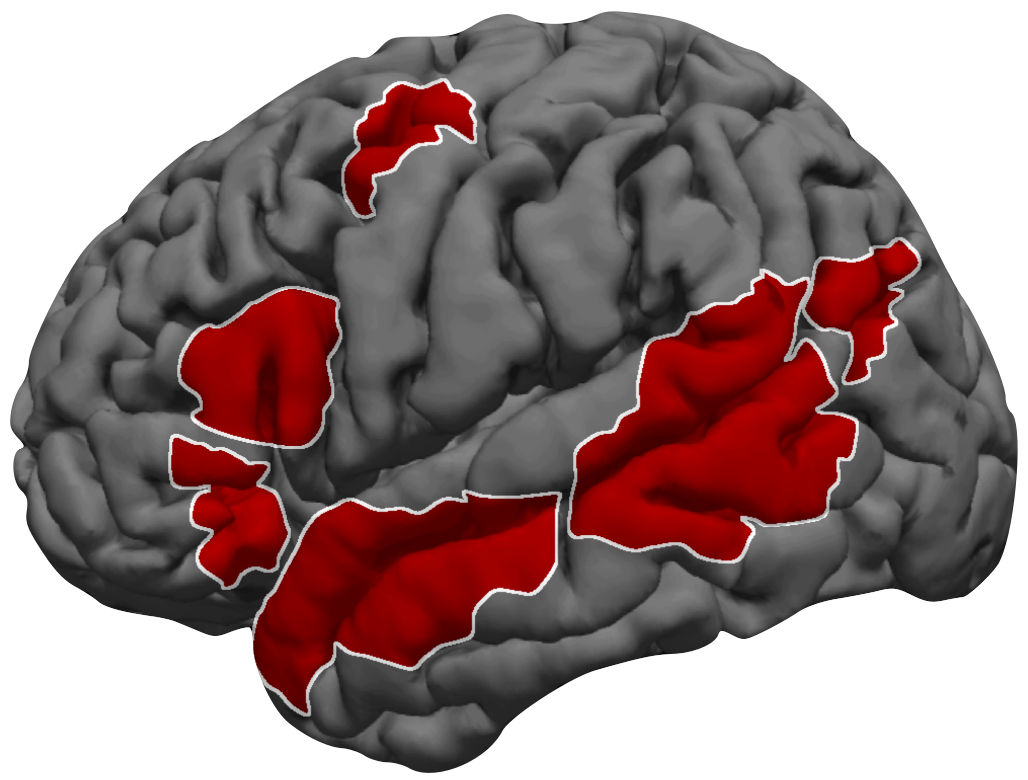


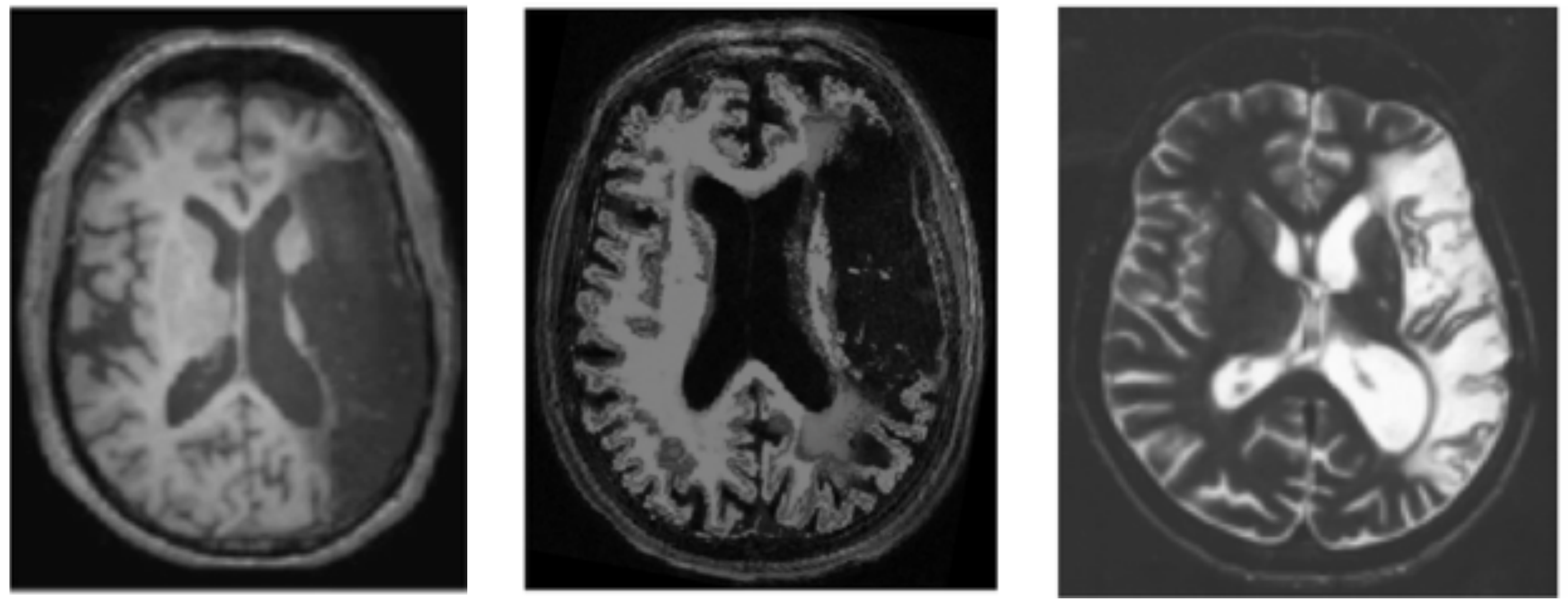
slide adapted from Ev Fedorenko; for a review, see Fedorenko, Ivanova & Regev, 2024

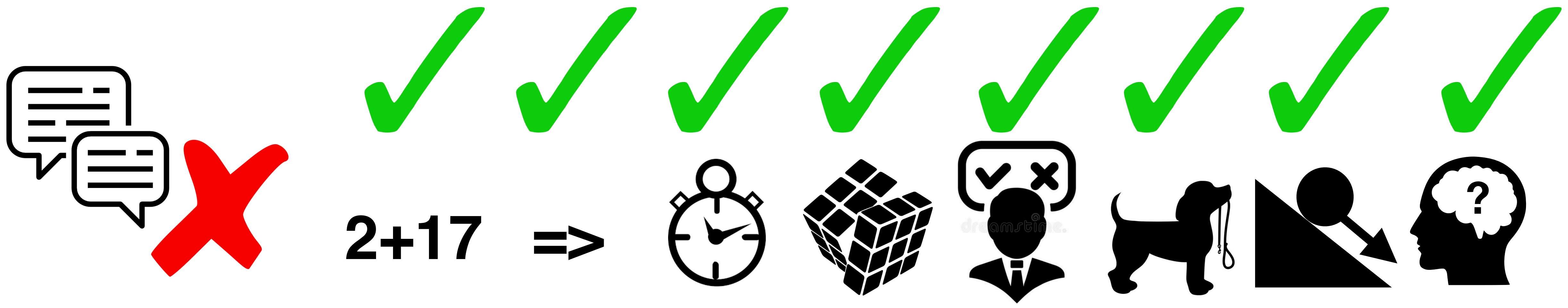
Language and the brain

Language areas can be **damaged** with little/no effect on thought-related activities.

Sample patients' lesions:







etc.

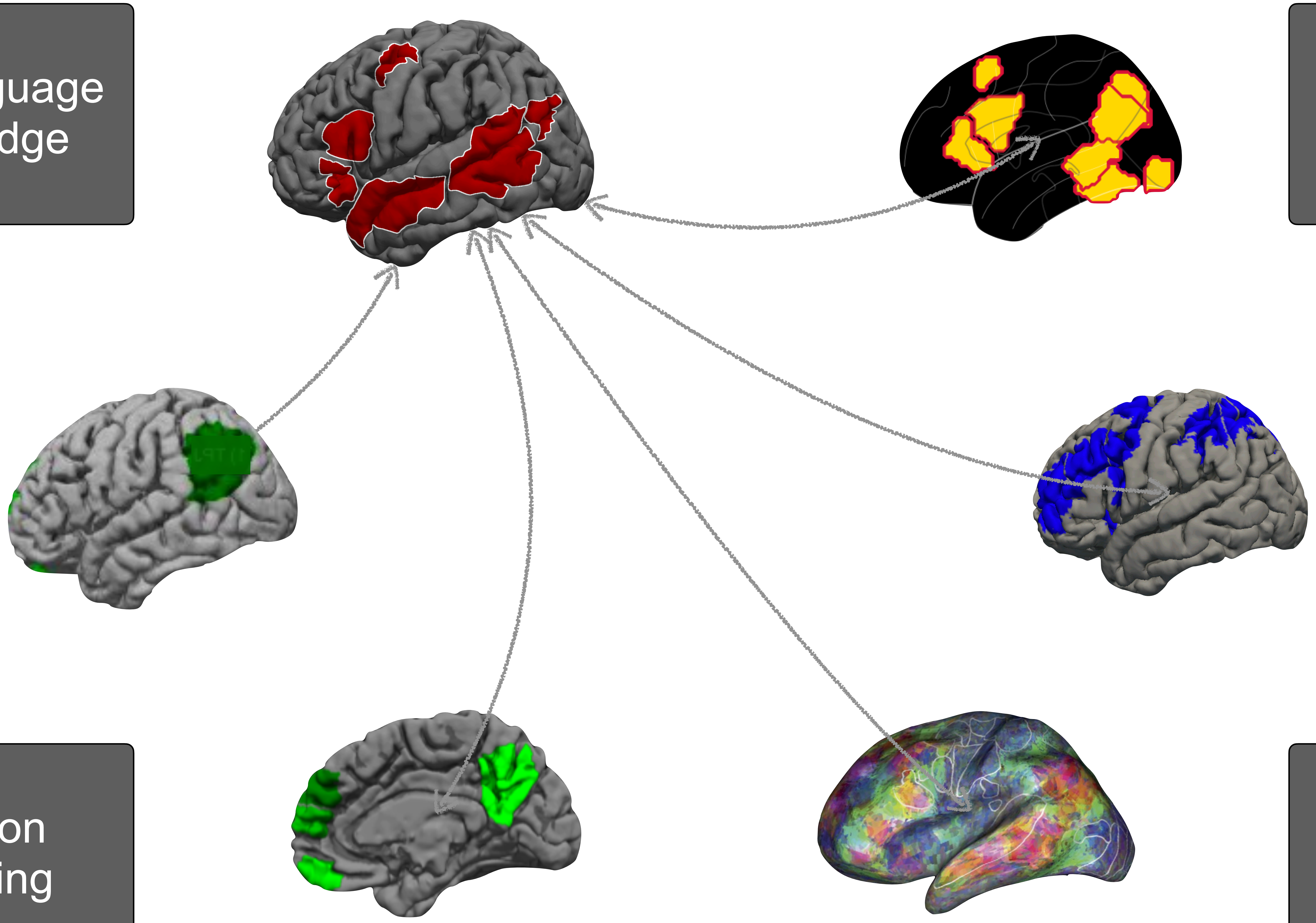
slide adapted from Ev Fedorenko; for a review, see Fedorenko, Ivanova & Regev, 2024

Formal and functional linguistic competence

core language knowledge

semantic tasks

social knowledge



general cognitive tasks

situation modeling

world knowledge

Mahowald, Ivanova et al, 2024

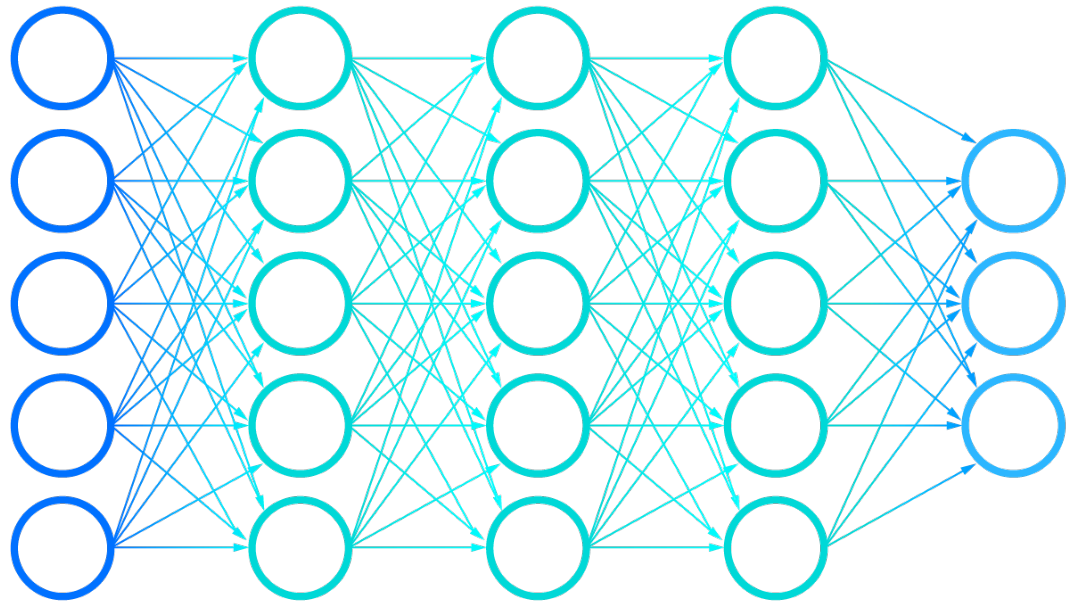
Formal and functional linguistic competence

core language knowledge

semantic tasks

social knowledge

general cognitive tasks



situation modeling

world knowledge

Mahowald, Ivanova* et al, 2024*

Formal and functional linguistic competence

FORMAL COMPETENCE
(language-specific)

core language knowledge

FUNCTIONAL COMPETENCE
(non-language-specific)

semantic tasks

social knowledge

general cognitive tasks

situation modeling

world knowledge

Mahowald, Ivanova* et al, 2024*

Formal and functional linguistic competence

FORMAL COMPETENCE
(language-specific)

The keys to the cabinet **are** on the table.

Easy for language models starting with GPT2/3

A remarkable scientific and engineering breakthrough

Not something linguists were expecting

FUNCTIONAL COMPETENCE
(non-language-specific)

Six birds were sitting on a tree. Three flew away, but then one came back. There are now **four** birds.

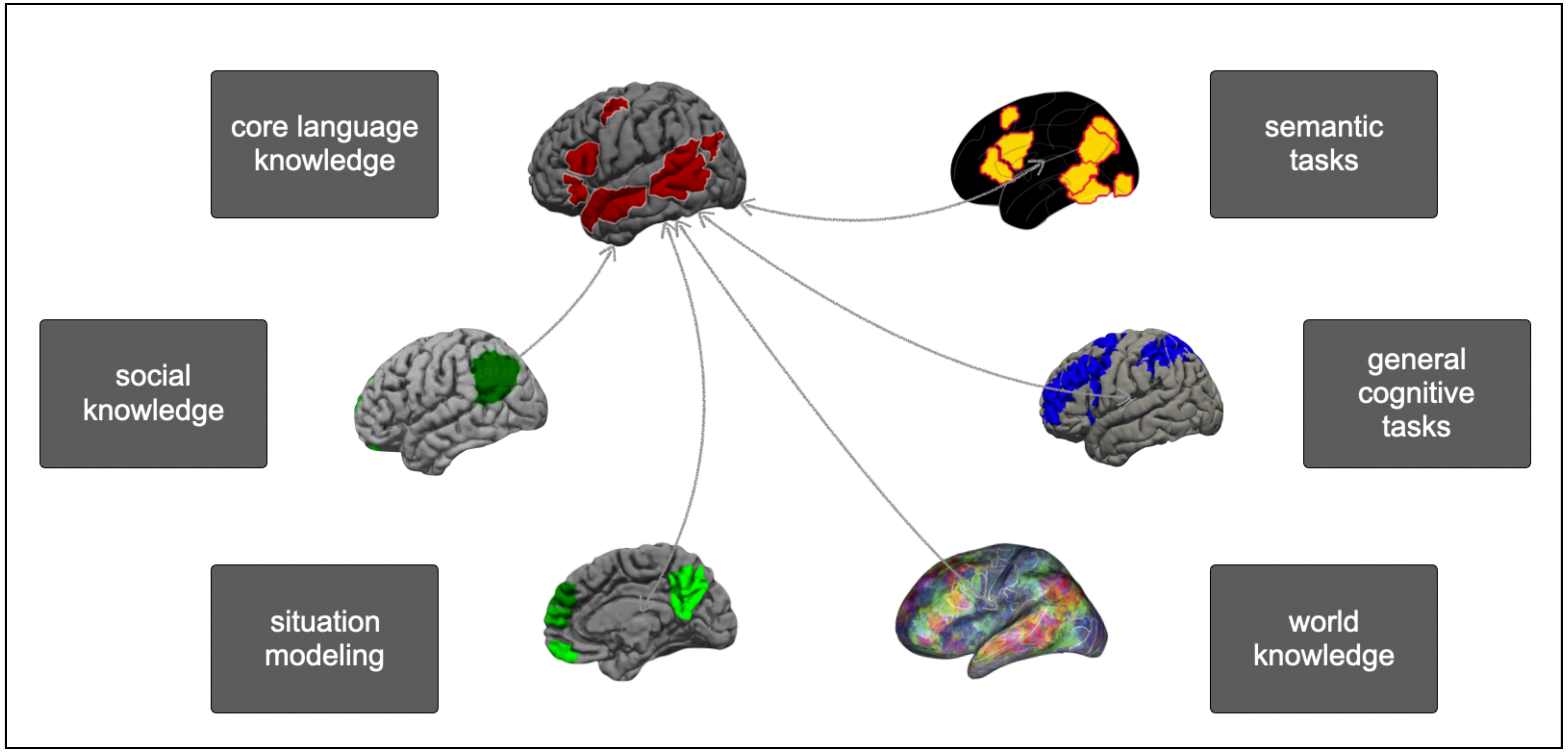
Can be challenging for language models!

Performance might rely on memorization or heuristics

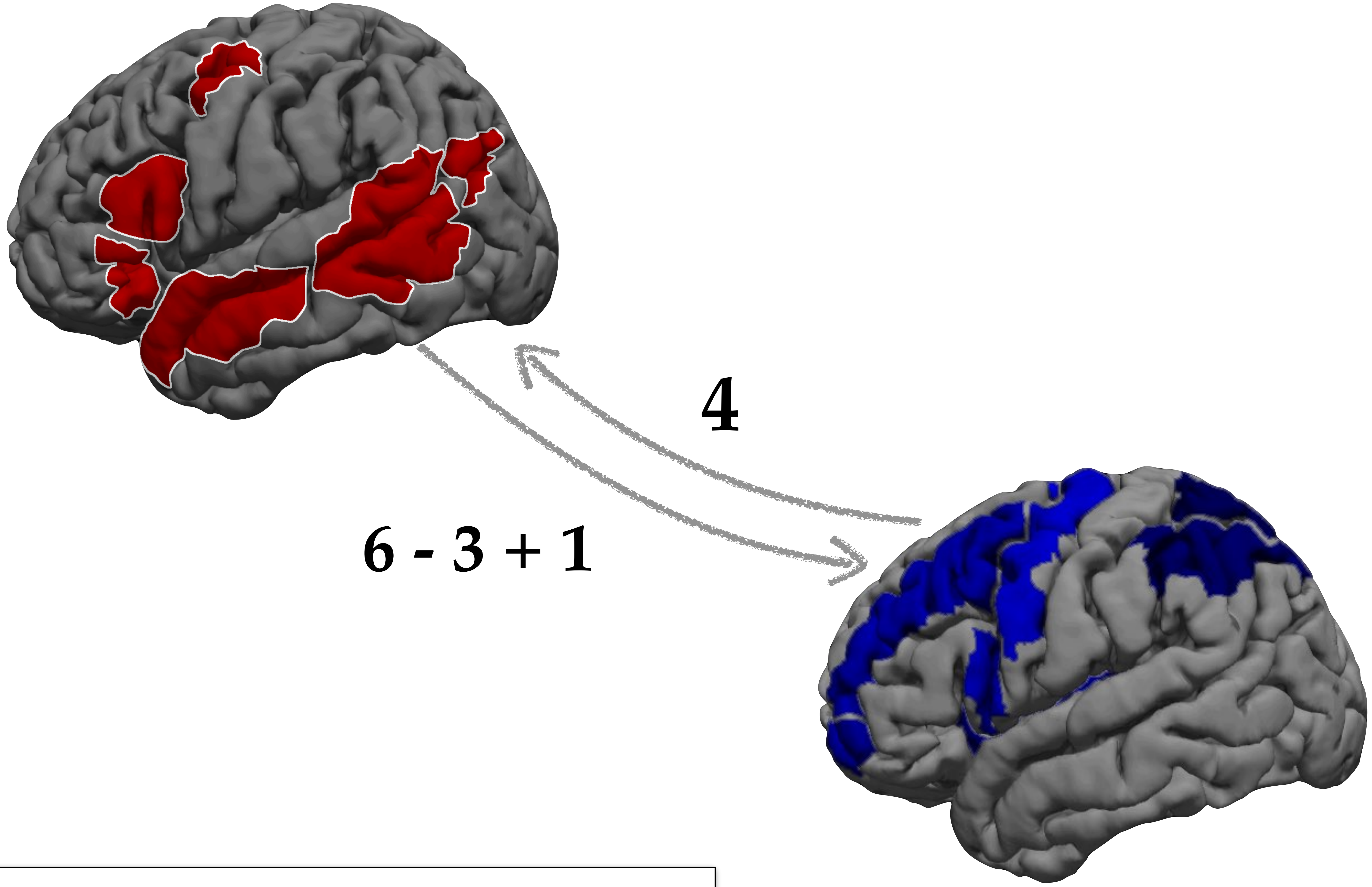
Progress requires shifting away from pure next-word prediction to fine-tuning or additional modules

Mahowald, Ivanova* et al, 2024*

A humanlike AI system would look like this…



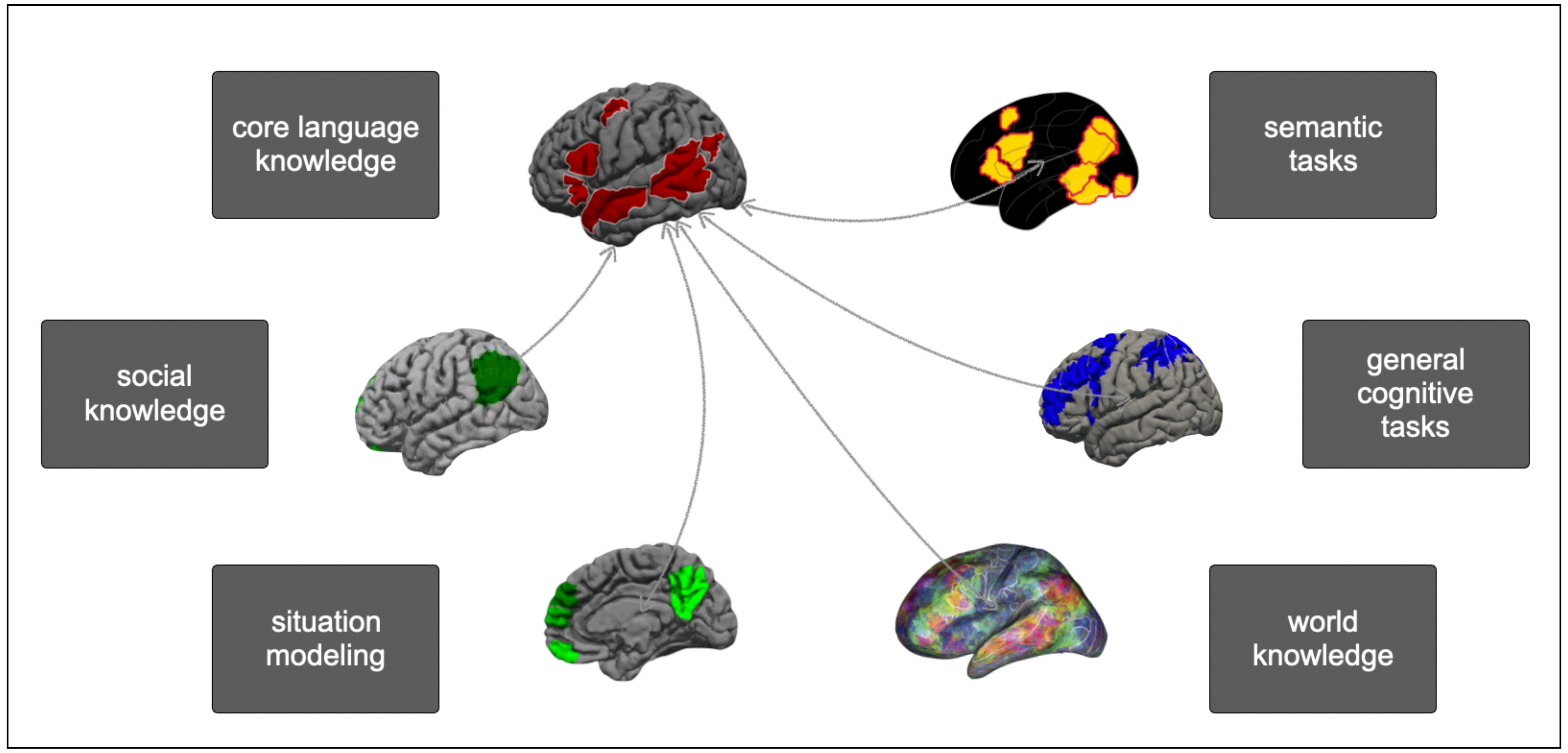
Mahowald, Ivanova et al, 2024



Six birds were sitting on a tree. Three flew away, but then one came back.

How many birds are there now?

A humanlike AI system would look like this…



Options:

- **architectural modularity**: macro structure is built in
- **emergent modularity**: macro structure arises during training (MoE-like)
- **status quo**: possible emergence of implicit structure?

Mahowald, Ivanova et al, 2024

Roadmap

Formal vs functional linguistic competence

It gets complicated: generalized world knowledge

Moving forward

Roadmap

It gets complicated:
generalized world knowledge

1. **Why it's complicated**
2. **Generalized event knowledge**
3. **Elements of World Knowledge (EWoK)**
4. **Yes-bias**

Large language models and world knowledge

Language contains a wealth of information about the world

FACTUAL

Paris is the capital of France

Birds lay eggs

DISTRIBUTIONAL

The sky is blue today

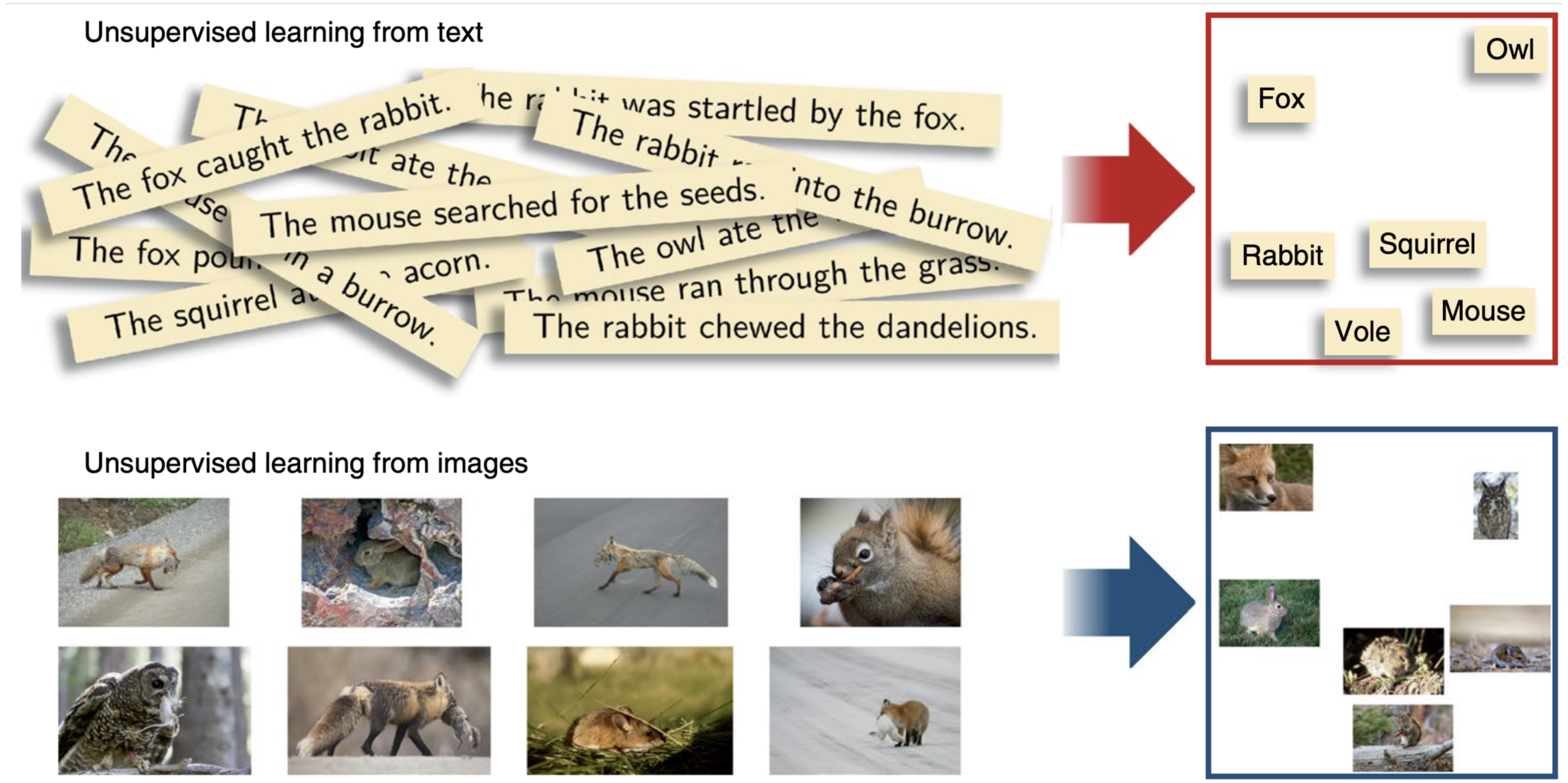
The sky was pitch black

The sky is pink

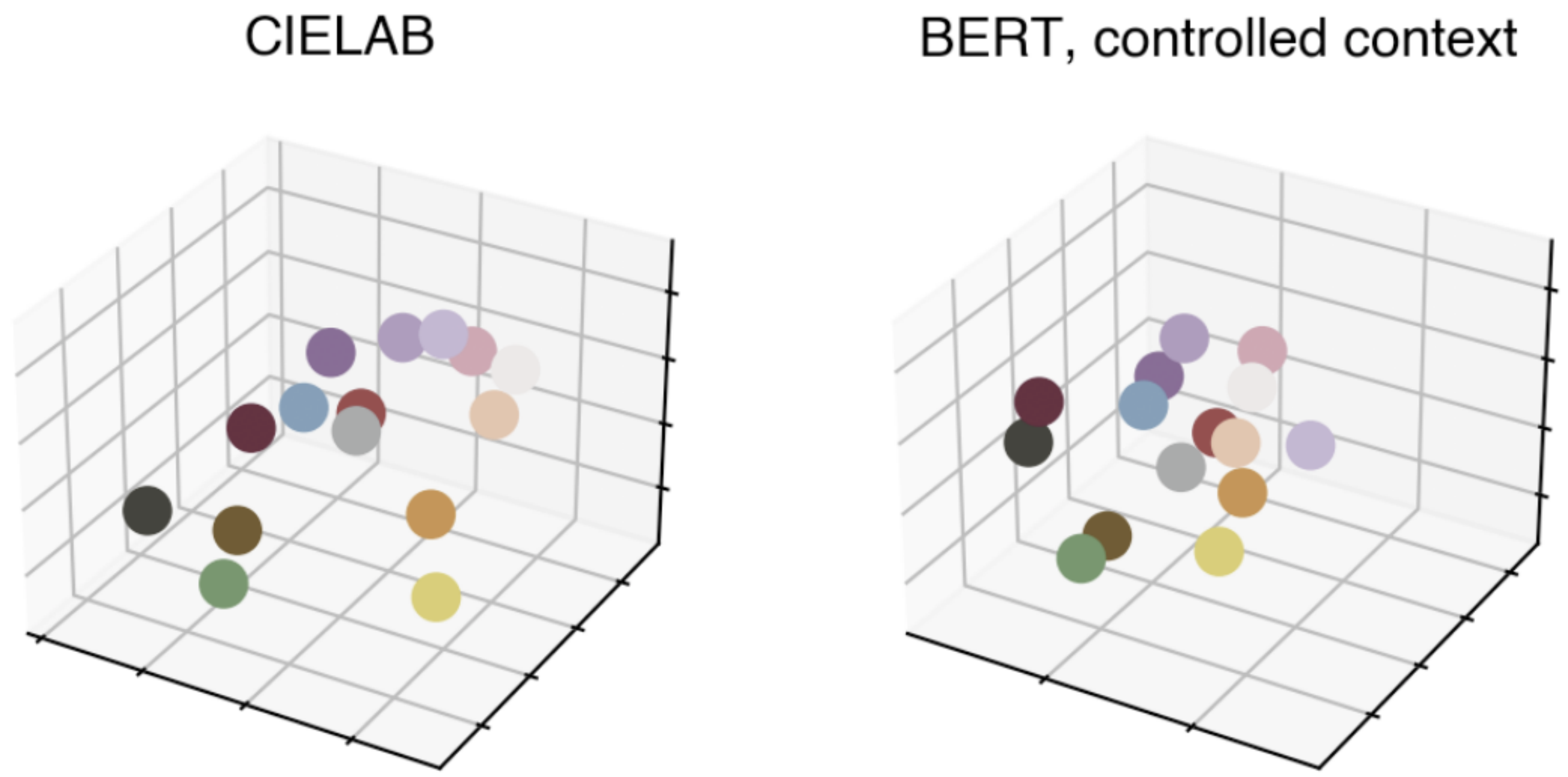
Large language models and world knowledge

Distributional information from language aligns with that from other domains

Roads & Love, 2020; Luo et al, 2024



Abdou et al, 2021



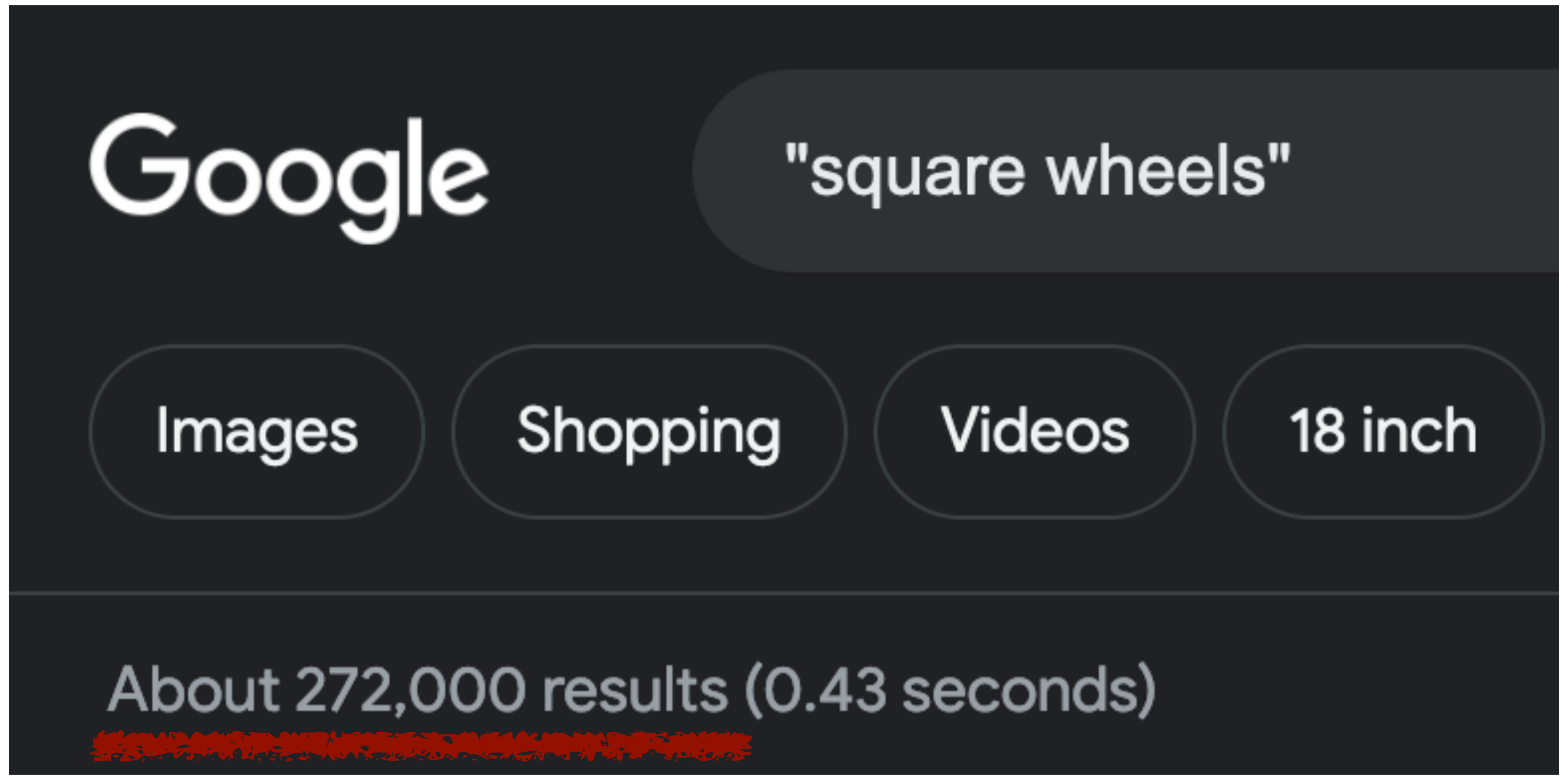
Large language models and world knowledge

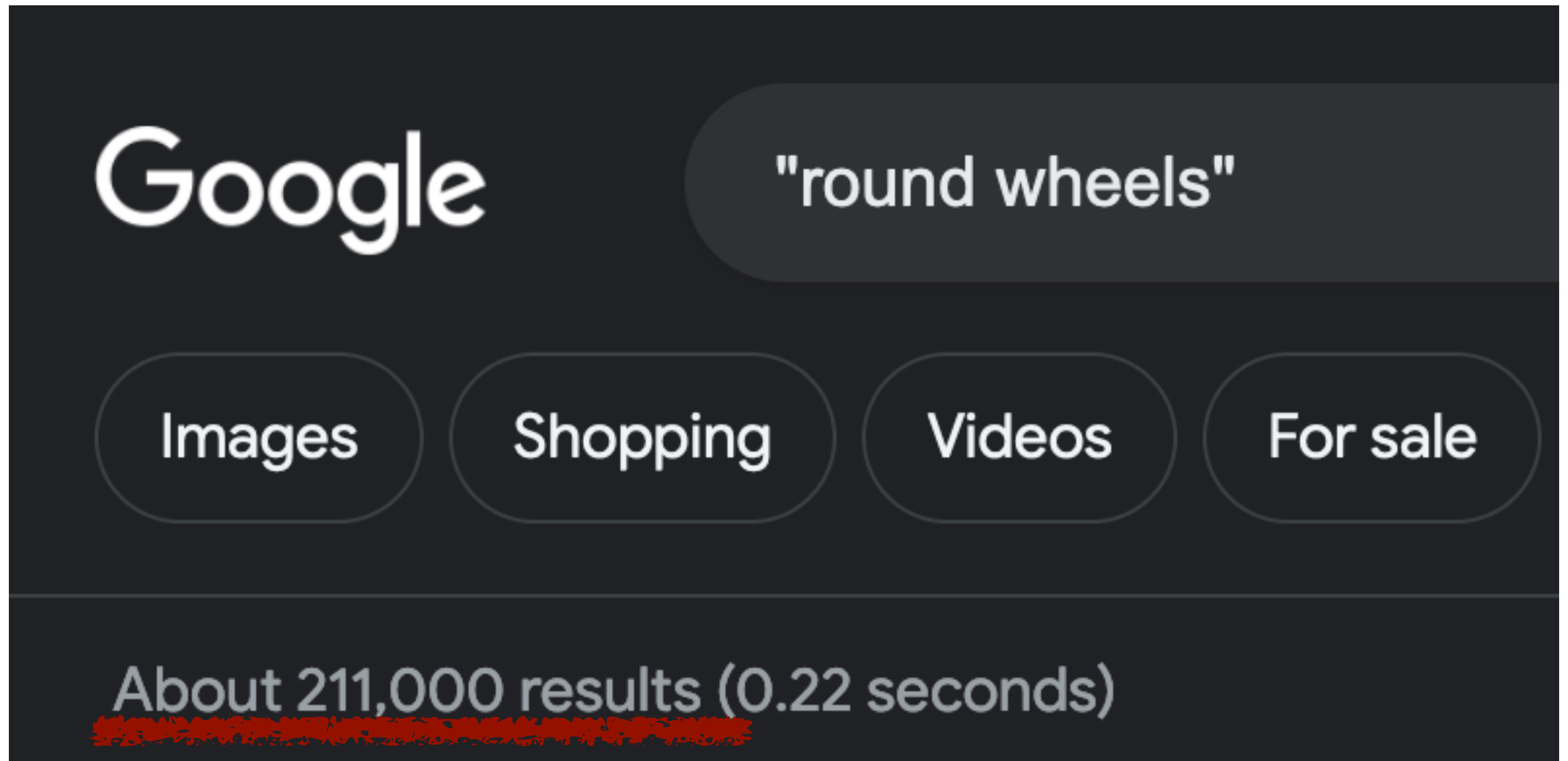
Distributional information from language aligns with that from other domains

… but it’s biased

reporter bias

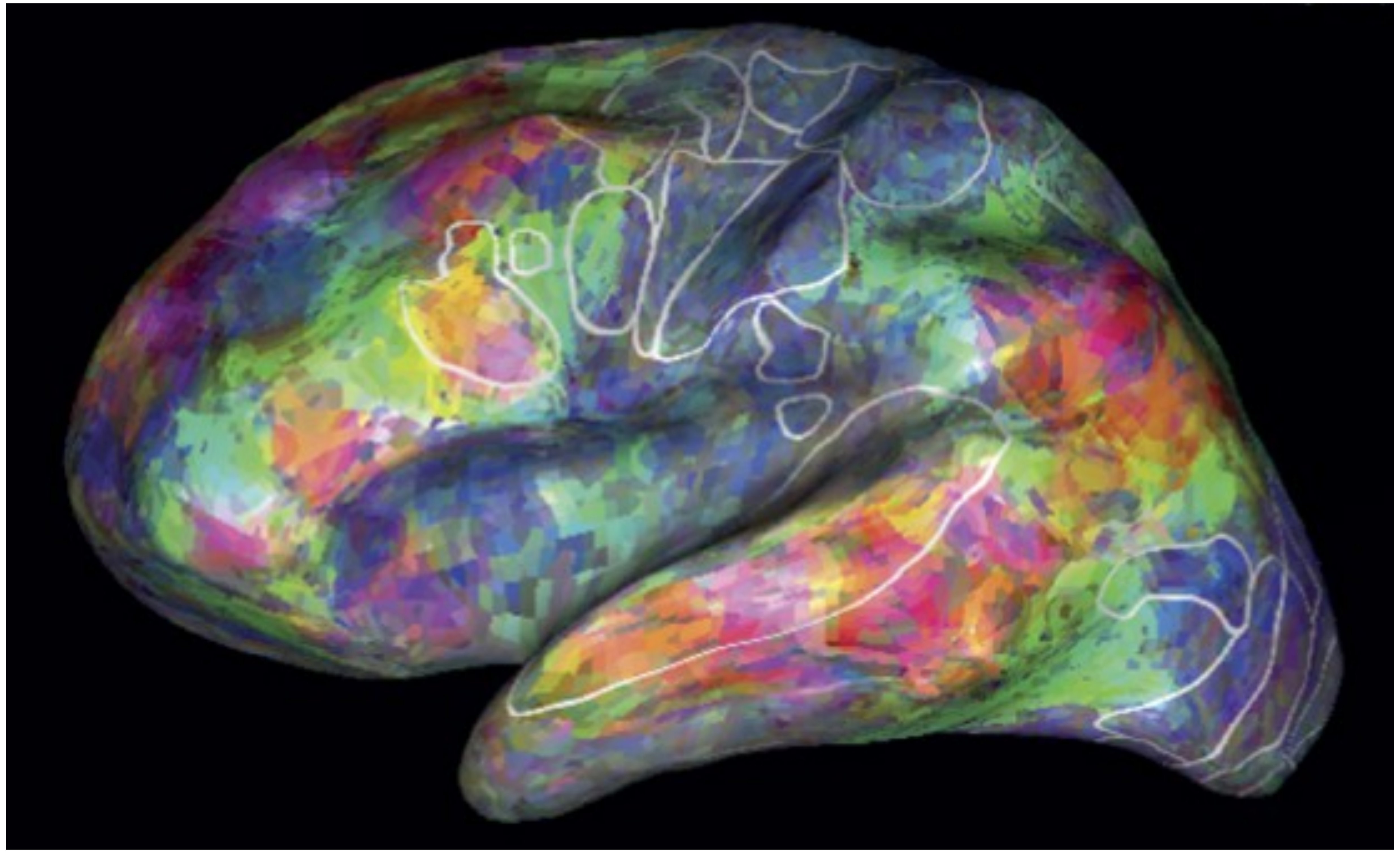
Gordon & van Durme, 2013; Schwartz & Choi, 2020





Large language models and world knowledge

Huth et al., 2016



Roadmap

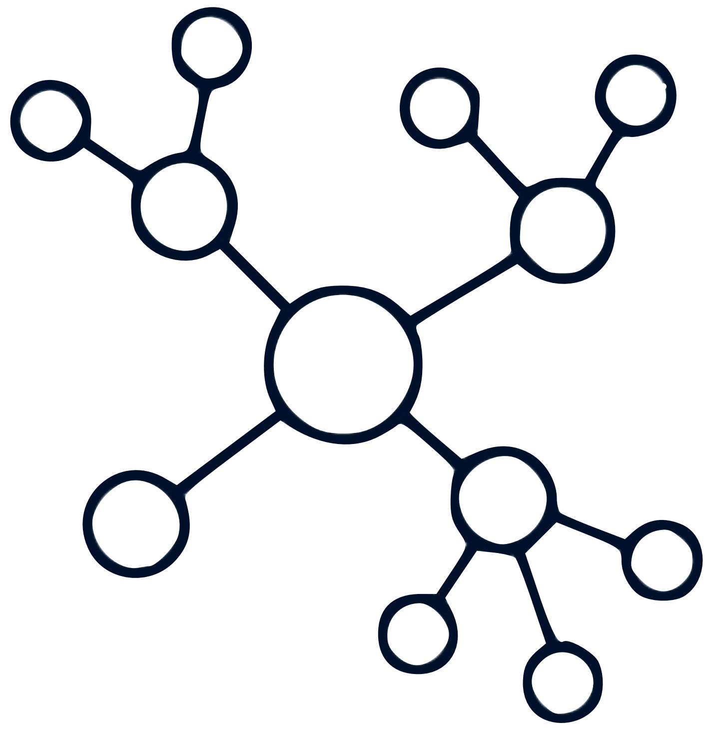
It gets complicated:
generalized world knowledge

1. Why it’s complicated
2. Generalized event knowledge
3. Elements of World Knowledge (EWoK)
4. Yes-bias

Language and event knowledge

Generalized Event Knowledge (GEK; McRae & Matsuki 2009)

- storage of **templates** of **common events** observed in the world



The fox chased the rabbit.

The rabbit chased the fox.

The fox chased the planet.

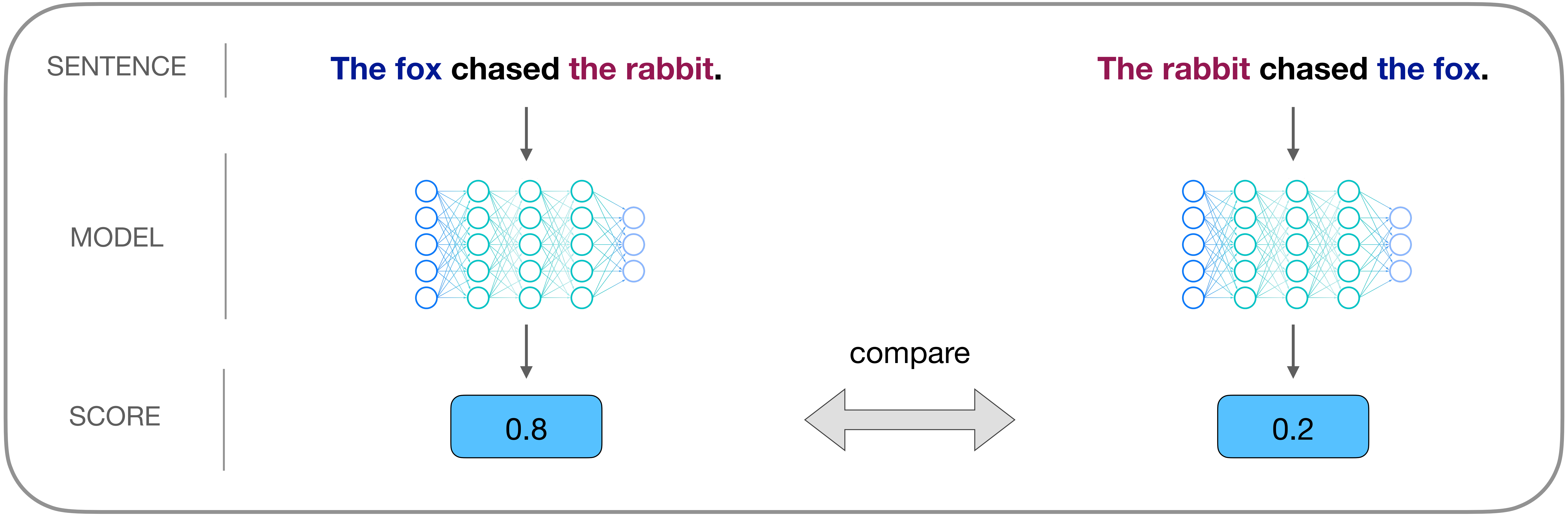
Does generalized event knowledge naturally arise
in pretrained language models?

Language models and event knowledge

co-lead:
Carina Kauf

Does generalized event knowledge naturally arise
in pretrained language models?

Approach: minimal sentence pairs



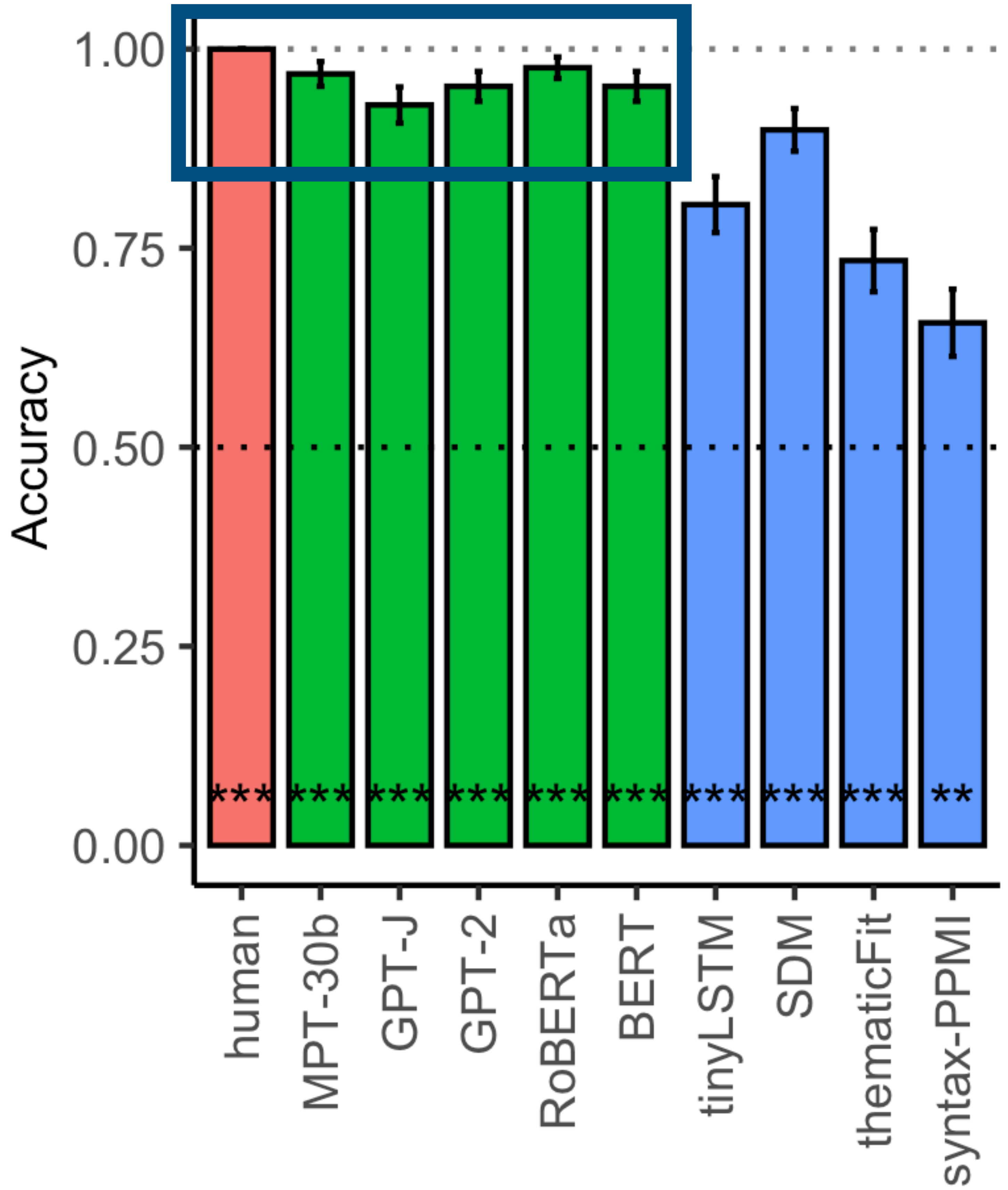
Kauf, Ivanova* et al, 2023; Kauf & Ivanova, 2023, ACL*

Language models and event knowledge

Animate-Inanimate, impossible

The teacher bought the laptop.

The laptop bought the teacher.

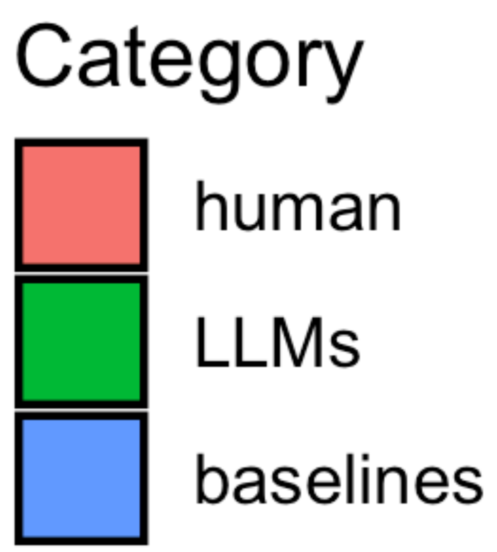


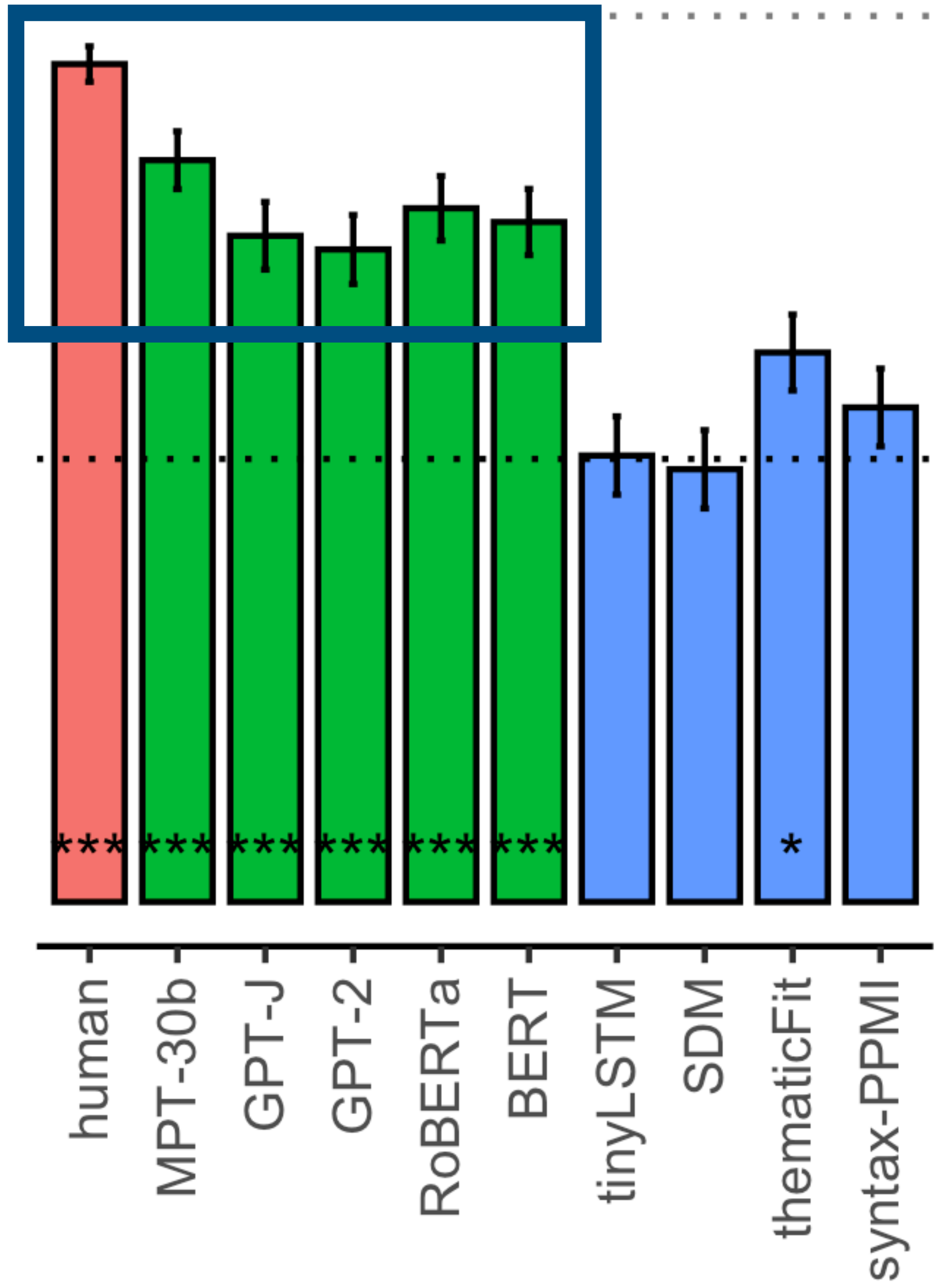
Animate-Animate, unlikely

The fox chased the rabbit.

The rabbit chased the fox.

"the gap between the impossible and the unlikely"





Language models and event knowledge

Animate-Inanimate, impossible

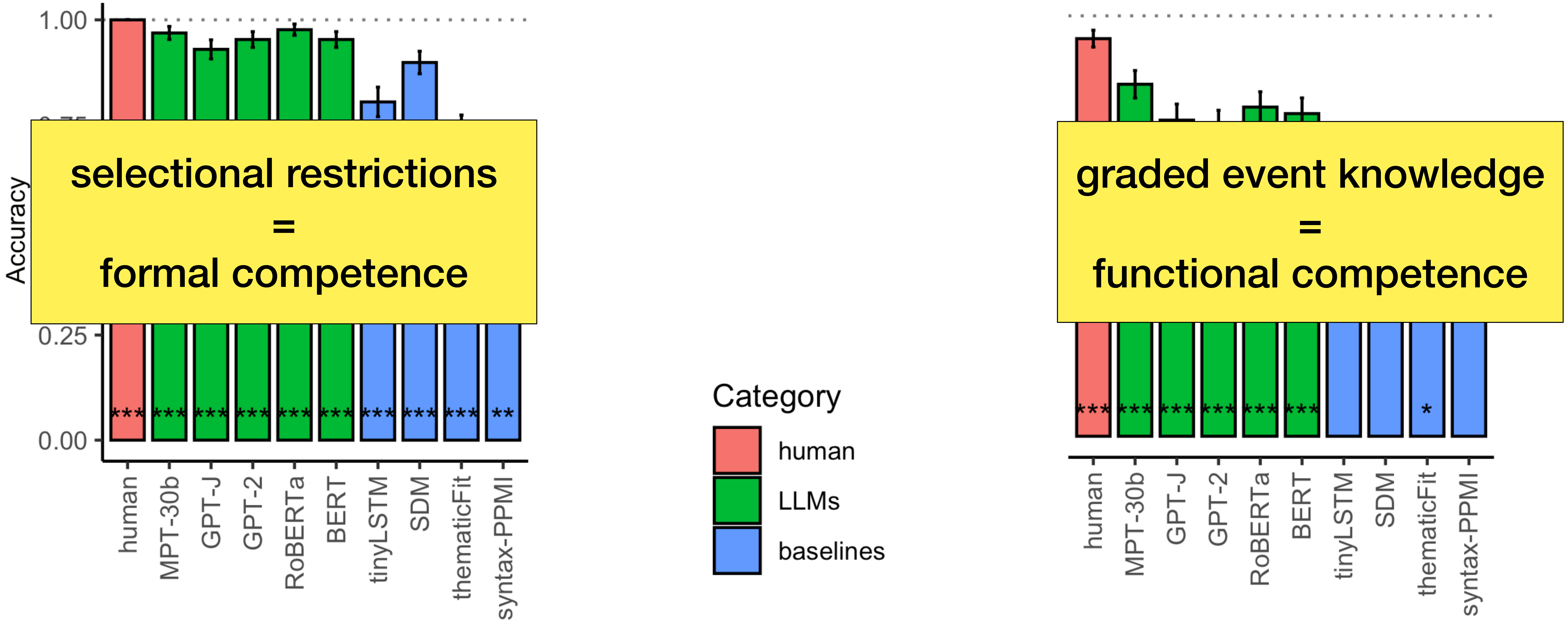
The teacher bought the laptop.

The laptop bought the teacher.

Animate-Animate, unlikely

The fox chased the rabbit.

The rabbit chased the fox.



selectional restrictions
=
formal competence

graded event knowledge
=
functional competence

Event semantics in language models

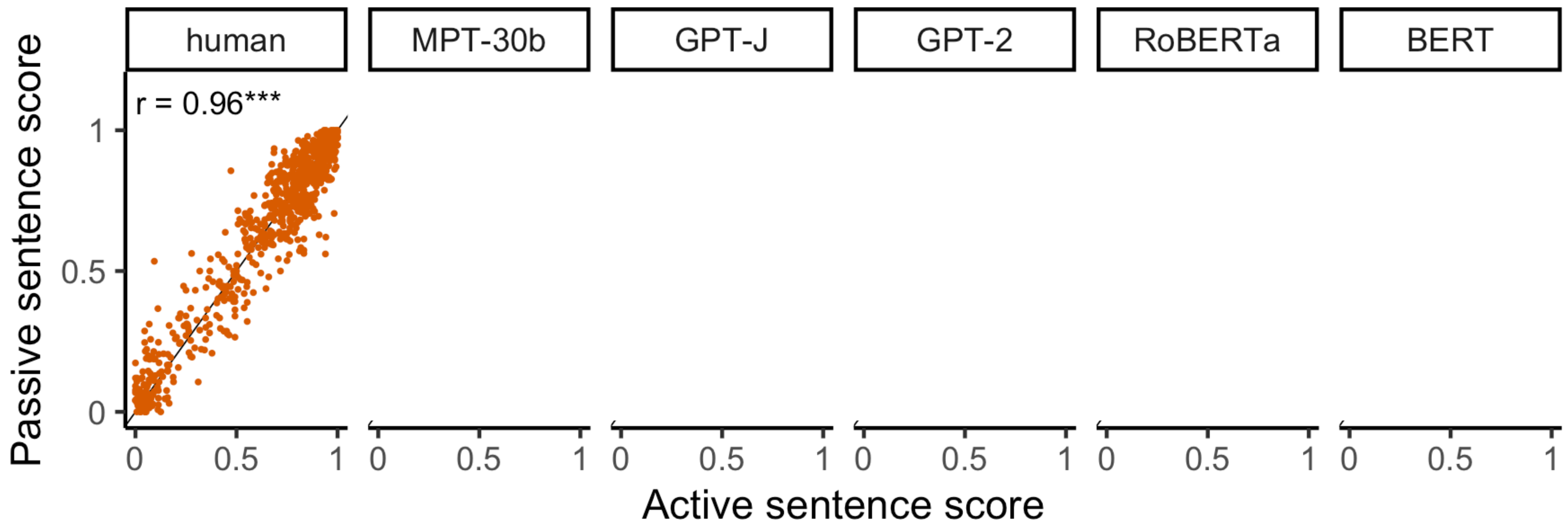
Generalizability

syntactic generalization ✓

The author finished the novel.

vs.

The novel was finished by the author.

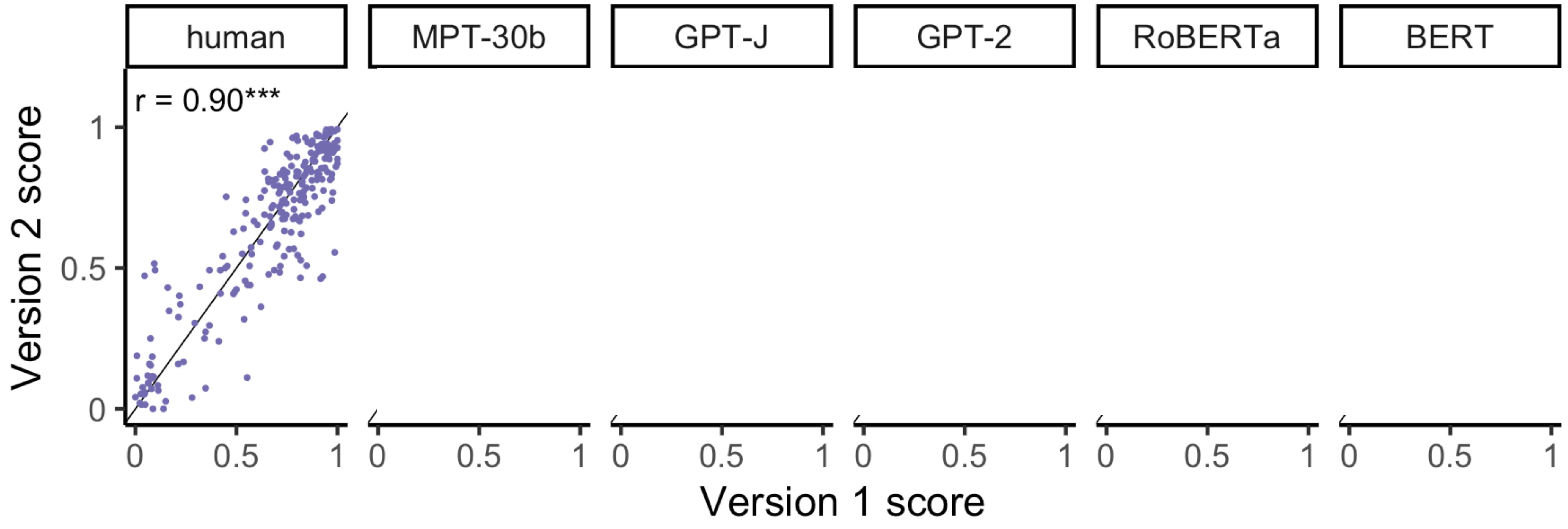


semantic generalization ?

The author finished the novel.

vs.

The writer completed the book.



Kauf, Ivanova* et al, 2023*

Language models and event knowledge

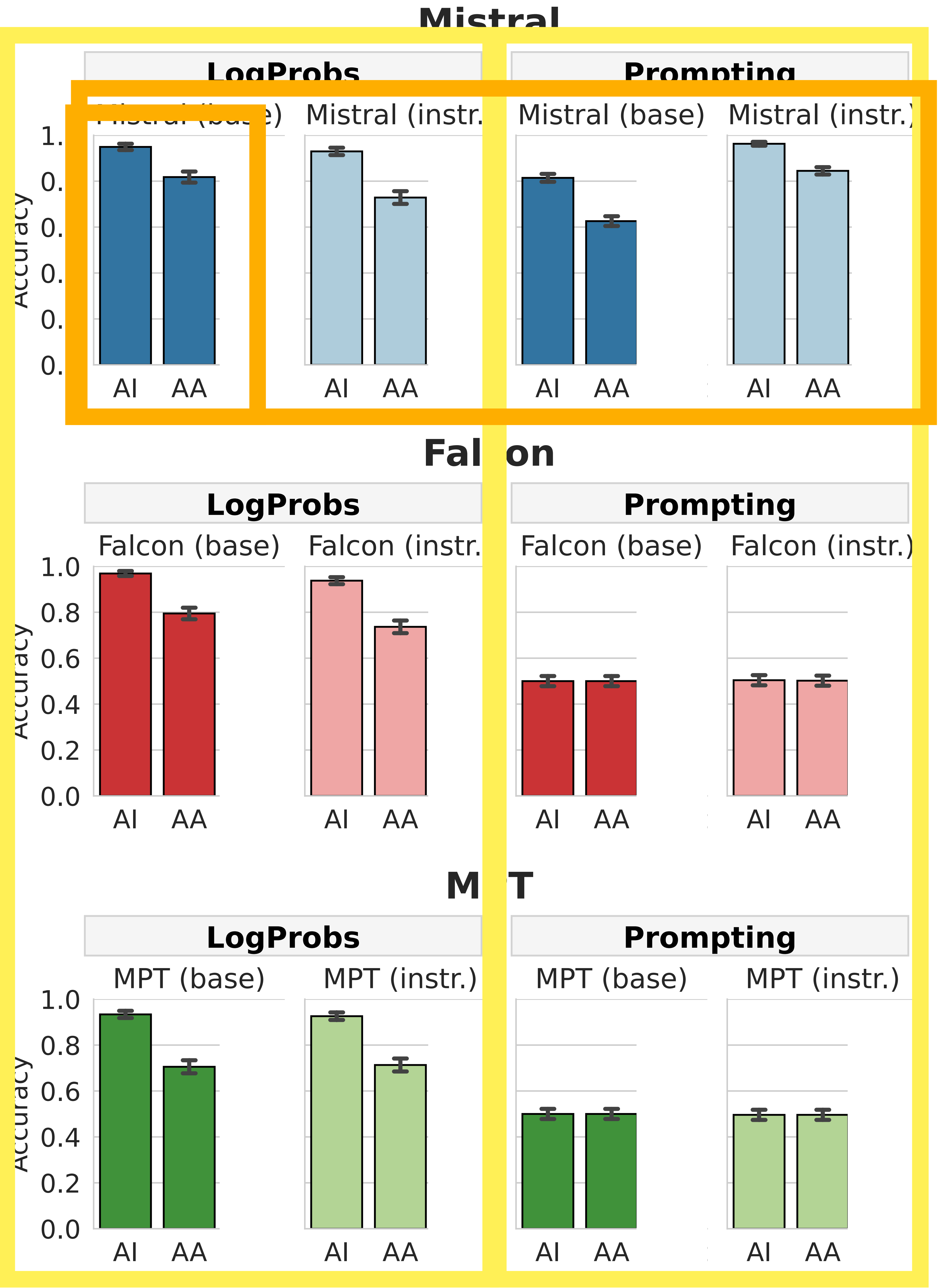
What if, instead of evaluating the *LogProb* of the sentence under the model, we ask the models directly (*Prompting*)?

And what if we evaluate not just pretrained (*base*) models, but also *instruction-tuned* models?

The 'impossible-unlikely' (AI-AA) gap remains.

LogProbs are consistent across models, whereas prompting is hit-or-miss.

Conclusions from base models hold for instruction-tuned models.



Kauf et al, 2024, best paper at BlackBoxNLP workshop

Language models and event knowledge

LLMs systematically distinguish
possible and impossible events but are less consistent with
likely vs. unlikely events.

Roadmap

It gets complicated:
generalized world knowledge

1. Why it's complicated
2. Generalized event knowledge
3. Elements of World Knowledge (EWoK)
4. Yes-bias

Elements of World Knowledge (EWoK)

ewok-core.github.io

co-lead:
Aalok Sathe

co-lead:
Ben Lipkin

Ivanova ,Sathe*, Lipkin*, and a team of cognitive scientists, arXiv*

Social Interactions

Social Properties

Social Relations

Physical Interactions

Physical Dynamics

Physical Relations

Material Dynamics

Material Properties

Agent Properties

Quantitative Properties

Spatial Relations

Domains *have* ***Concepts*** *that are tested using several* ***Templates***

concept: TEACHER

concept: STUDENT

template:

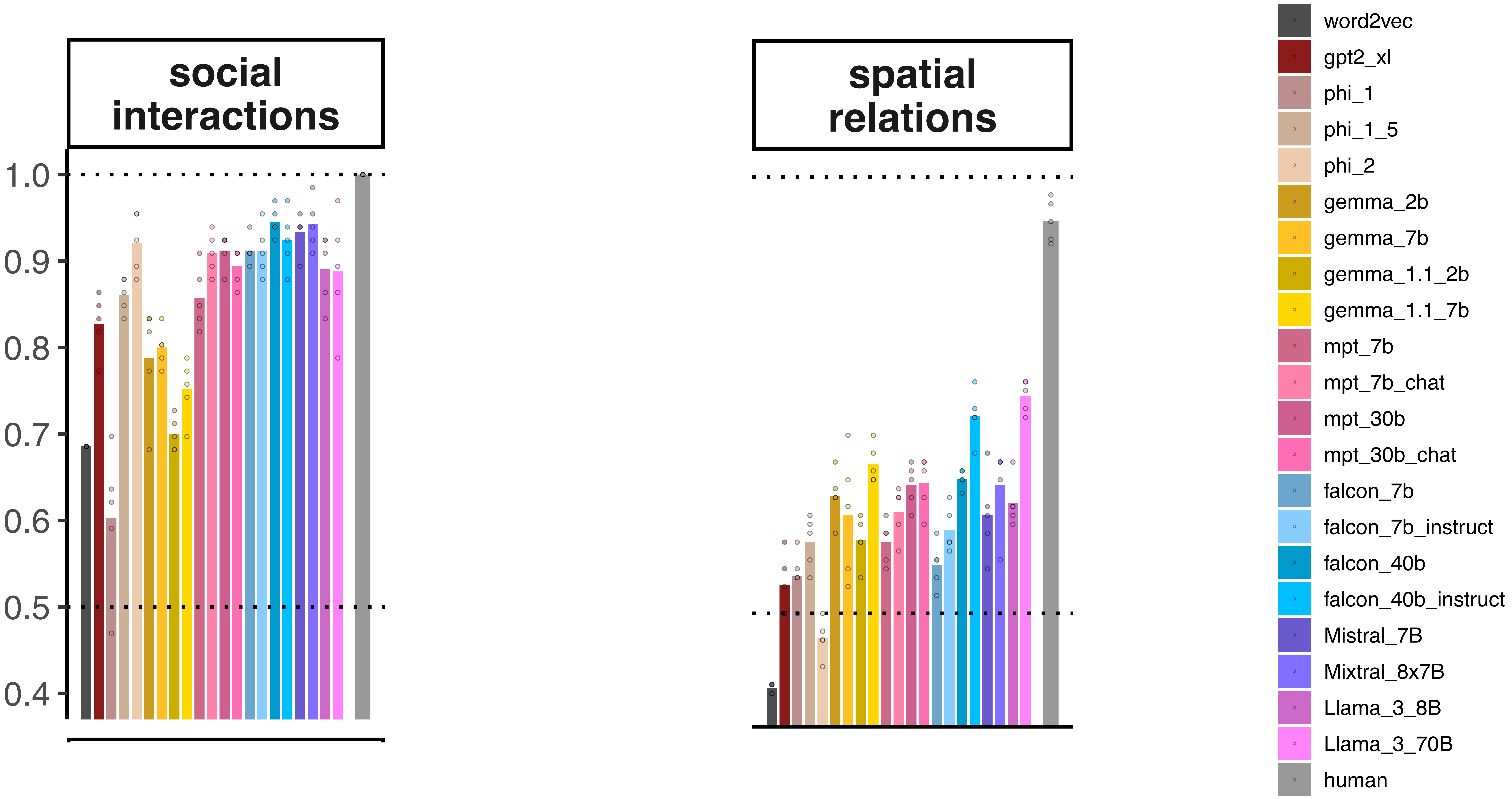
C1: AGENT-1 *assigns* homework to AGENT-2.
C2: AGENT-1 *submits* homework to AGENT-2.

T1: AGENT-1 is AGENT-2's ***teacher***.
T2: AGENT-1 is AGENT-2's ***student***.

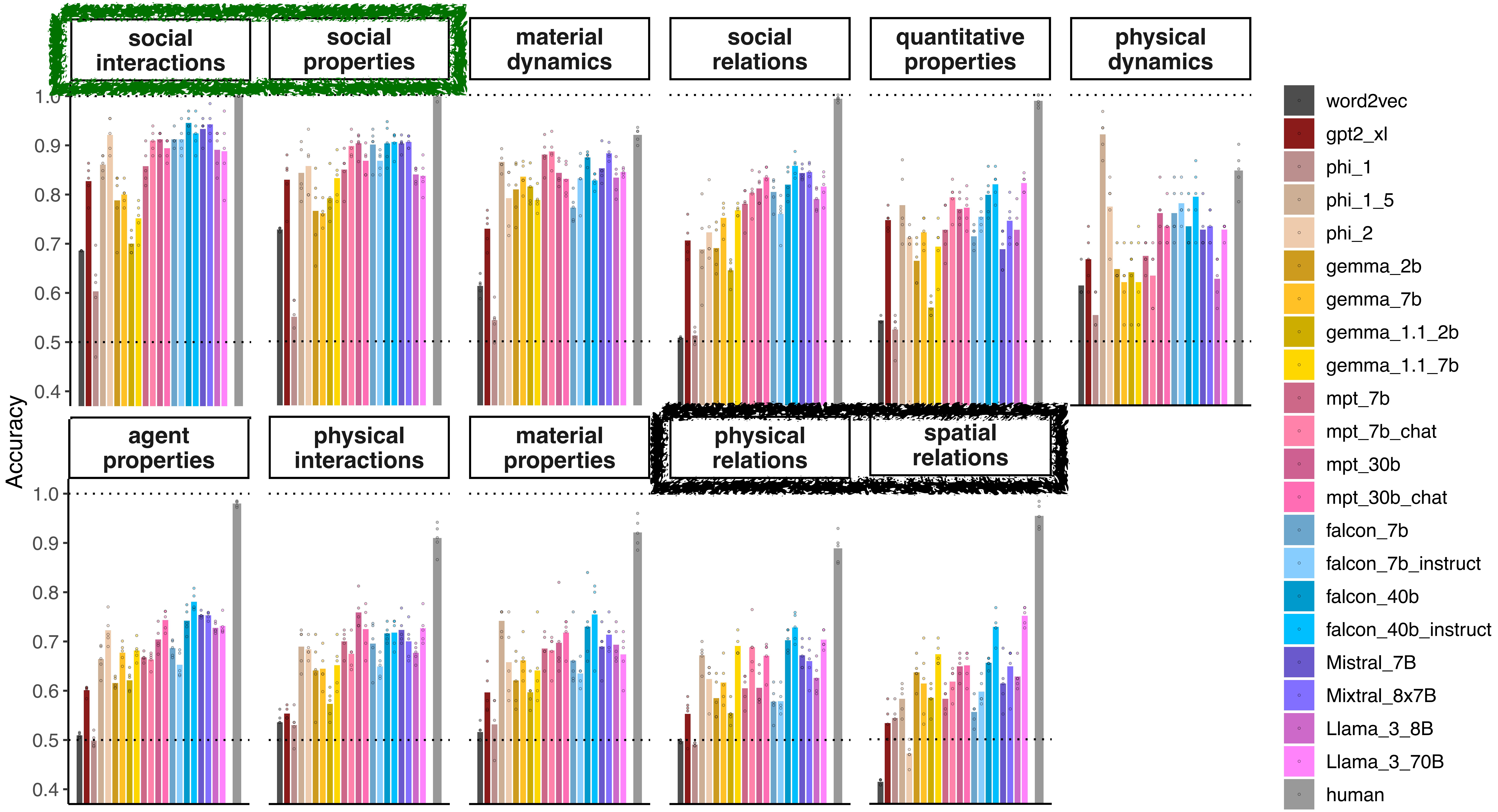
✓ **C1-T1**: Fatima *assigns* homework to Jose. Fatima is Jose's ***teacher***.
✗ **C1-T2**: Fatima *assigns* homework to Jose. Fatima is Jose's ***student***.
✓ **C2-T2**: Fatima *submits* homework to Jose. Fatima is Jose's ***student***.
✗ **C2-T1**: Fatima *submits* homework to Jose. Fatima is Jose's ***teacher***.

Templates *give rise to* ***Items***

Ivanova ,Sathe*, Lipkin*, and a team of cognitive scientists, arXiv*



Ivanova ,Sathe*, Lipkin* et al, arXiv*



Ivanova ,Sathe*, Lipkin* et al, arXiv*

Elements of World Knowledge (EWoK)

Ivanova ,Sathe*, Lipkin* et al, arXiv*

EWoK in BabyLM

- EWoK-core-1.0 served as a test-only benchmark

	Model	BLiMP	BLiMP Supplement	(Super)GLUE	EWoK	*Text Average*	VQA	Winoground	DevBench	*Vision Average*
Strict	GPT-BERT	**86.1**	**76.8**	**81.5**	**58.4**	**75.7**	–	–	–	–
	BabbleGPT	77.9	69.5	71.7	52.0	67.8				
	MLSM	69.6	65.4	74.8	52.6	65.6	–	–	–	–
	Best baseline: LTG-BERT	69.2	66.5	68.4	51.9	64.8	–	–	–	–
Strict-small	GPT-BERT	81.2	69.4	76.5	54.6	70.4	–	–	–	–
	DeBaby	74.2	63.7	73.7	54.3	66.5	–	–	–	–
	BabyLlama-2	71.8	63.4	70.2	51.5	64.2	–	–	–	–
	Best baseline: BabyLlama	69.8	59.5	63.3	50.7	61.6	–	–	–	–
Multimodal	GIT-1vd125	66.5	60.9	65.6	52.2	61.3	51.9	**57.8**	48.1	52.6
	Wake/Sleep	73.6	55.6	64.7	51.4	61.3	42.0	50.9	22.8	38.6
	$\text{Flamingo}_{\text{CL}}$	60.1	53.3	64.3	50.7	57.1	40.9	50.8	47.3	46.3
	Best baseline: Flamingo	70.9	65.0	69.5	52.7	65.2	**52.3**	51.6	**59.5**	**54.5**

Table 3: Macro averages for each benchmark across the top-performing systems (by overall score), best baseline, and skylines.

Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora

Michael Y. Hu[1] Aaron Mueller[2,3] Candace Ross[4]
Adina Williams[4,7] Tal Linzen[1] Chengxu Zhuang[6] Ryan Cotterell[8]
Leshem Choshen[5,6] Alex Warstadt[8] Ethan Gotlieb Wilcox[9]
[1]New York University [2]Northeastern University [3]Technion [4]Meta AI (FAIR)
[5]IBM Research [6]MIT [7]ML Commons
[8]ETH Zürich [9]Georgetown University
michael.hu@nyu.edu

Elements of World Knowledge (EWoK)

ewok-core.github.io

Basic world knowledge in LLMs varies drastically by domain, with social knowledge > physical and spatial knowledge.

EWoK is not just one dataset; it's a framework. So do consider adding to it!

Ivanova ,Sathe*, Lipkin* et al, arXiv*

Roadmap

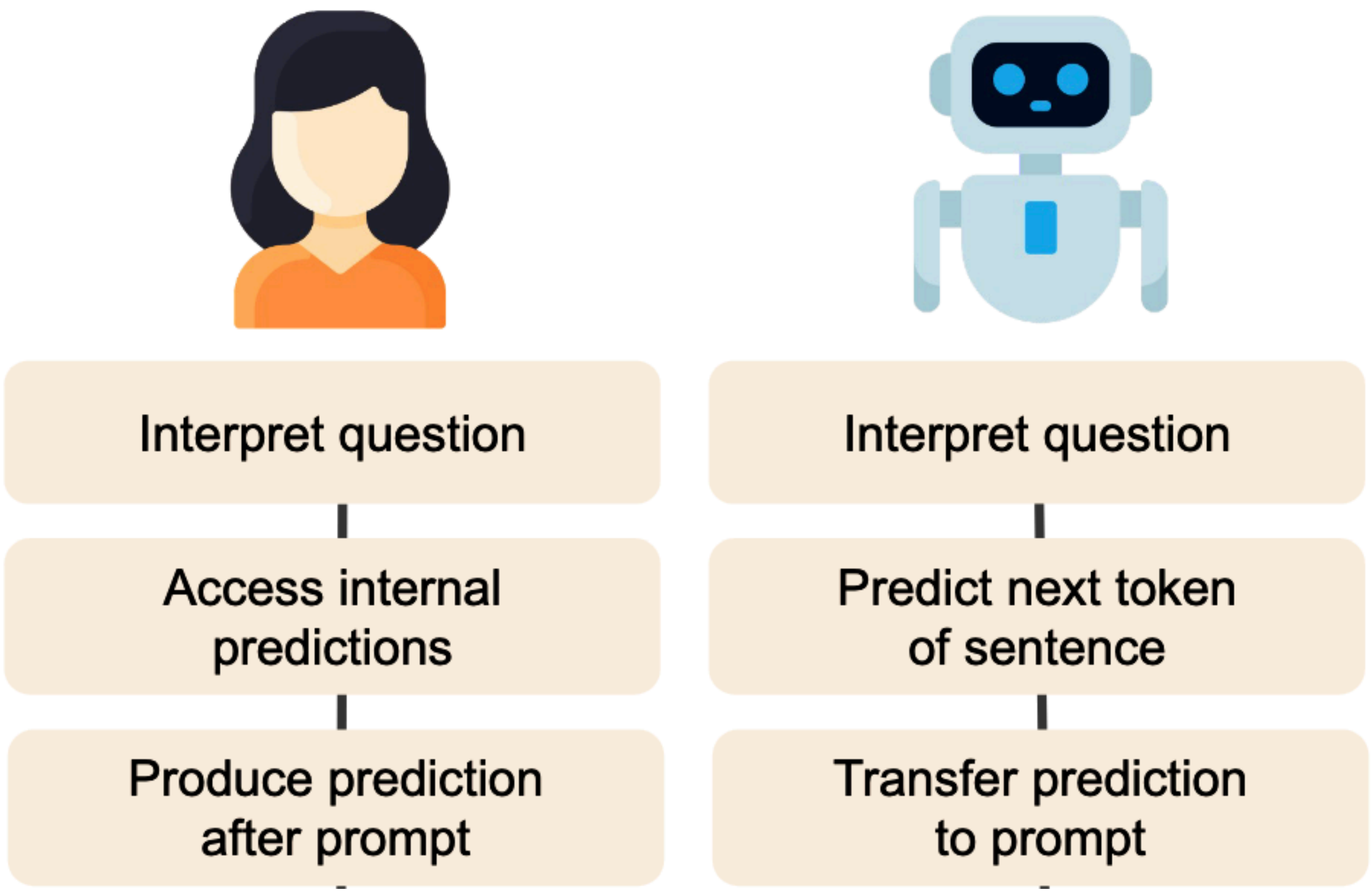
It gets complicated:
generalized world knowledge

1. Why it's complicated
2. Generalized event knowledge
3. Elements of World Knowledge (EWoK)
4. Yes-bias

Yes-no bias in language models

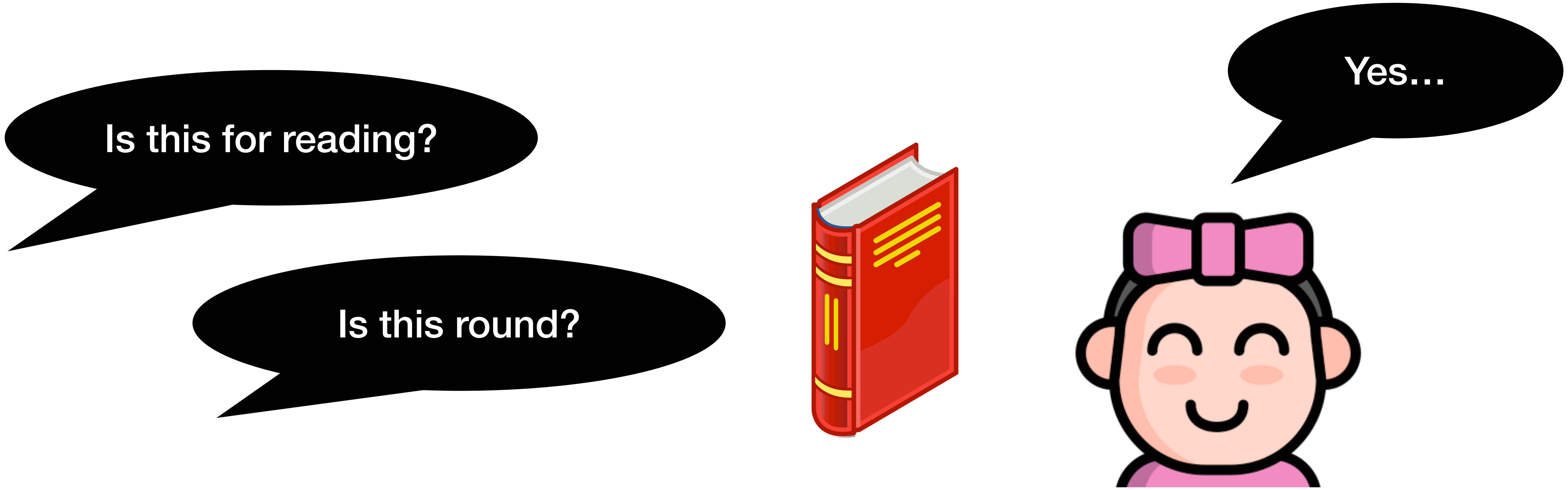
- Background: task demands affect performance in LLMs and humans

Hu & Frank, 2024



Yes-no bias in language models

- Humans tend to exhibit a yes-bias (acquiescence bias)



example from Okanda & Itakura, 2008

Yes-no bias in language models

Om Bhatt

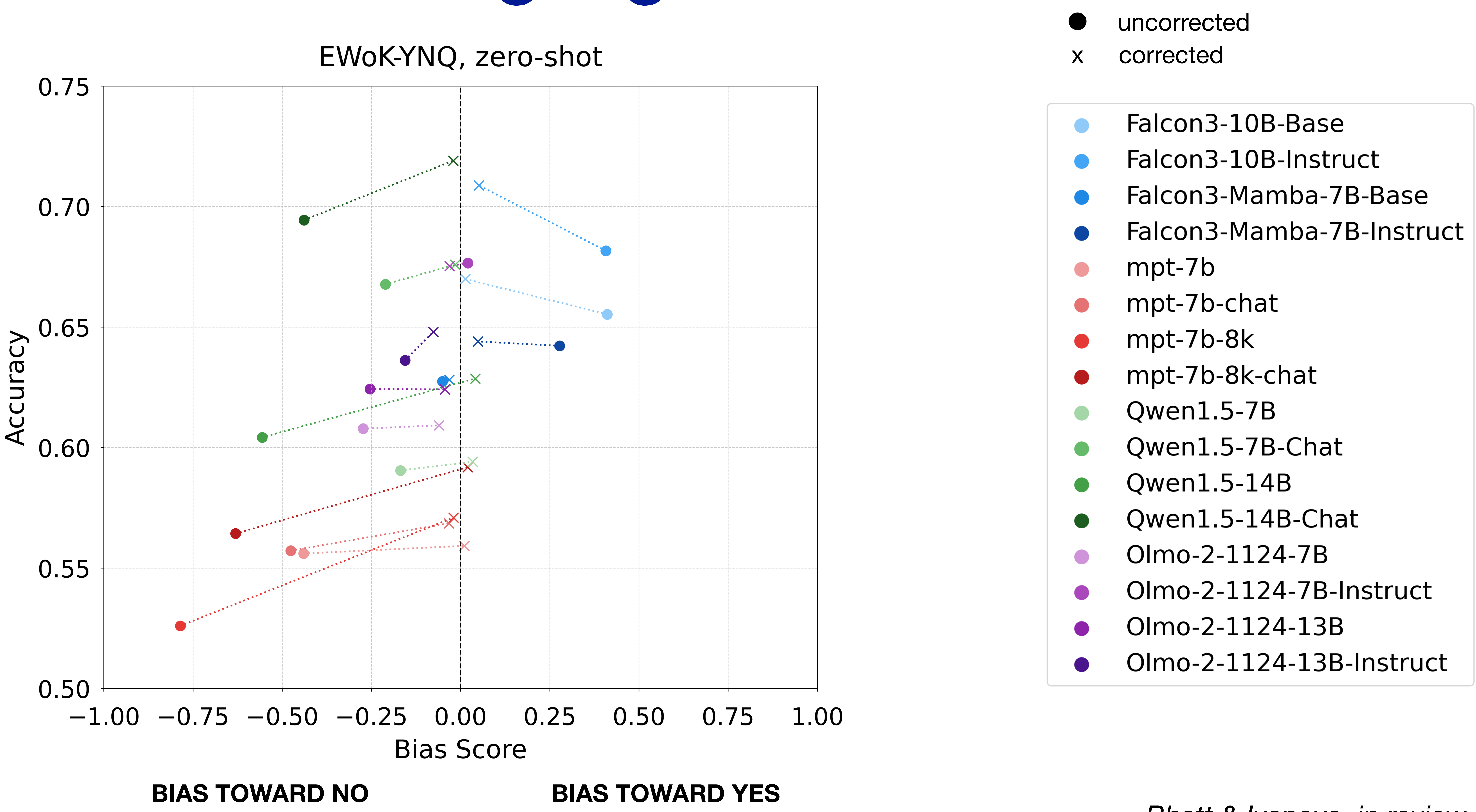
- Humans tend to exhibit a yes-bias (acquiescence bias)

Does yes-bias arise in LMs as a result of statistical learning on language inputs and/or instruction tuning?

Does the yes-bias mask existing model knowledge,
such that
correcting for this bias will improve model performance?

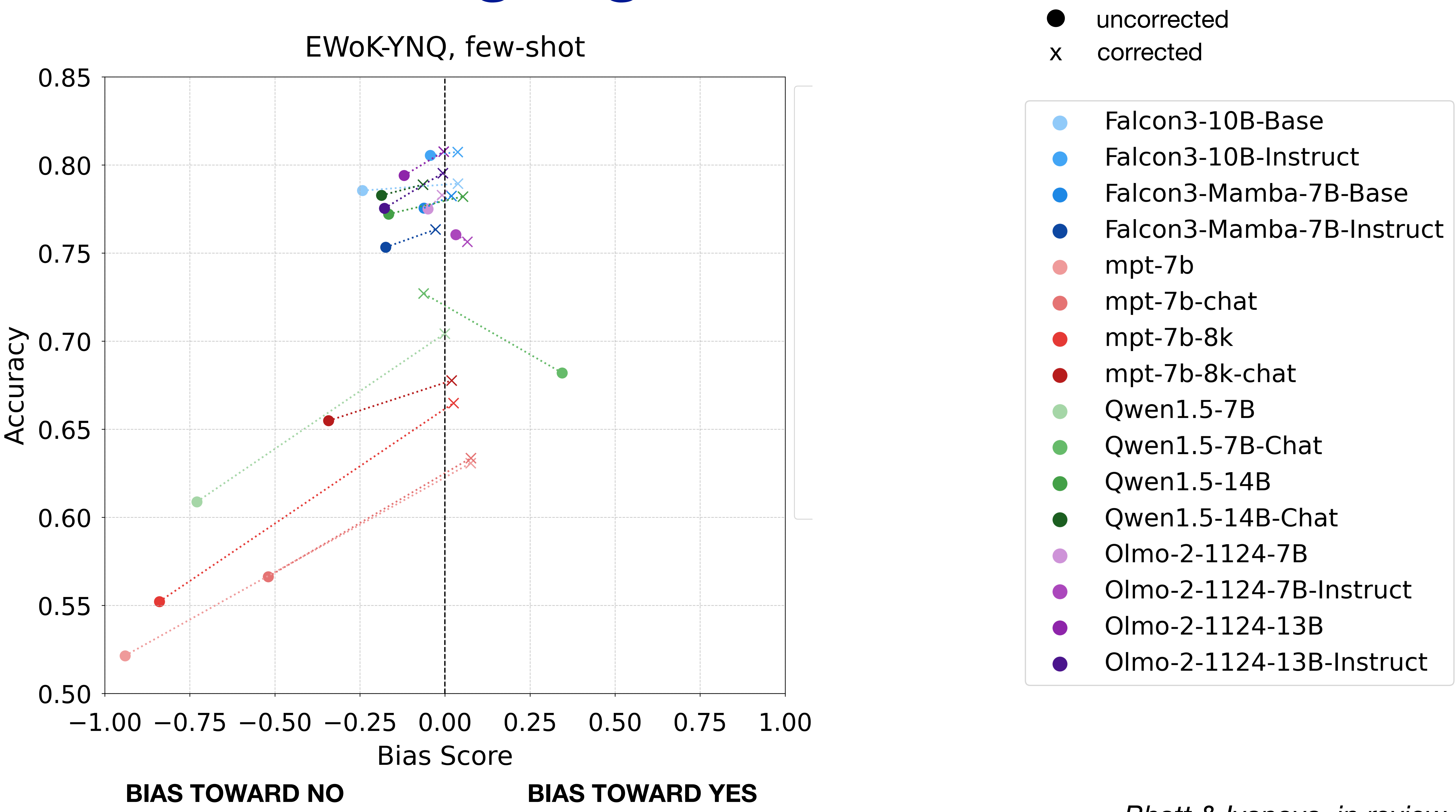
Bhatt & Ivanova, in review

Yes-no bias in language models



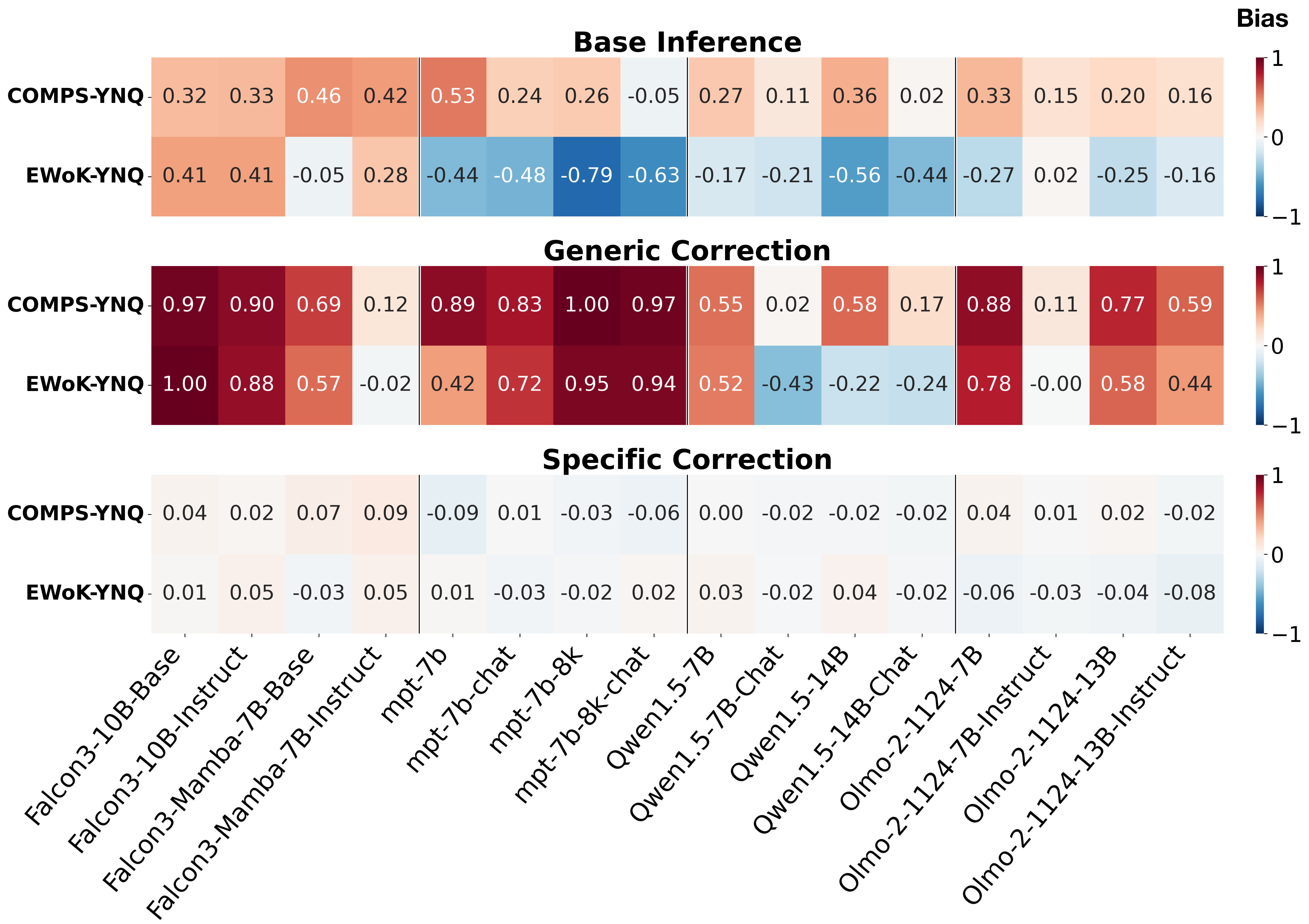
Bhatt & Ivanova, in review

Yes-no bias in language models



Bhatt & Ivanova, in review

Yes-no bias in language models



Generic correction

Subtract LogProbs of sequence-initial Yes/No

Specific correction

Subtract average LogProbs of Yes/No for other Qs in a (class-balanced) portion of the same dataset

Bhatt & Ivanova, in review

Yes-no bias in language models

Does yes-bias in LMs arise as a result of statistical learning on language inputs and/or instruction tuning?

No: language models are often biased, but the bias direction varies depending on the model and testing conditions.

Does the yes-bias mask existing model knowledge, such that correcting for this bias will improve model performance?

Yes: correcting for the bias typically improves model performance.

(our bias correction method works at the level of LogProbs: need open models)

Bhatt & Ivanova, in review

Roadmap

Formal vs functional linguistic competence

It gets complicated: generalized world knowledge

Moving forward

conceptual insights

formal and functional competence

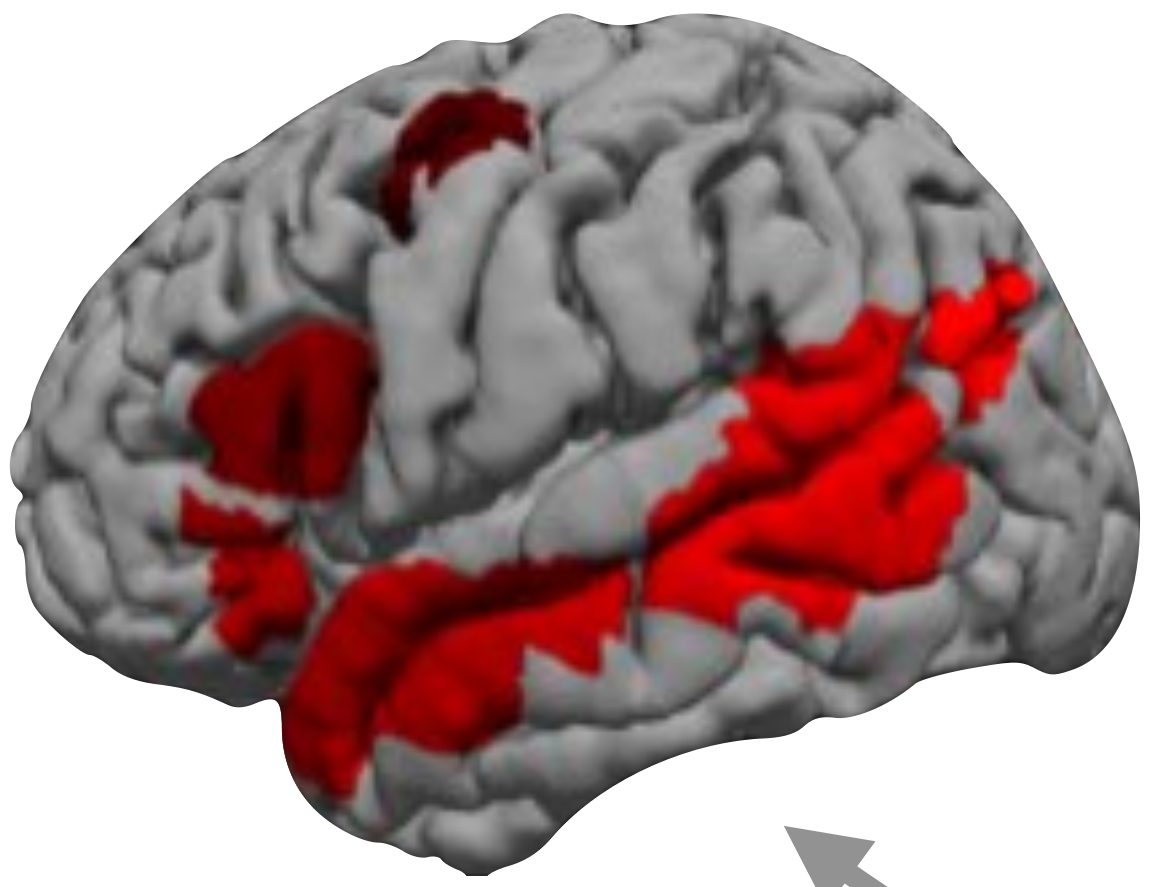
elements of world knowledge (EWoK) benchmark

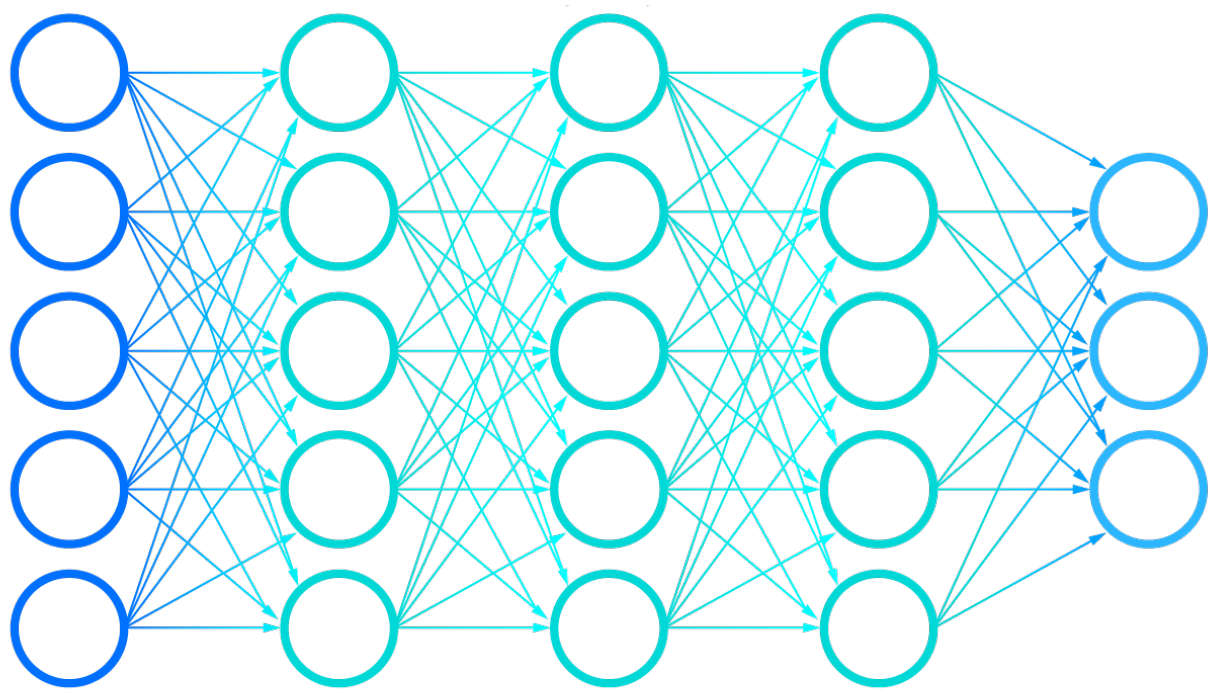
methodological insights

the do’s and the don’ts of experimental design

(Ivanova, 2025, Nat Hum Behav)

from brain to AI





from AI to brain

LLMs as model organisms

generalized event knowledge

yes/no bias

LLMs as computational tools

encoding models of the brain

(ongoing work)

Thanks to…

Kyle Mahowald

Carina Kauf

Aalok Sathe

Ben Lipkin

Om Bhatt

Ev Fedorenko

Jacob Andreas

all other co-authors
my lab members
and all who provided feedback

Thank you for listening!

DO

- determine what the model might have learned about your test during training
- consider alternative strategies a model might use to arrive at the correct answer
- incorporate careful control conditions
- evaluate models under conditions similar to humans
- examine the effect of culture-specific aspects of the prompt on model behavior (y'all / you)
- compare model and human performance (85% / 87%)
- be explicit about the experimental settings when reporting the results
- check whether a model generalizes beyond a single test

DON'T

- rely on minor changes in the test items (This is Sally / This is Megan)
- overly trust crowd-sourced / automatically generated items (My nam is Mark)
- overly trust LLM item scoring
- assume that model responses reflect "universal" human behavior
- assume that models solve the task in the same way as humans
- jump to conclusions (A+ / F-)

Ivanova, 2025, Nat Hum Behav