COGNITION IN LLMS

KARIN DE LANGIS

OUTLINE

Today:

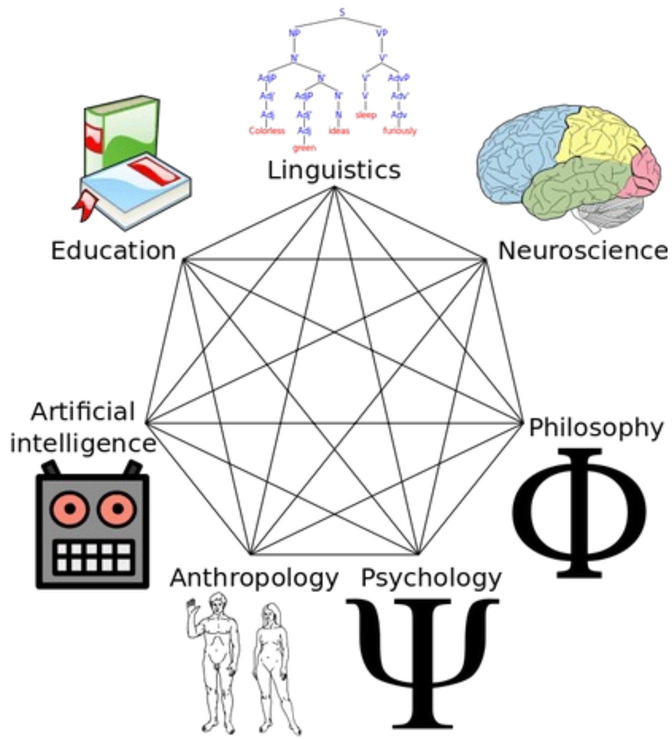
What is **cognition**?

What is the evidence for **LLM cognition?**

- Preliminaries: The Cognitive Sciences
- [Part 1] Cognitive processes in humans and LLMs
- [Part 2] Comprehension in LLMs
- [Part 3] Supporting high-level cognition in human workers

THE COGNITIVE SCIENCES

- Cognitive revolution (1950s)
- Highly interdisciplinary
- Goal: understand the mind



The irony that it took a **machine** to arouse psychologists to an active interest in mental processes has been frequently noted.

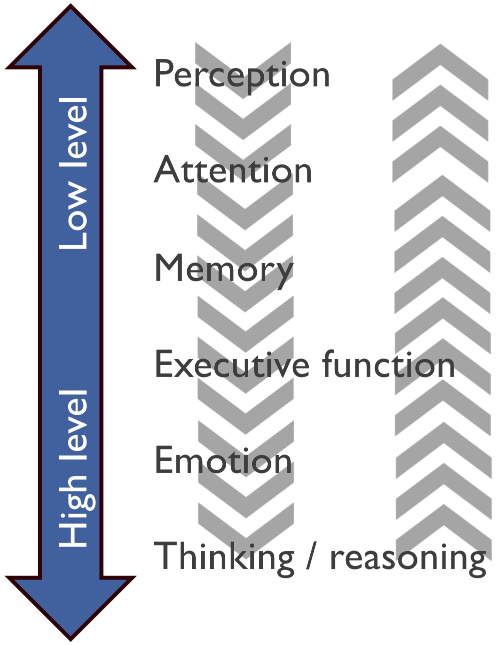
George Miller

WHAT IS COGNITION?

Cognitive Processes

Cognition refers to the processes by which sensory input is transformed, reduced, elaborated, stored, recovered, and used.

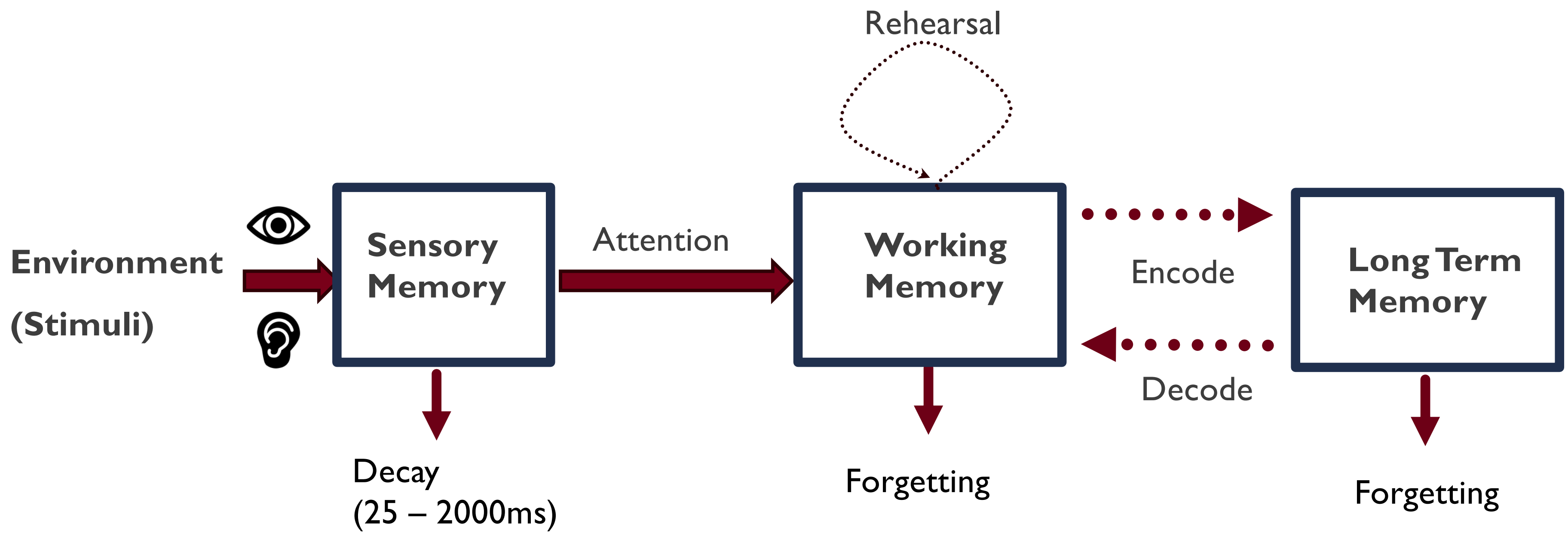
-- *Cognitive Psychology,* Ulric Niesser (1967)



No conscious experience

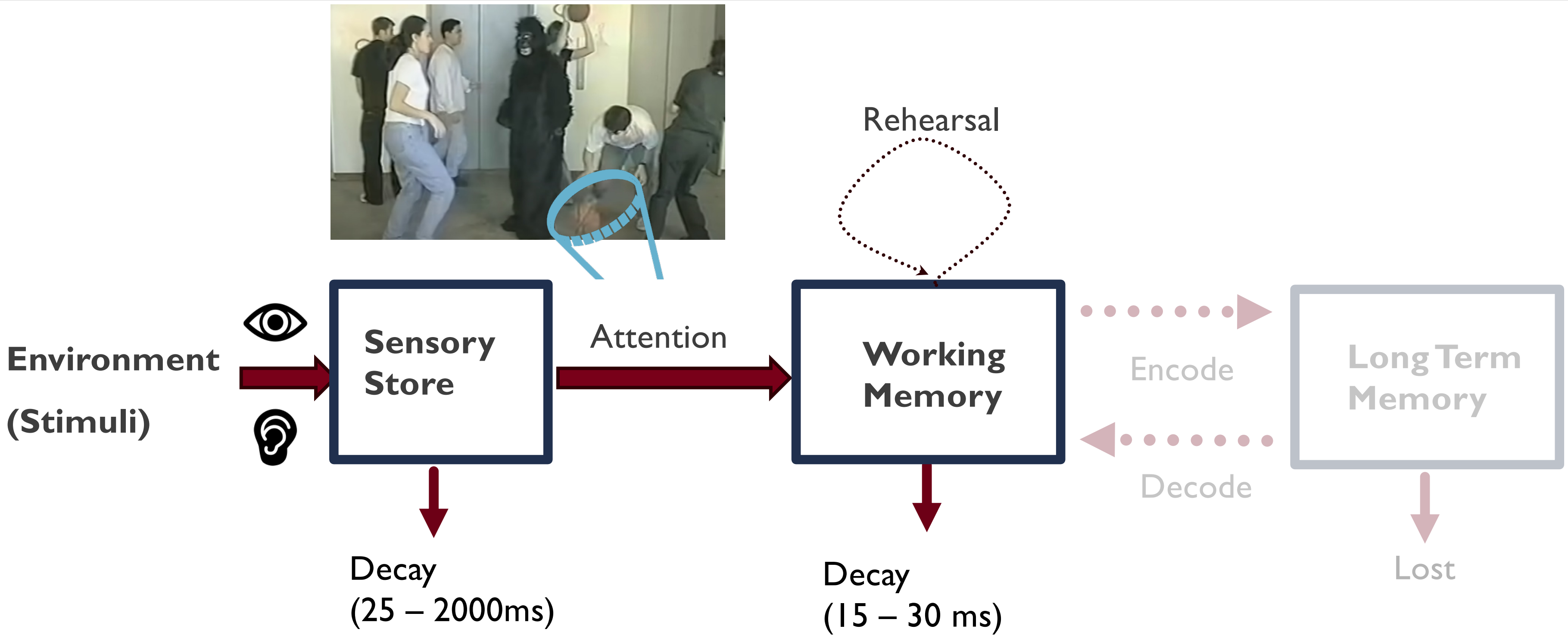
Some conscious experience

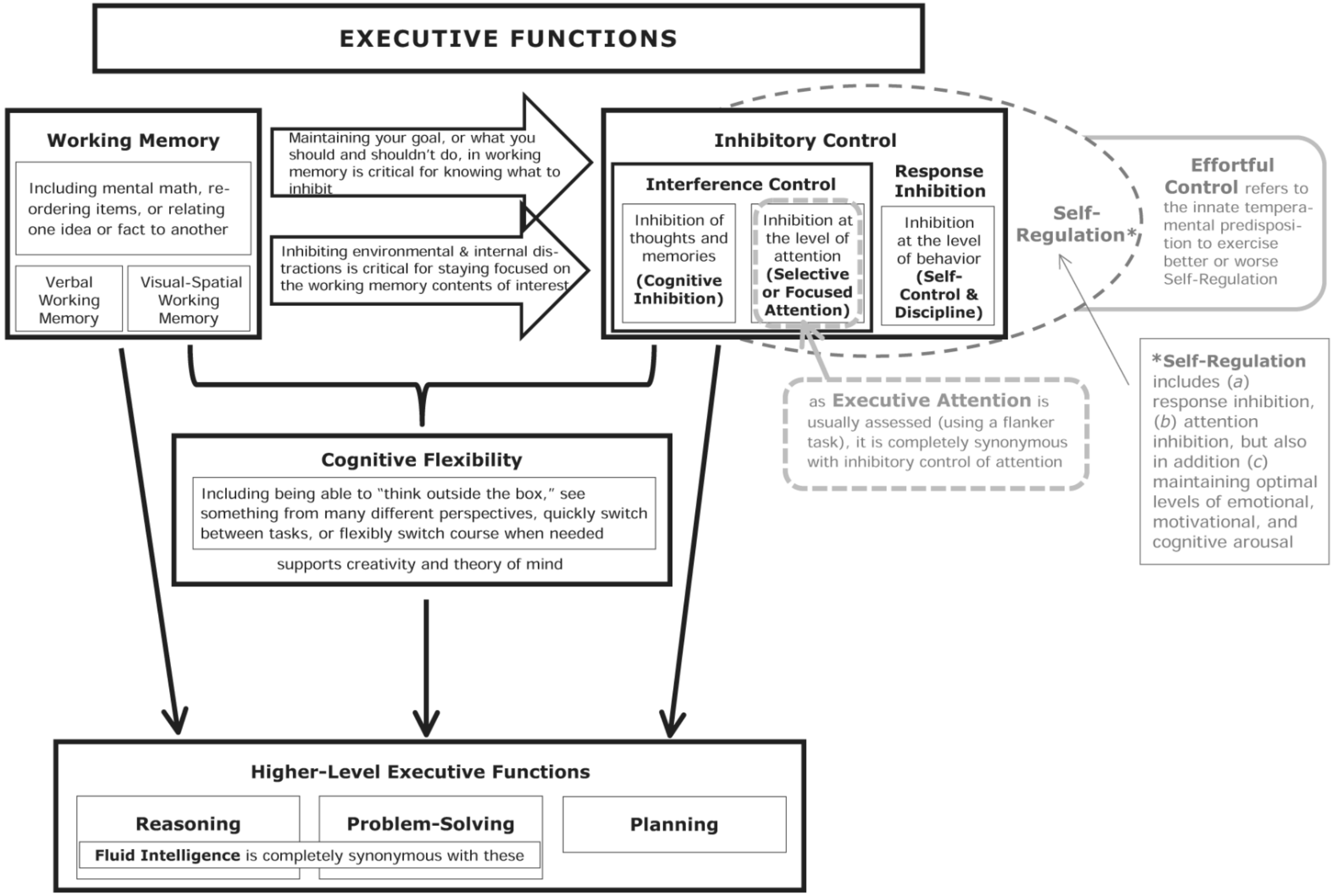
COGNITION AS INFORMATION PROCESSING: MEMORY



QUICK EXAMPLE

COGNITION AS INFORMATION PROCESSING





Diamond 2001

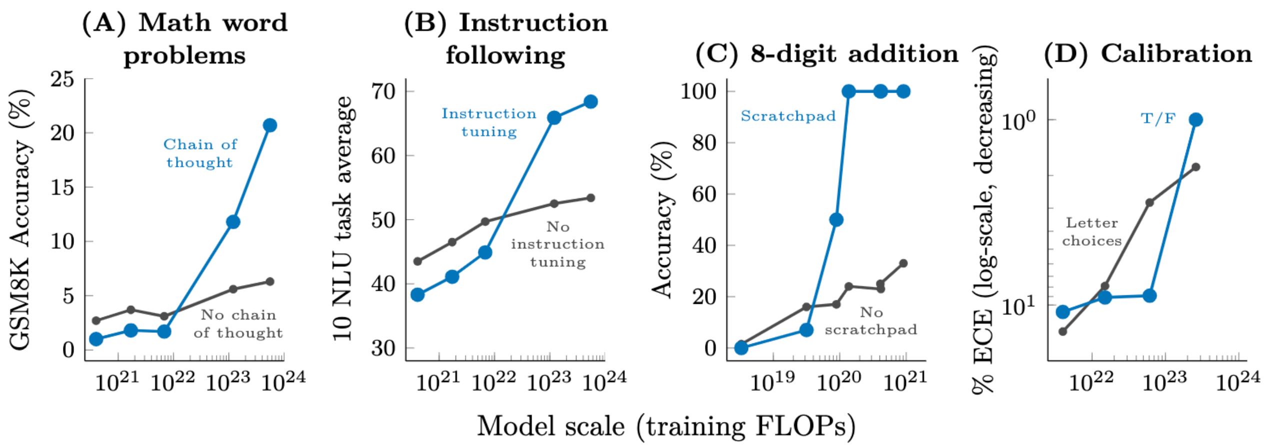
SUMMARY

1. Cognition is supported by complex, multi-faceted cognitive processes
2. Cognitive processes ←→ information processing

   -- Example: sensory store into working memory, attentional filtering
3. Low-level processes (attention, working memory) support high-level processes (reasoning, planning)

HOW DO LANGUAGE MODELS COMPARE?

EMERGENT COGNITION IN LANGUAGE MODELS?



Wei et al., 2022

Language models learn to complete **non-linguistic tasks** as a side effect of training.

COGNITIVE ABILITIES IN LANGUAGE MODELS

But ability to perform language doesn't entail ability to reason.

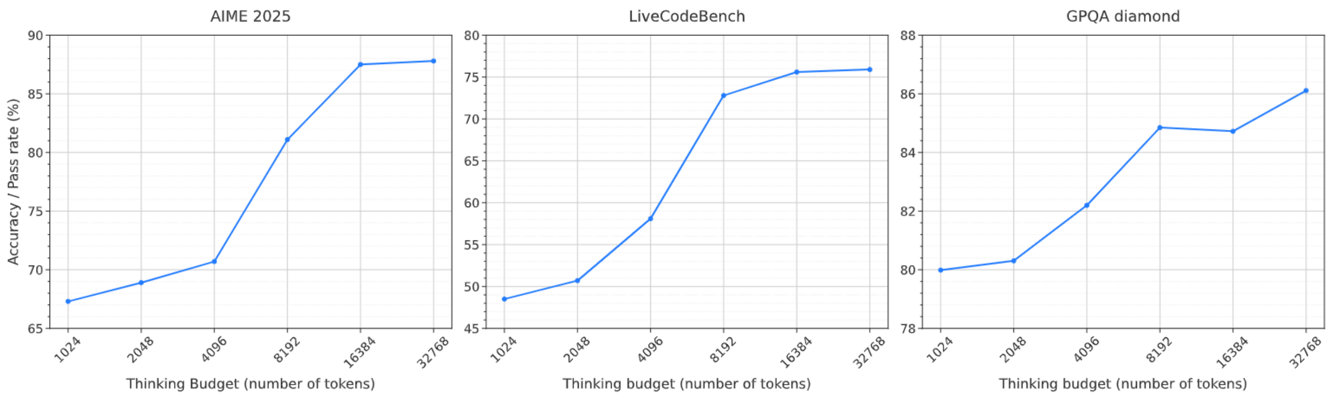
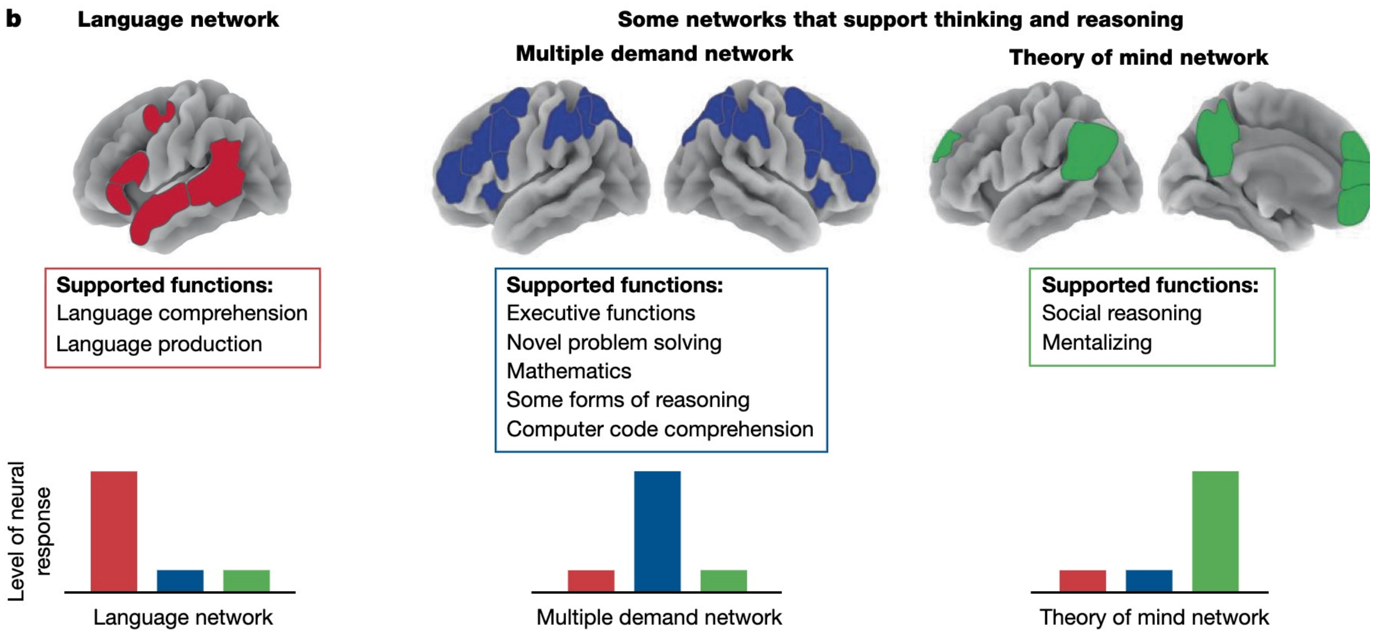


Figure 4 | Impact of thinking budget on performance on AIME 2025 (Balunović et al., 2025), Live-CodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

Gemini Team, 2025

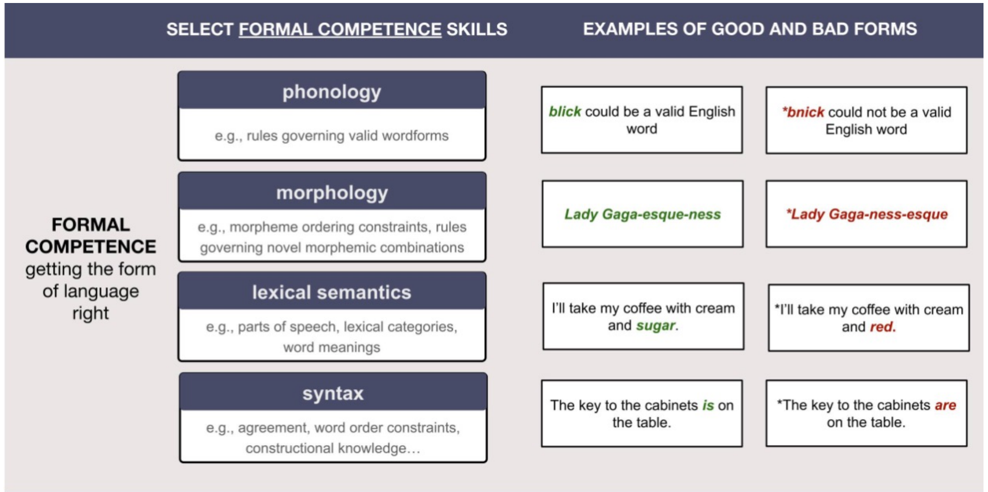
Generating "think-aloud" strings can further improve performance.

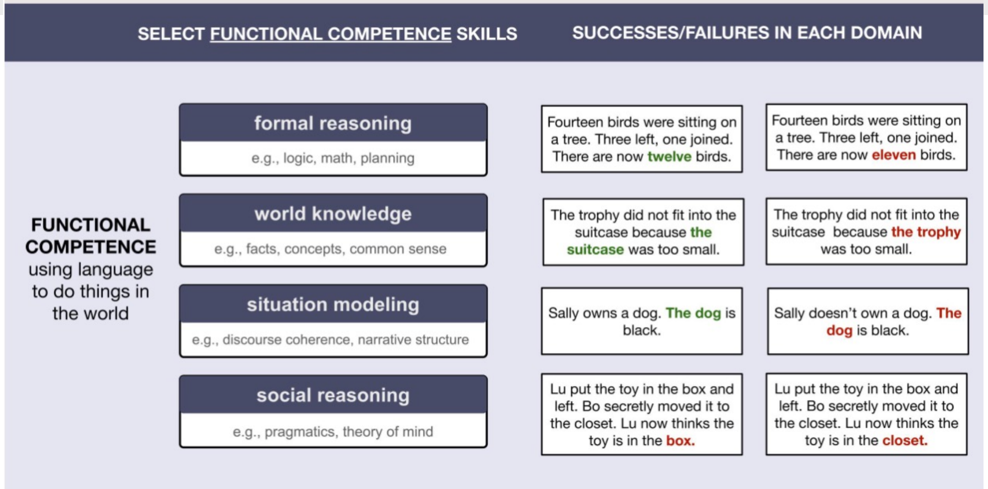
LINGUISTIC ABILITY DOES NOT ENTAIL THINKING ABILITY



Fedorenko, Piantadosi, and Gibson (2023)

FORMAL VS FUNCTIONAL LINGUISTIC COMPETENCE

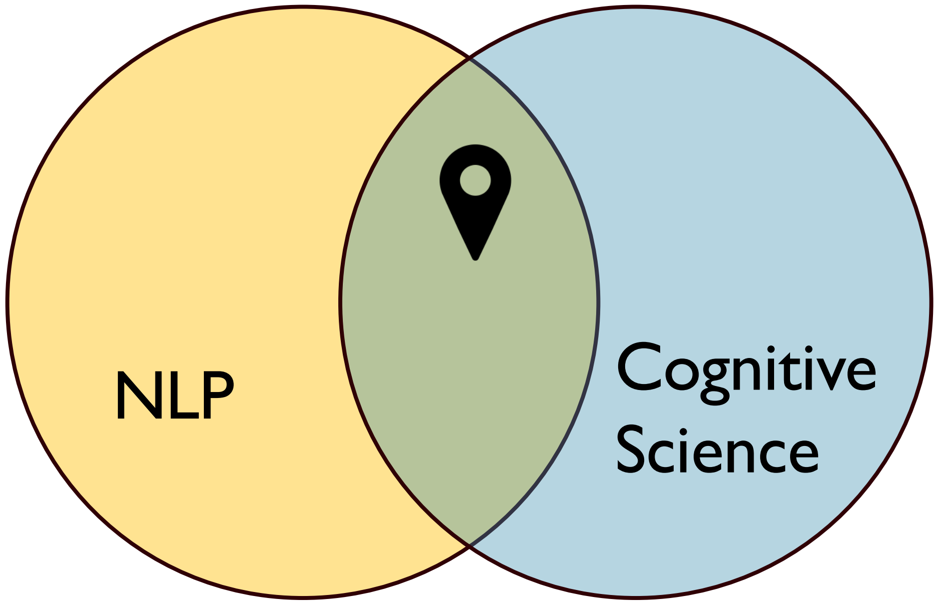




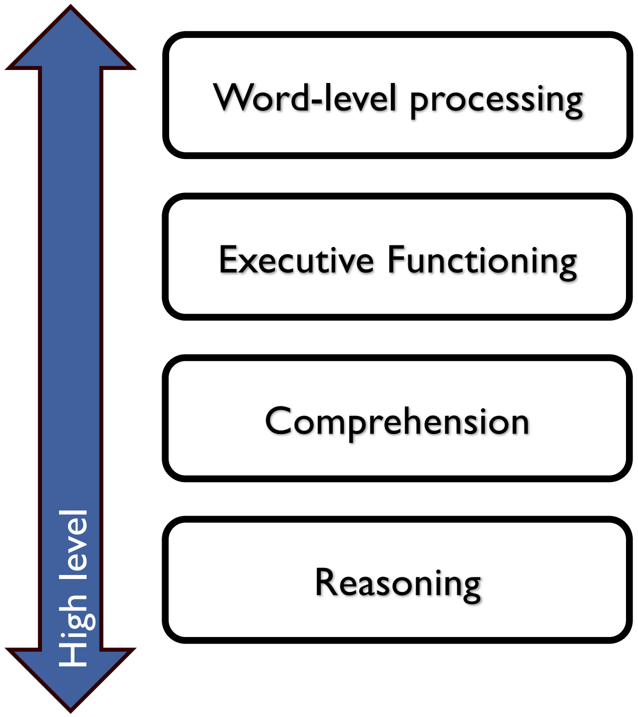
Mahowald et al., 2023

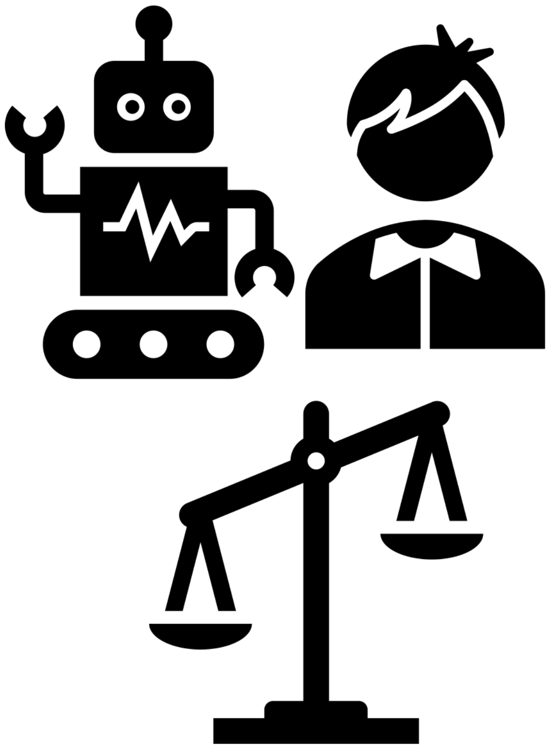
BROADER IMPACTS

- Develop more cognitively-aligned language models
  - These are more helpful for linguists, neuroscientists, cognitive psychologists
- Complementary explanations of LLM behavior…
  - at a higher level of abstraction than neural architecture
  - with frameworks that have predictive power

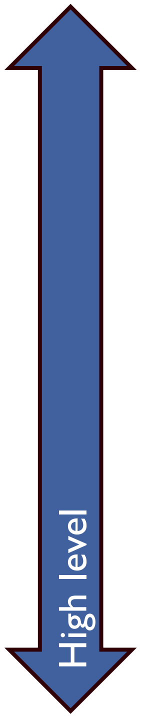


Cognitive Processes in Humans and LLMs





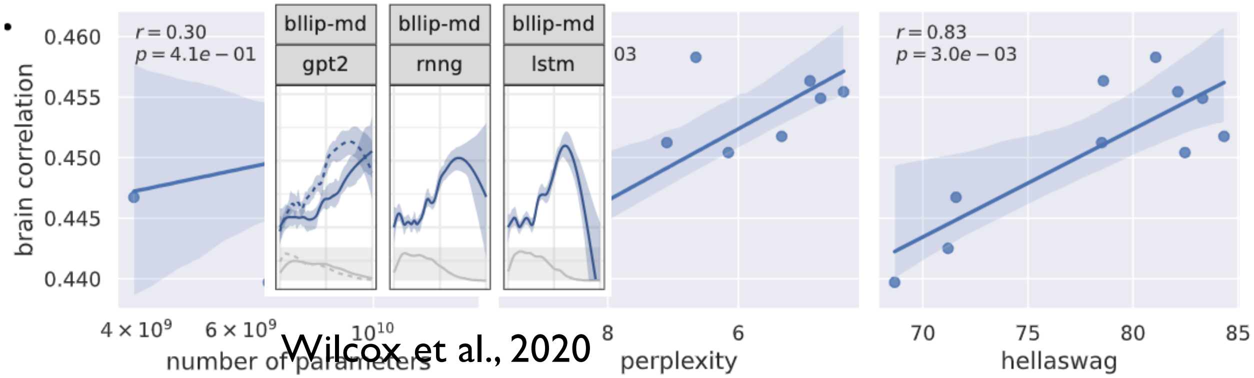
LLMS MATCH WORD-LEVEL PROCESSING



- Word-level processing
- Executive Functioning
- Comprehension
- Reasoning

Larger, better performing language models match brains more closely

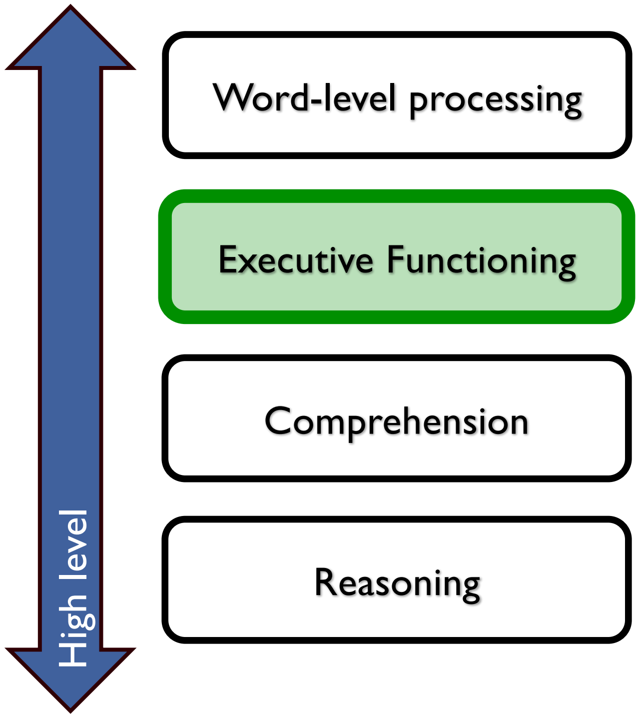
Transformers have high predictive power for eye-tracking data



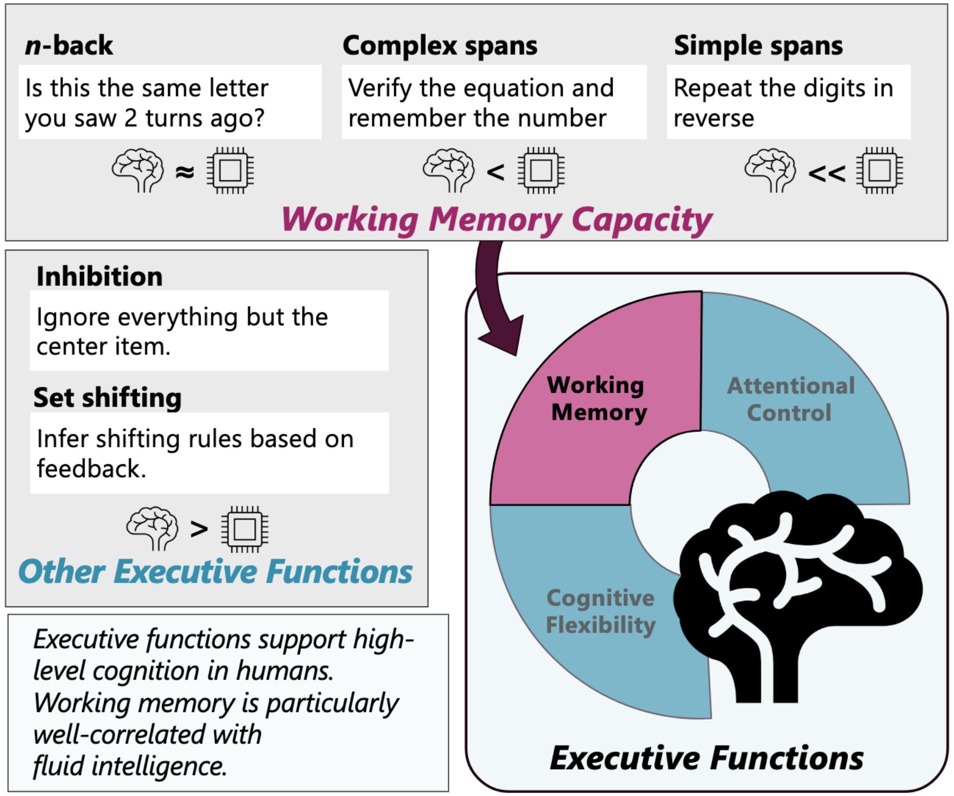
Wilcox et al., 2020

Bonnasse-Gahot and Pallier (2024)

WORKING MEMORY AND EXECUTIVE FUNCTION



WORKING MEMORY AND EXECUTIVE FUNCTION



de Langis et al (2025)

MODELS (USUALLY) HAVE HIGHER WMC

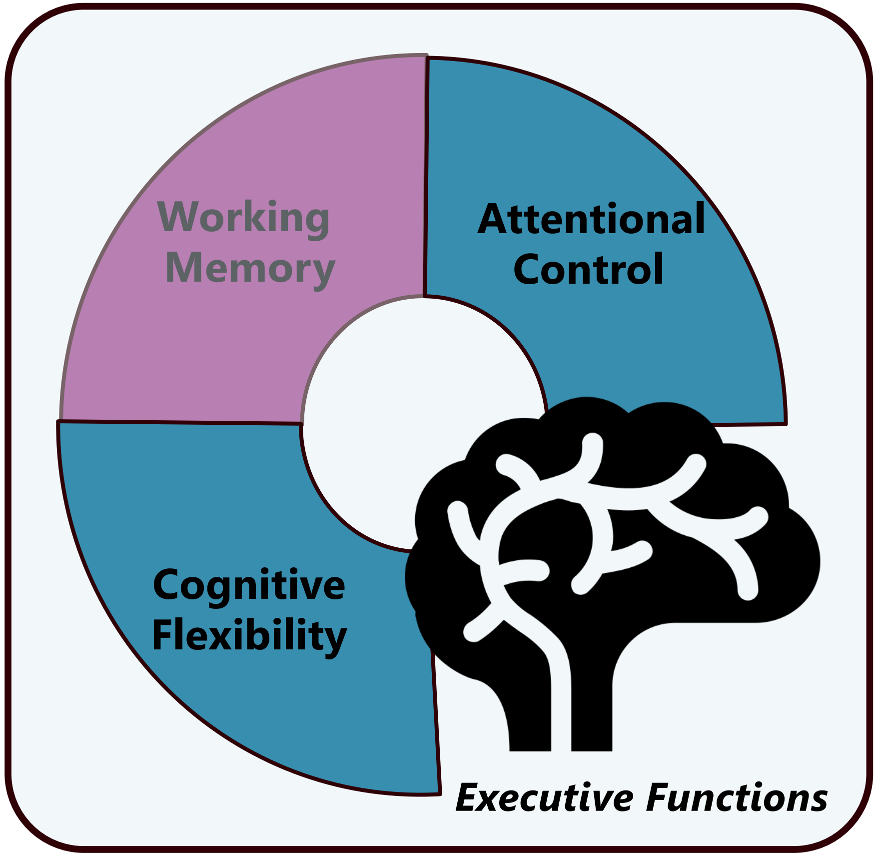
Model	**1-back**	**2-back**	**3-back**	**O-SPAN**	**R-SPAN**	**BDS** ($d = 15$)	**FDS** ($d = 50$)
Gemma-2-9B	0.99	0.75	0.72	0.93	0.97	0.21	0.99
Gemma-2-27B	0.91	0.72	0.69	0.92	0.98	0.59	1.00
Llama-3.1-8B	0.76	0.68	0.67	0.99	0.92	0.18	1.00
Llama-3.1-70B	0.93	0.82	0.82	0.99	0.94	0.83	1.00
Qwen2-7B	0.99	0.89	0.85	0.96	0.66	0.00	1.00
Qwen2-72B	0.78	0.74	0.70	0.93	0.97	0.51	1.00
Human (approx)	0.98	0.91	0.75	0.53	0.48	0.00	0.00

Table 1: Average model accuracy on each test. Rough estimates of typical human scores are provided for reference (see §A.4). Note that for the digit span tasks, we include accuracies for very long strings ($d = 15$ and $d = 50$), while the typical human span is 5 (BDS) to 7 (FDS).

HOW DO MODELS UTILIZE THEIR WORKING MEMORY?

We only care about working memory capacity because it allows us to be smarter!

Does superhuman working memory mean superhuman executive function?



ATTENTION CONTROL

Rules:

- **Only pay attention to the center letter**
- If it is "C" or "X" raise your right hand
- If it is "V" or "B" raise your left hand

ATTENTION CONTROL

X X X X X

ATTENTION CONTROL

C C X C C

ATTENTION CONTROL

V V B V V

ATTENTION CONTROL

X X C X X

ATTENTION CONTROL

C C B C C

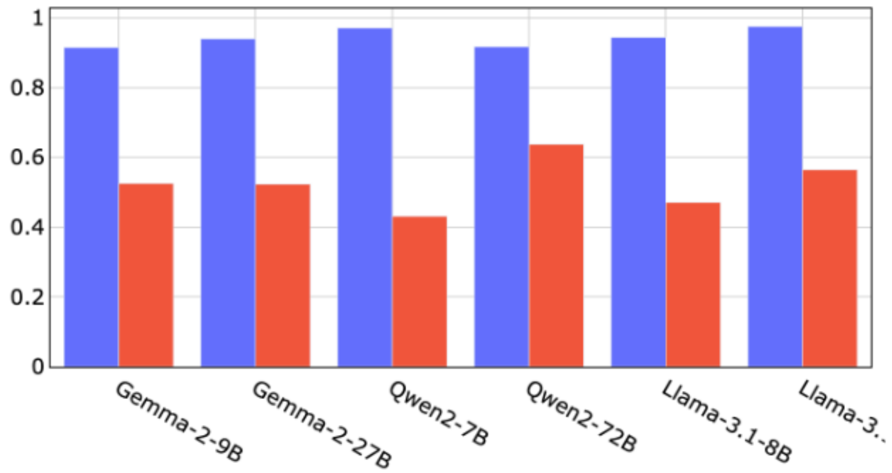
FLANKER TASK

People have ~100% accuracy

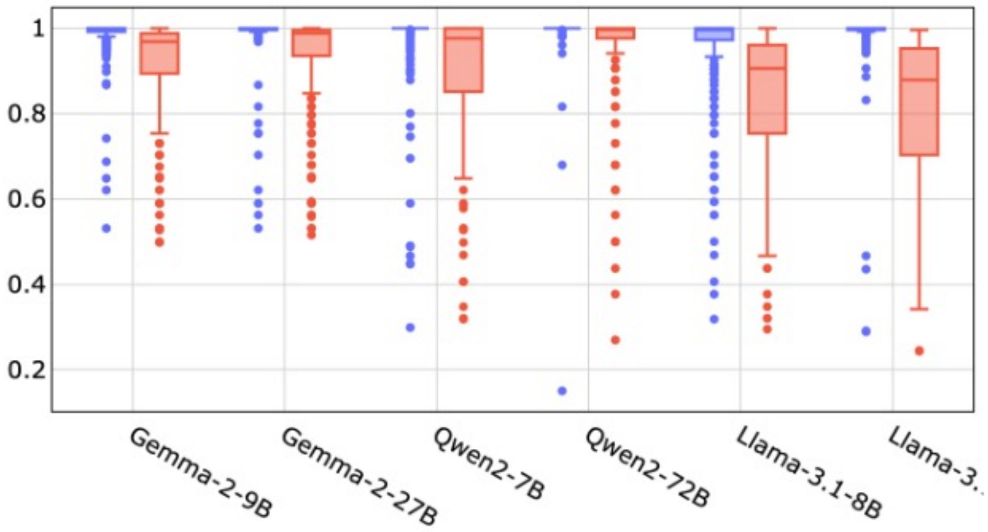
but when the flanker letters are "right" and the center letter is "left"

people require *cognitive processes* (inhibition) to ignore, slowing reaction time (300ms → 500ms)

LLMS ON FLANKER TASK



LLMS ON FLANKER TASK

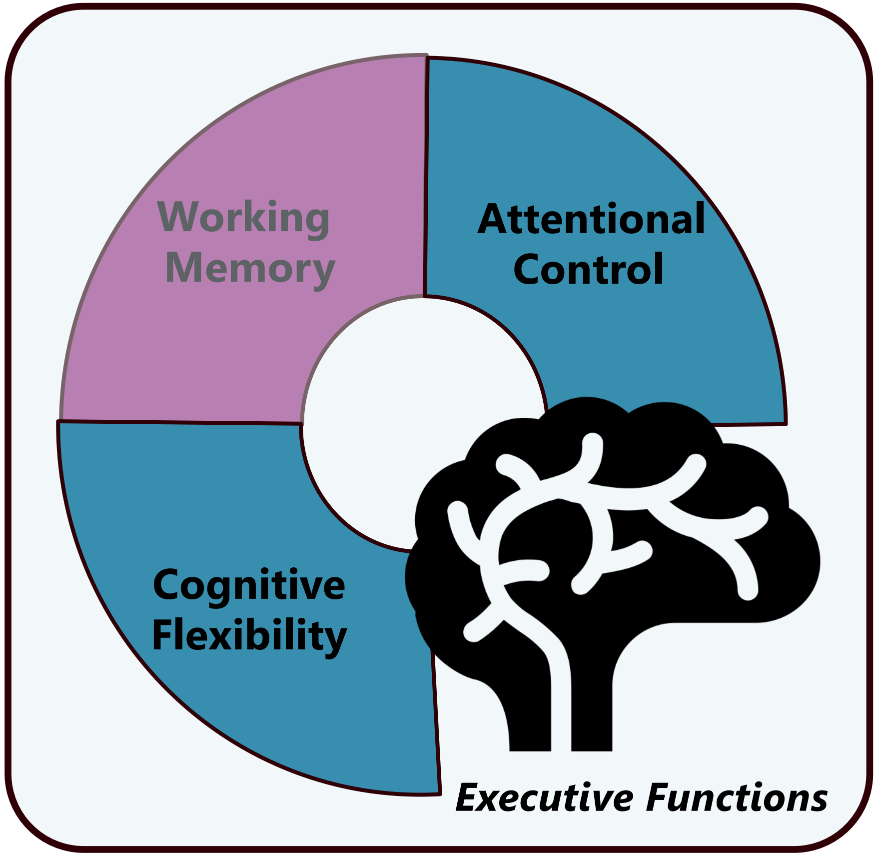


REASONING AUGMENTS

Model	**Congruent**	**Incongruent**
Llama-3.1-8B	0.944	0.471
+Reasoning	0.977	**0.910**
Qwen2-32B	0.917	0.638
+Reasoning	0.994	**0.985**

Table 3: Flanker accuracies (simple attentional control) on the congruent vs. incongruent conditions. Adding reasoning allows models to achieve much higher accuracies in this task. Thought strings are several hundred tokens long, despite the straightforward and simple task.

WHAT ABOUT COGNITIVE FLEXIBILITY?



WISCONSIN CARD SORTING TASK

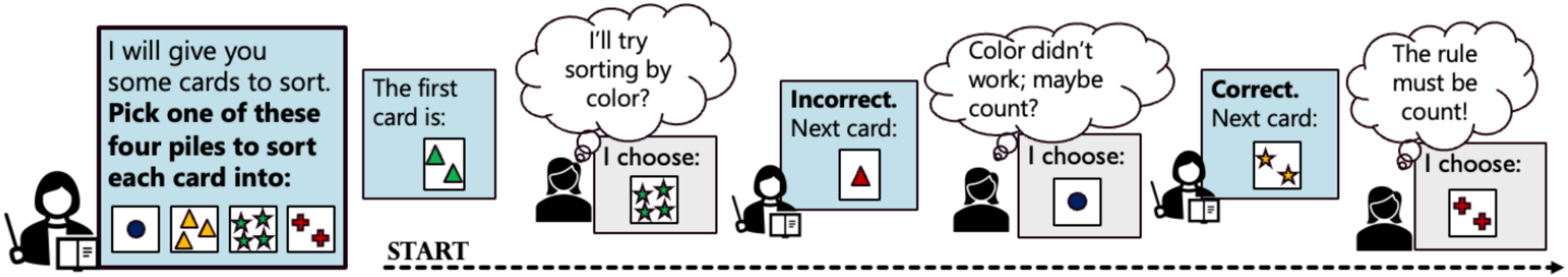
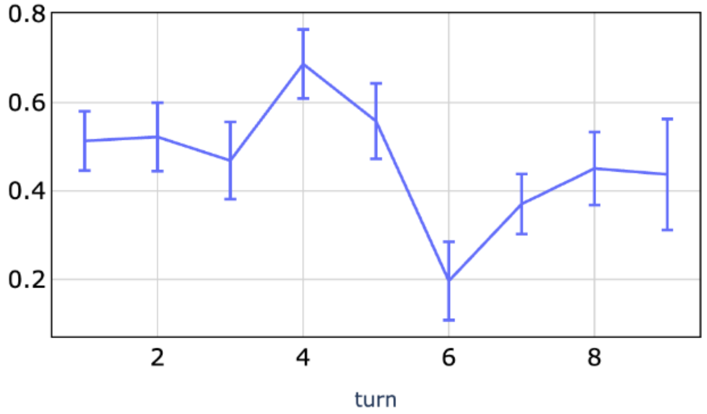


Figure 5: The Wisconsin Card Sorting Task (WCST). Participants are serially presented with cards to sort, and they must infer an underlying "sorting rule" based on feedback. *The rule will periodically change without warning*, and people must detect the change and adjust to the new rule. Participants make sorting decisions as quickly as possible. The WCST tests various executive functions, primarily cognitive flexibility as participants adapt from one set of rules to another.

LLMS DO NOT ESTABLISH AN EFFECTIVE STRATEGY

		Accuracy (↑)	Preservation Error (↓)	Other Error (↓)
Gemma2	9B	0.29	0.21	0.51
	27B	0.49	0.20	0.30
Llama3.1	8B	0.53	0.32	0.15
	70B	0.50	0.26	0.24
Qwen2	7B	0.52	0.33	0.14
	72B	0.51	0.29	0.21
Healthy adults		0.77	0.12	0.09

LLMS DO NOT ESTABLISH AN EFFECTIVE STRATEGY



Llama-3.1-70B-Instruct also fails to maintain the current rule

REASONING DOES NOT HELP

Llama-8B-Instruct:
56% accuracy

R1-Distill-Llama-8B:
23% accuracy

Qwen2.5-7B-Instruct
53% accuracy

Qwen3-8B:
26% accuracy

DIFFICULTY CONVERGING

Wait, maybe the rule is that the item's **color** matches the option's color. …

Alternatively, maybe the rule is based on the **count**. …

But in the first case, the item was two red squares. If the rule is count, …

Wait, maybe the rule is that the item's shape matches the option's **shape**. …

Alternatively, maybe the rule is that the item's count matches the option's **count**. …

Wait, maybe the rule is that the item's color matches the option's **color**. …

…This is getting a bit tangled.

KEY INSIGHTS

Models have **very strong working memory**

But it does not assist in **attentional control** and **cognitive flexibility** as in humans

Reasoning **may help** but it is **not there yet**

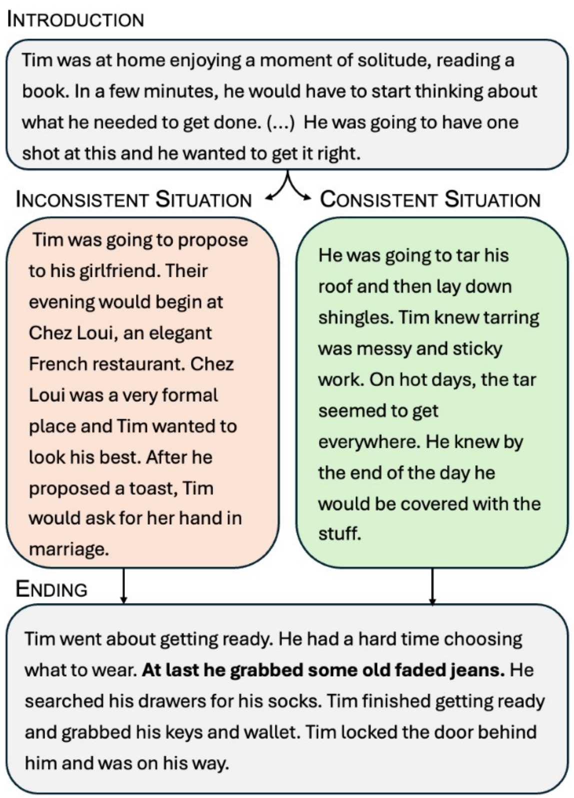
Comprehension in LLMs

STORY COMPREHENSION

Humans build meaning by building a model of the story world

Incoherence: a story becomes *incoherent* if one "world model" cannot explain the whole story.

APPROACH: PAIRED NARRATIVES



de Langis et al (2025)

PROBES REVEAL DISTINCTION ONLY AT EVENT

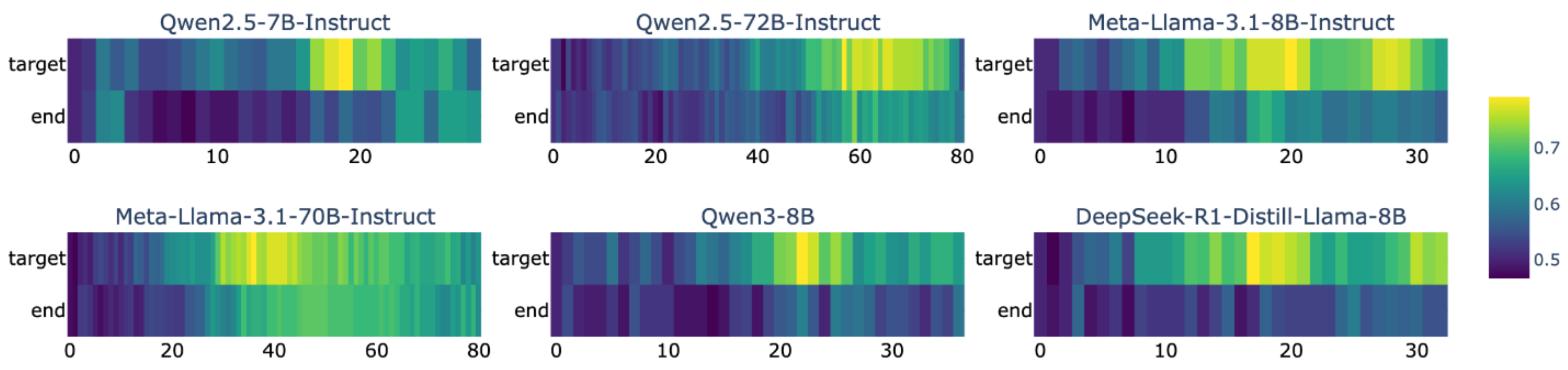


Figure 4: Mean accuracies across 10-fold cross-validation for probing hidden state representations to identify incoherent narratives. We probe both at the end of the target sentence that contains the incoherence in the incoherent story version (target), and at the conclusion of the story (end). The x axis denotes model layer. All models show strong separation at the target location, but by the story's end, separation is notably weaker, with smaller models in particular near chance ($\approx 50 - 60\%$ accuracy). Llama3.1-70B has the best performance at the story's end, and it also demonstrates the best understanding of coherence in responses to rating questions.

RESPONSES SHOW POOR SEPARATION

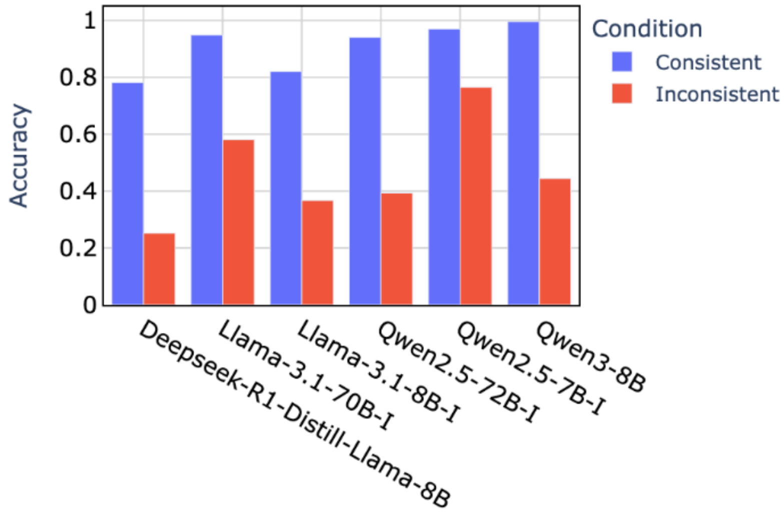


Figure 6: Model accuracies by condition when responding True or False to whether a story is coherent.

TAKEAWAYS

Models do not appear to have a strong mastery of "coherence" in a story

Models may be responding to **unexpectedness** rather than **incoherence**
→ statistical, rather than meaning-based, understanding

COLLABORATORS AND ADVISORS

Dr Dongyeop Kang

Dr Stephen Guy

Dr Andreas Schramm

Dr Andrew Elfenbein

Dr Laura Allen

Puren Oncel

Khanh Chi Le

Jong Inn Park

William Walker

Ryan Koo

Bin Hu

Minnesota NLP

Thank you!!