

COGNITION IN LLMS

KARIN DE LANGIS



OUTLINE

Today:

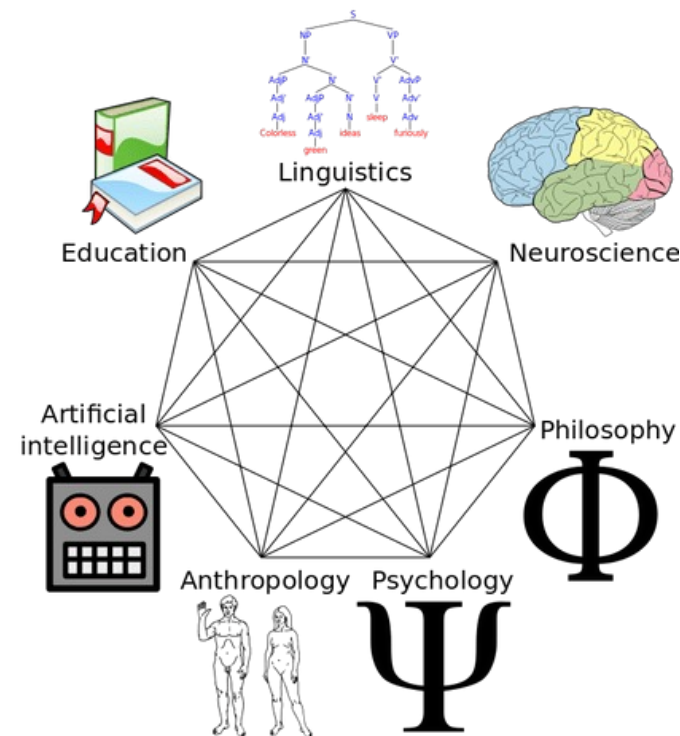
What is **cognition**?

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



Preliminaries

Cognitive Processes

Comprehension

Annotation

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

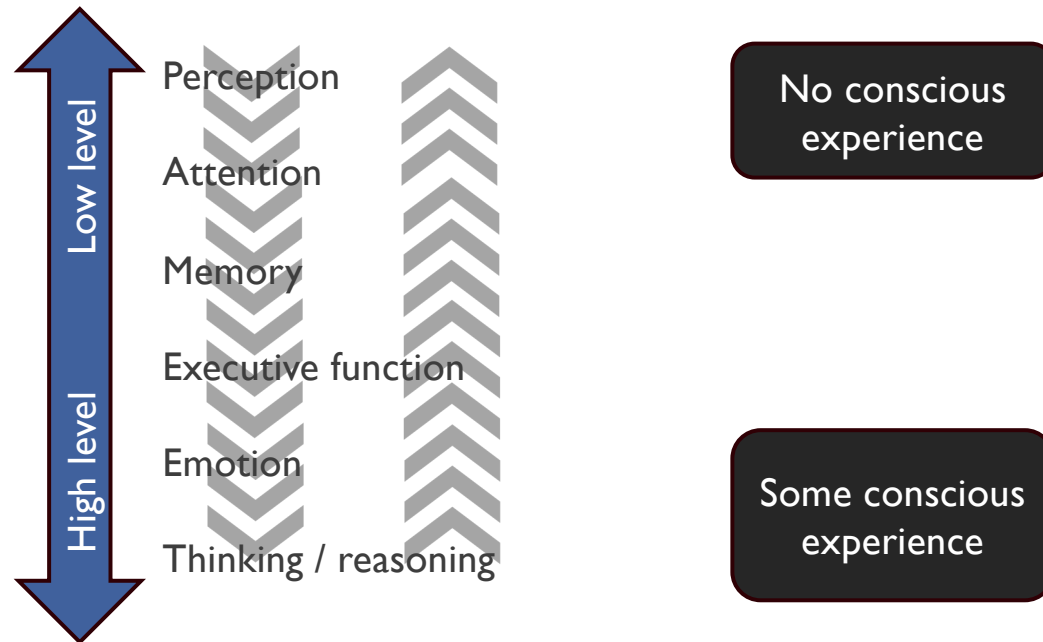
George Miller

Cognitive Processes

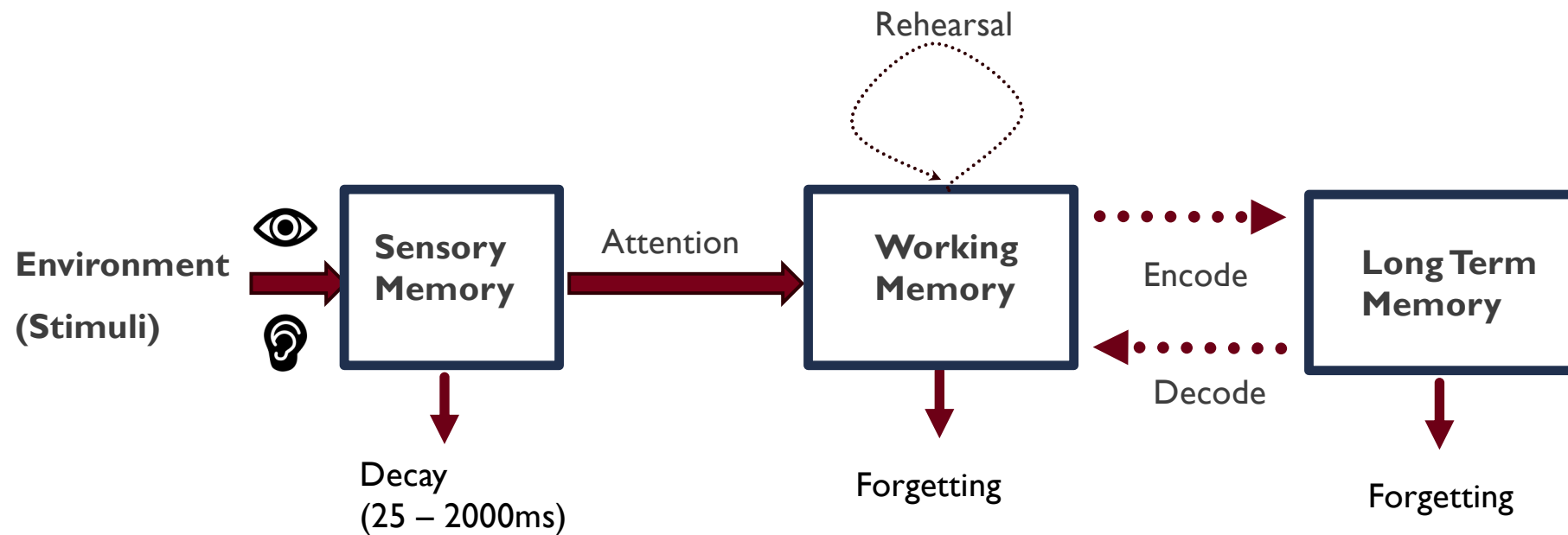
Cognition refers to the processes by which sensory input is transformed, reduced, elaborated, stored, recovered, and used.

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

WHAT IS COGNITION?



COGNITION AS INFORMATION PROCESSING: MEMORY



Preliminaries

Cognitive Processes

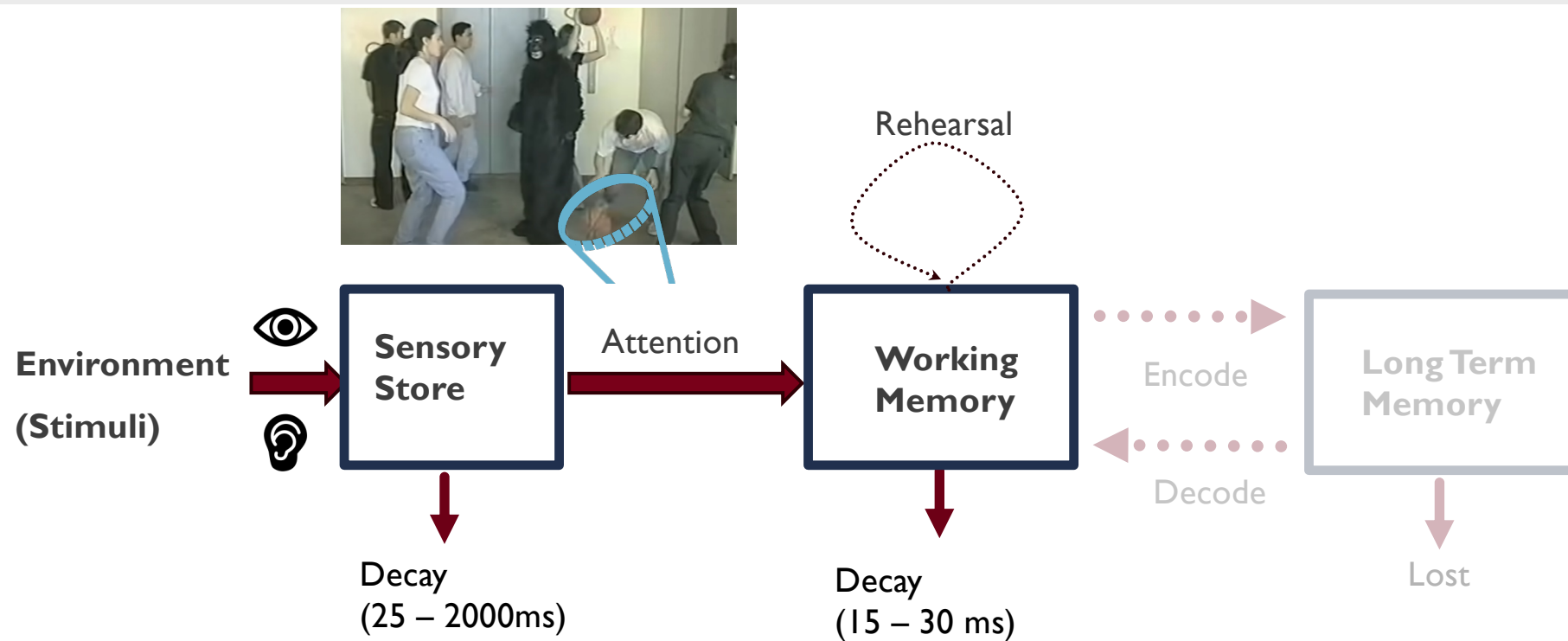
Comprehension

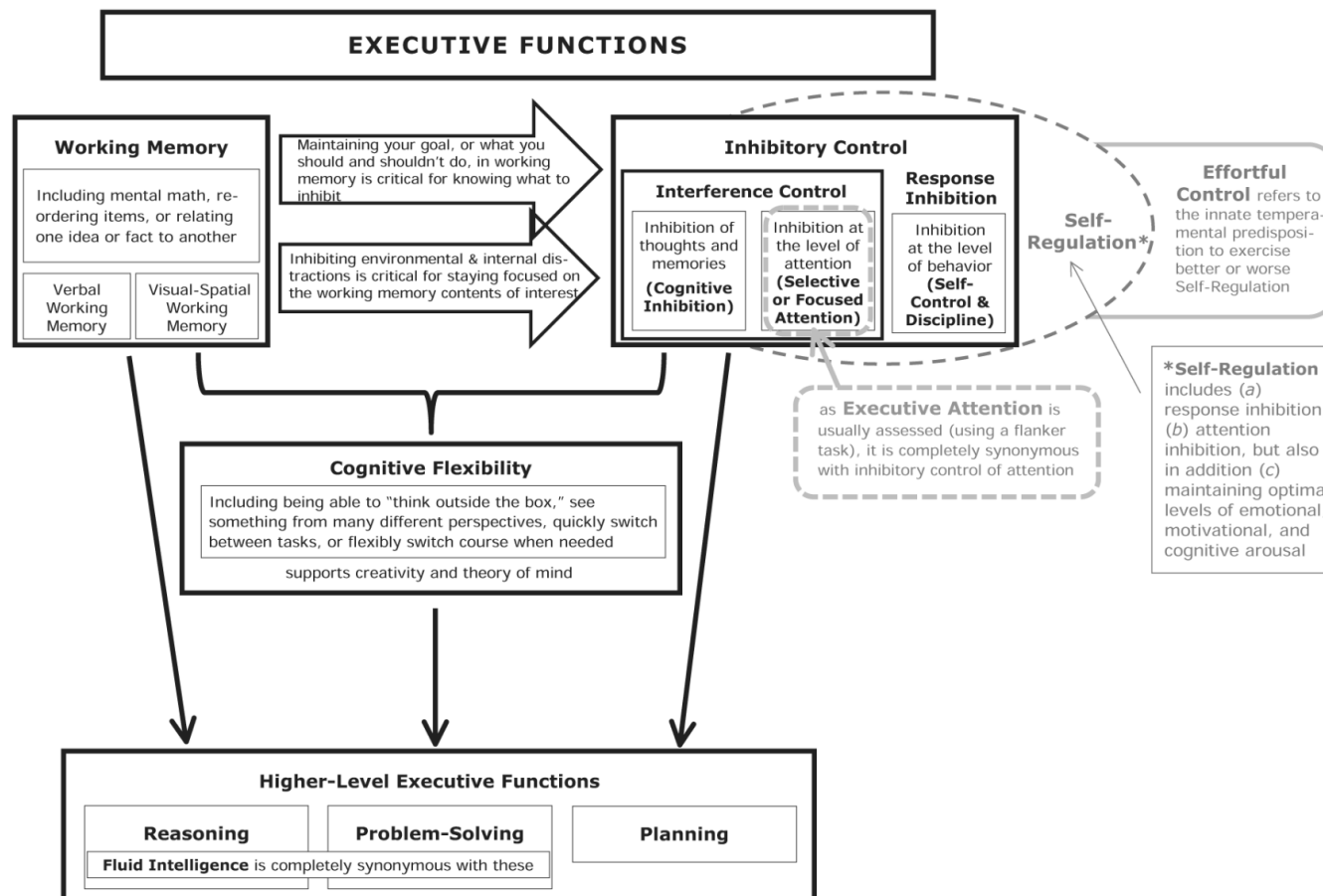
Annotation

QUICK EXAMPLE



COGNITION AS INFORMATION PROCESSING





Diamond 2001

SUMMARY

1. Cognition is supported by complex, multi-faceted cognitive processes
2. Cognitive processes $\leftrightarrow$ information processing
 - Example: sensory store into working memory, attentional filtering
3. Low-level processes (attention, working memory) support high-level processes (reasoning, planning)

Preliminaries

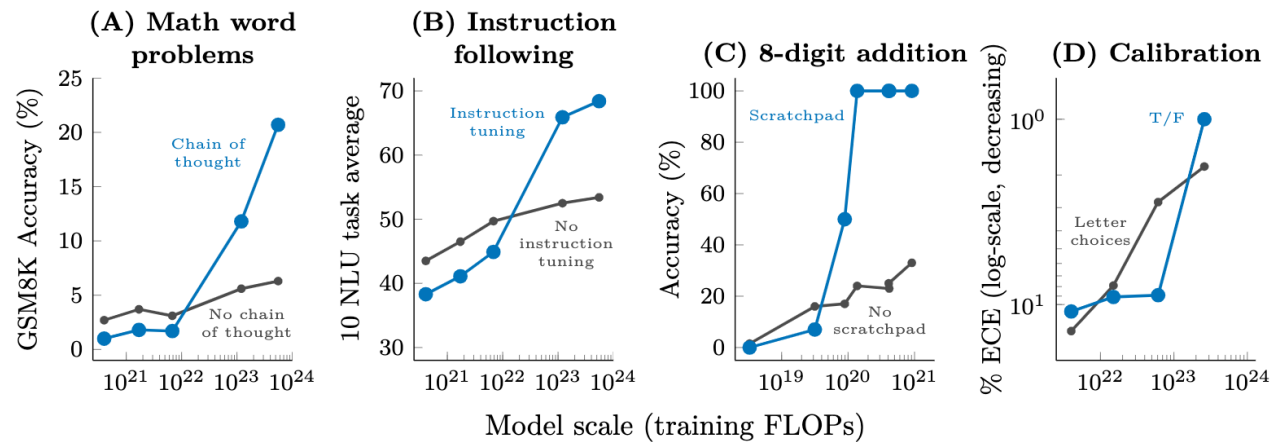
Cognitive Processes

Comprehension

Annotation

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.

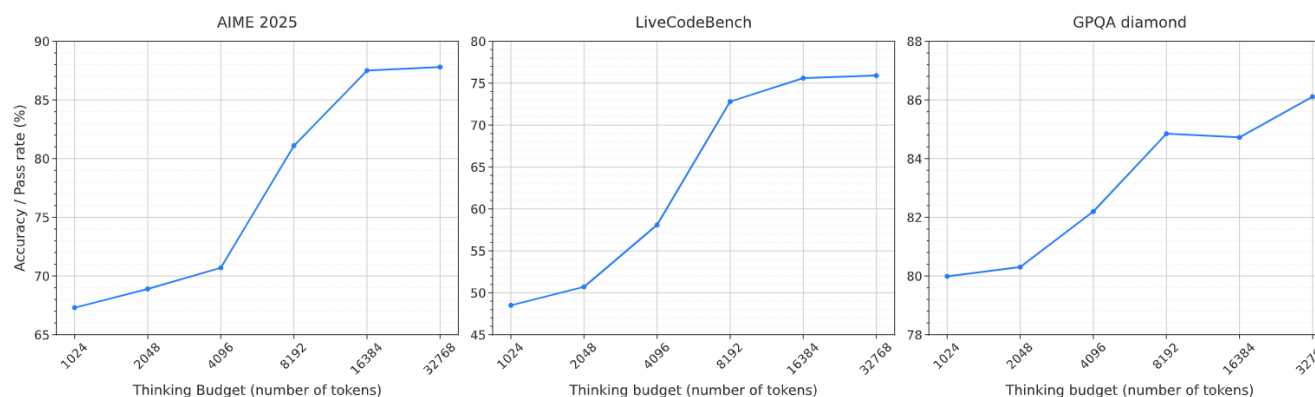


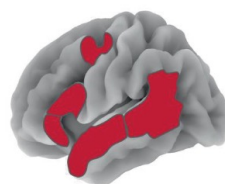
Figure 4 | Impact of thinking budget on performance on AIME 2025 ([Balunović et al., 2025](#)), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) ([Jain et al., 2024](#)) and GPQA diamond ([Rein et al., 2024](#)) benchmarks.

Gemini Team, 2025

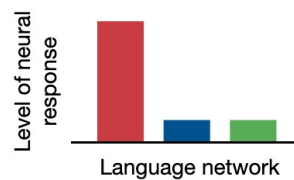
Generating “think-aloud” strings can further improve performance.

LINGUISTIC ABILITY DOES NOT ENTAIL THINKING ABILITY

b Language network

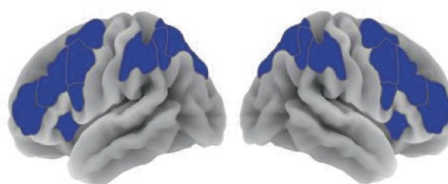


Supported functions:
Language comprehension
Language production

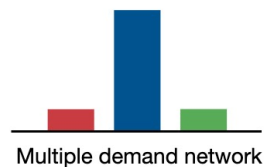


Some networks that support thinking and reasoning

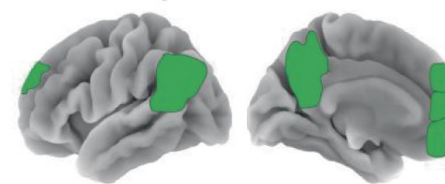
Multiple demand network



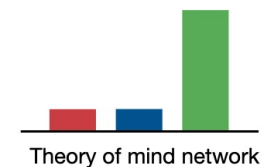
Supported functions:
Executive functions
Novel problem solving
Mathematics
Some forms of reasoning
Computer code comprehension



Theory of mind network



Supported functions:
Social reasoning
Mentalizing



Fedorenko, Piantadosi, and Gibson (2023)

Preliminaries

Cognitive Processes

Comprehension

Annotation

FORMAL VS FUNCTIONAL LINGUISTIC COMPETENCE

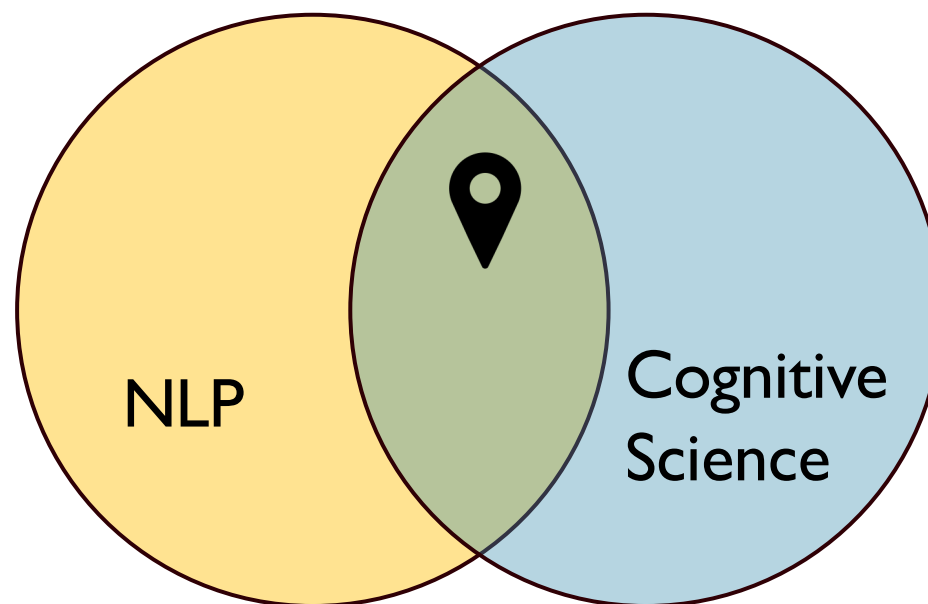
FORMAL COMPETENCE getting the form of language right	SELECT FORMAL COMPETENCE SKILLS	EXAMPLES OF GOOD AND BAD FORMS	
	phonology e.g., rules governing valid wordforms	<i>blick</i> could be a valid English word	<i>*bnick</i> could not be a valid English word
	morphology e.g., morpheme ordering constraints, rules governing novel morphemic combinations	<i>Lady Gaga-esque-ness</i>	<i>*Lady Gaga-ness-esque</i>
	lexical semantics e.g., parts of speech, lexical categories, word meanings	I'll take my coffee with cream and <i>sugar</i> .	*I'll take my coffee with cream and <i>red</i> .
	syntax e.g., agreement, word order constraints, constructional knowledge...	The key to the cabinets <i>is</i> on the table.	*The key to the cabinets <i>are</i> on the table.

FUNCTIONAL COMPETENCE using language to do things in the world	SELECT FUNCTIONAL COMPETENCE SKILLS	SUCCESSSES/FAILURES IN EACH DOMAIN	
	formal reasoning e.g., logic, math, planning	Fourteen birds were sitting on a tree. Three left, one joined. There are now <i>twelve</i> birds.	Fourteen birds were sitting on a tree. Three left, one joined. There are now <i>eleven</i> birds.
	world knowledge e.g., facts, concepts, common sense	The trophy did not fit into the suitcase because <i>the suitcase</i> was too small.	The trophy did not fit into the suitcase because <i>the trophy</i> was too small.
	situation modeling e.g., discourse coherence, narrative structure	Sally owns a dog. <i>The dog</i> is black.	Sally doesn't own a dog. <i>The dog</i> is black.
	social reasoning e.g., pragmatics, theory of mind	Lu put the toy in the box and left. Bo secretly moved it to the closet. Lu now thinks the toy is in the <i>box</i> .	Lu put the toy in the box and left. Bo secretly moved it to the closet. Lu now thinks the toy is in the <i>closet</i> .

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



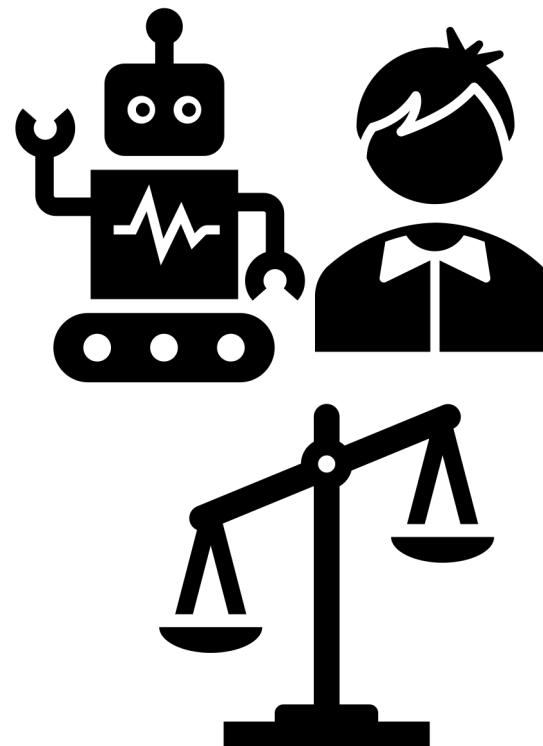
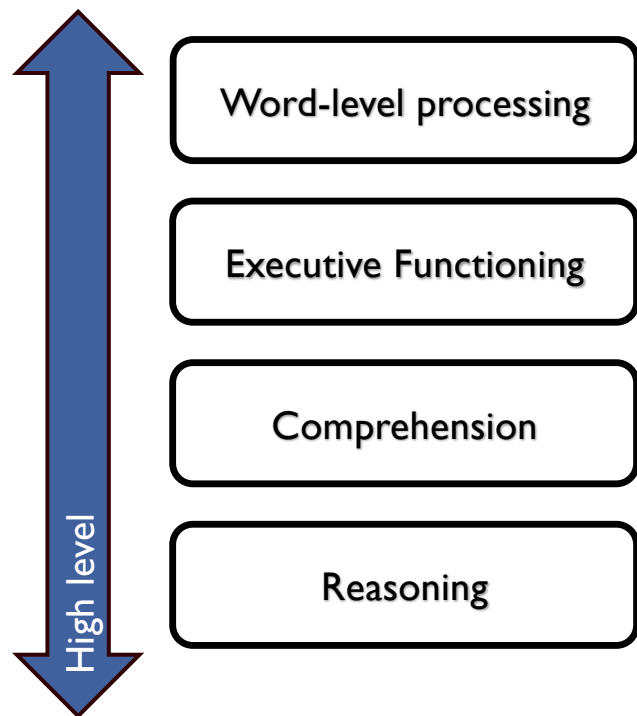
Preliminaries

Cognitive Processes

Comprehension

Annotation

Cognitive Processes in Humans and LLMs



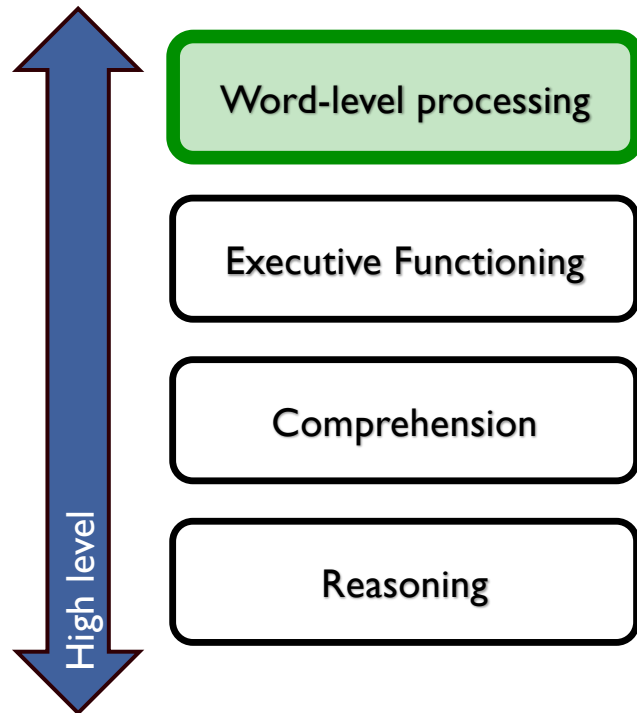
Preliminaries

Cognitive Processes

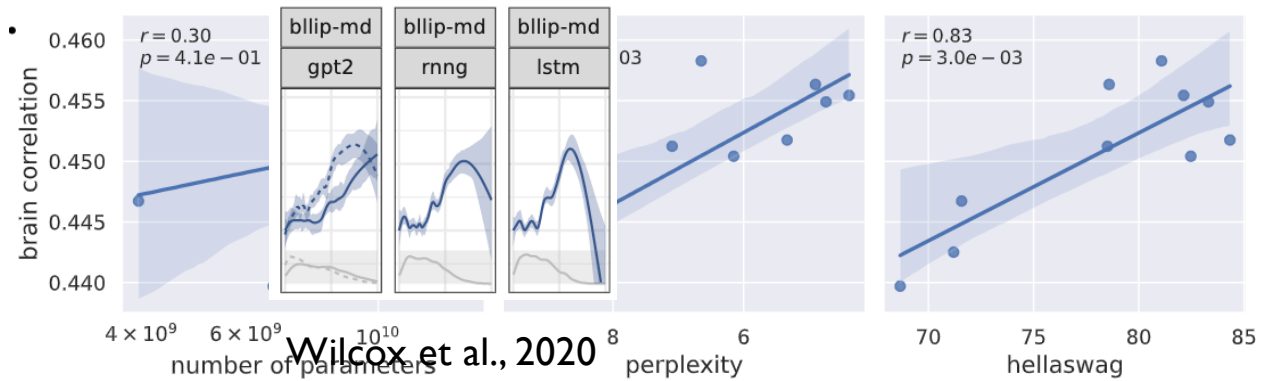
Comprehension

Annotation

LLMS MATCH WORD-LEVEL PROCESSING



Larger, better performing language models match brain tracking data



Wilcox et al., 2020

Bonnasse-Gahot and Pallier (2024)

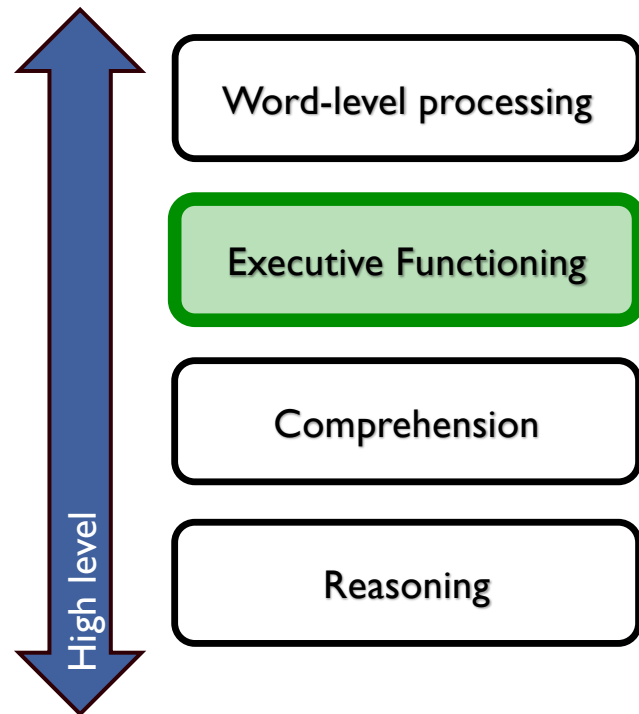
Preliminaries

Cognitive Processes

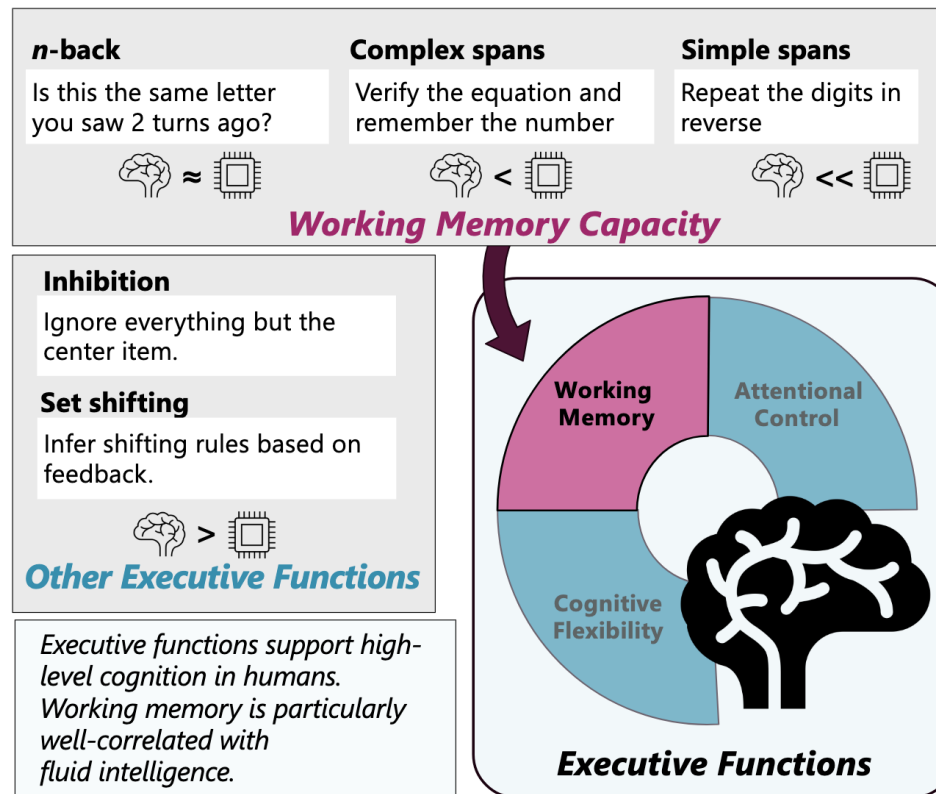
Comprehension

Annotation

WORKING MEMORY AND EXECUTIVE FUNCTION



WORKING MEMORY AND EXECUTIVE FUNCTION



de Langis et al (2025)

Preliminaries

Cognitive Processes

Comprehension

Annotation

MODELS (USUALLY) HAVE HIGHER WMC

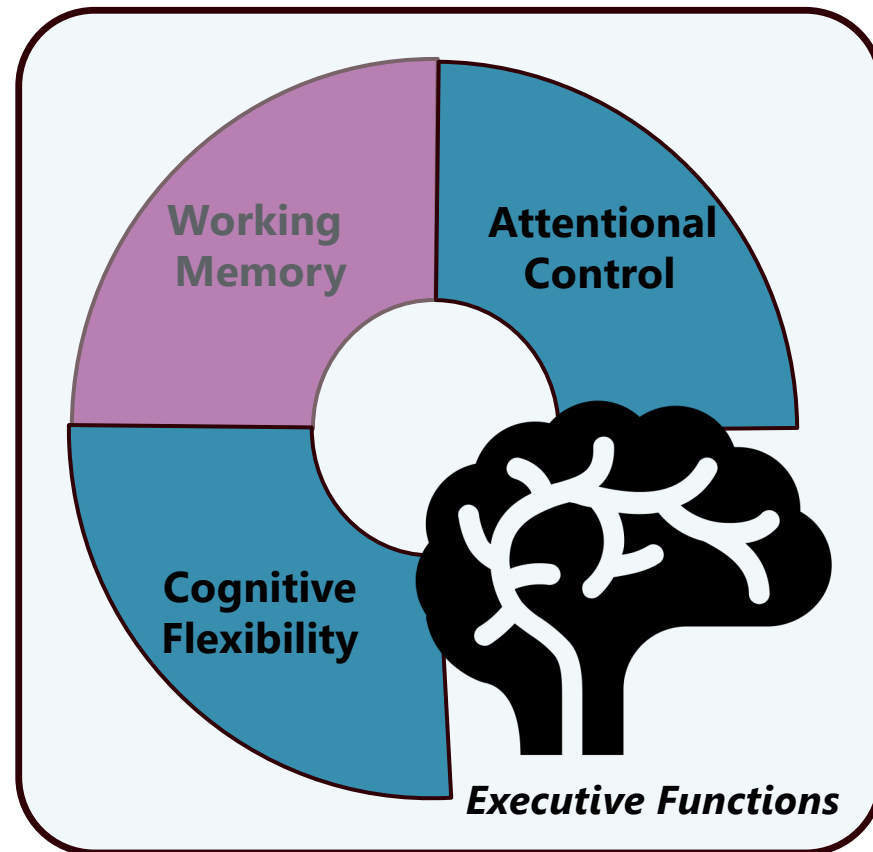
Model	1-back	2-back	3-back	O-SPAN	R-SPAN	BDS ($d = 15$)	FDS ($d = 50$)
Gemma-2-9B	0.99	0.75	0.72	0.93	0.97	0.21	0.99
Gemma-2-27B	0.91	0.72	0.69	0.92	0.98	0.59	1.00
Llama-3.1-8B	0.76	0.68	0.67	0.99	0.92	0.18	1.00
Llama-3.1-70B	0.93	0.82	0.82	0.99	0.94	0.83	1.00
Qwen2-7B	0.99	0.89	0.85	0.96	0.66	0.00	1.00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

Preliminaries

Cognitive Processes

Comprehension

Annotation

ATTENTION CONTROL

X X X X X

Preliminaries

Cognitive Processes

Comprehension

Annotation

ATTENTION CONTROL

C C X C C

Preliminaries

Cognitive Processes

Comprehension

Annotation

ATTENTION CONTROL

V V B V V

Preliminaries

Cognitive Processes

Comprehension

Annotation

ATTENTION CONTROL

X X C X X

Preliminaries

Cognitive Processes

Comprehension

Annotation

ATTENTION CONTROL

C C B C C

Preliminaries

Cognitive Processes

Comprehension

Annotation

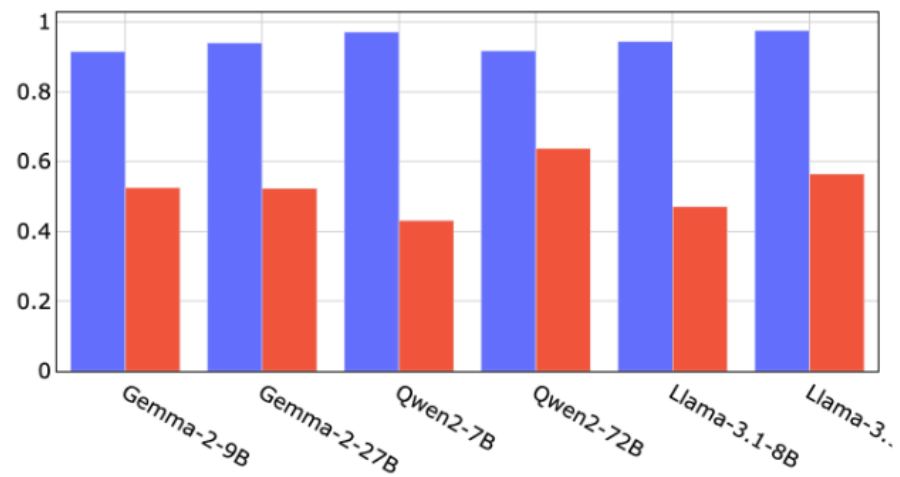
FLANKER TASK

People have ~100% accuracy

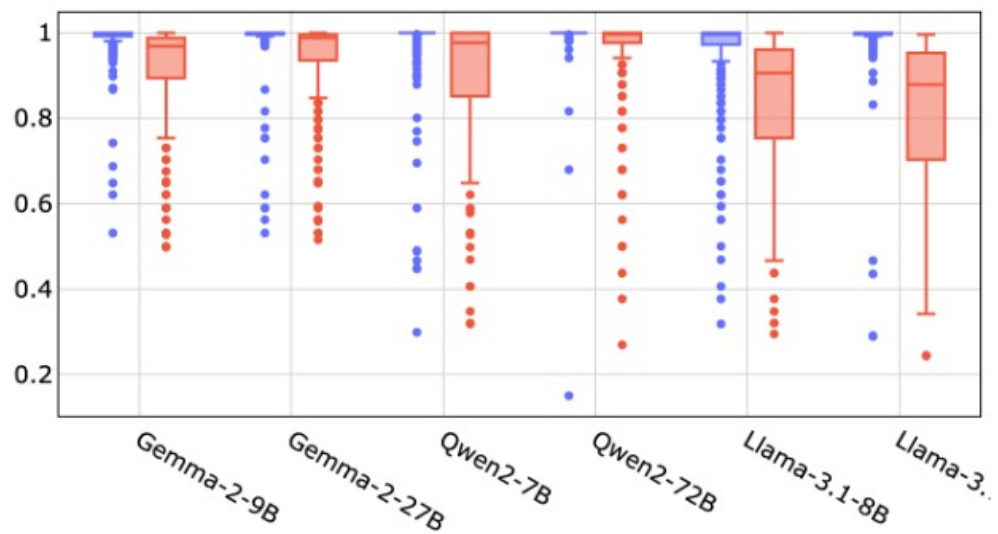
but when the flanker letters are “right” and the center letter is “left”

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.

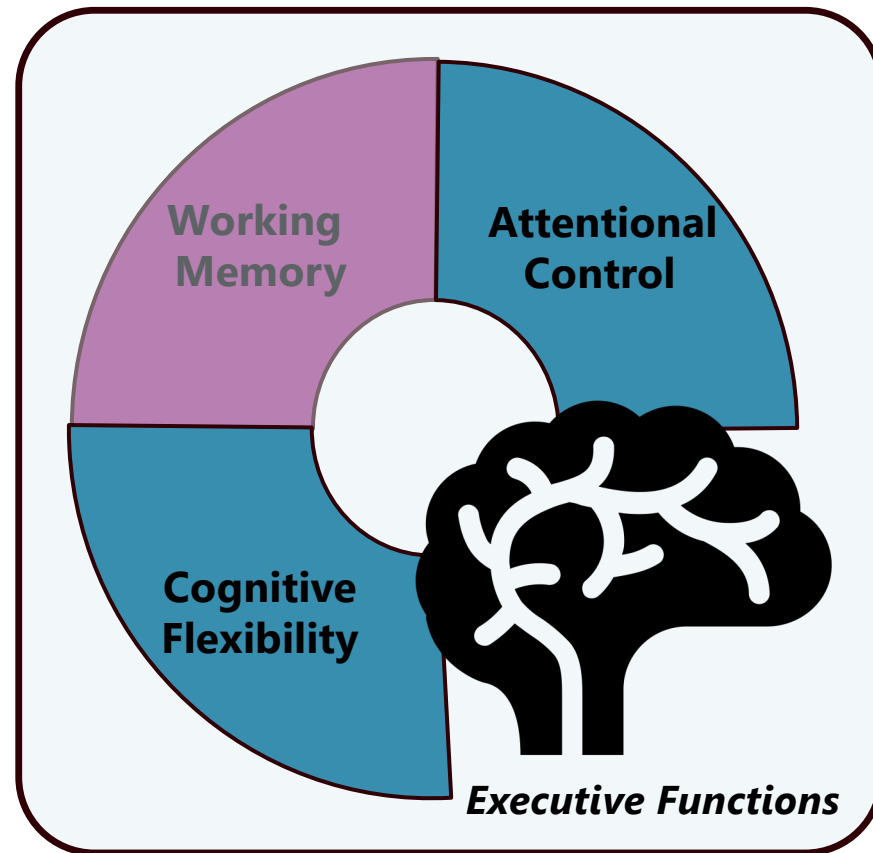
Preliminaries

Cognitive Processes

Comprehension

Annotation

WHAT ABOUT COGNITIVE FLEXIBILITY?



WISCONSIN CARD SORTING TASK

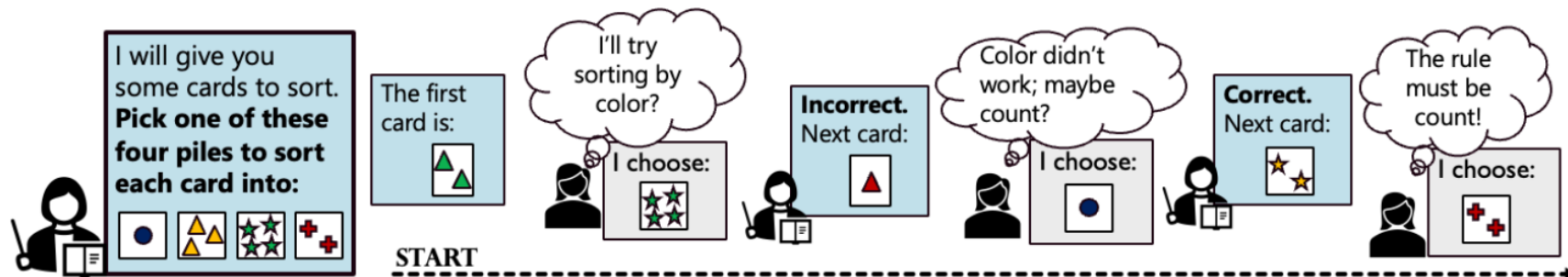
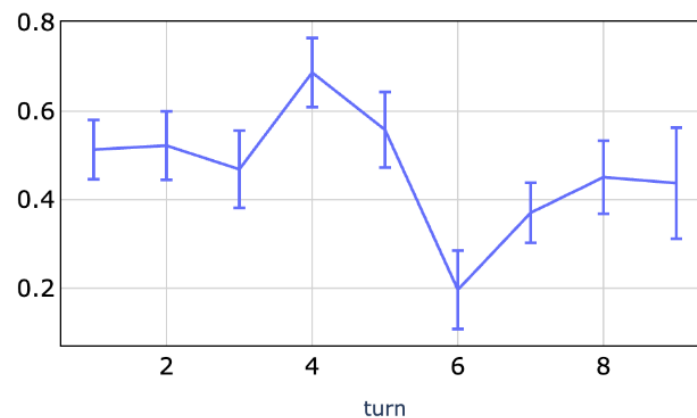


Figure 5: The Wisconsin Card Sorting Task (WCST). Participants are serially presented with cards to sort, and they must infer an underlying “sorting rule” based on feedback. *The rule will periodically change without warning*, and people must detect the change and adjust to the new rule. Participants 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
<i>Healthy adults</i>		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

Preliminaries

Cognitive Processes

Comprehension

Annotation

REASONING DOES NOT HELP

Llama-8B-Instruct:
56% accuracy



RI-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.

Preliminaries

Cognitive Processes

Comprehension

Annotation

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

INTRODUCTION

Tim was at home enjoying a moment of solitude, reading a book. In a few minutes, he would have to start thinking about what he needed to get done. (...) He was going to have one shot at this and he wanted to get it right.

INCONSISTENT SITUATION

Tim was going to propose to his girlfriend. Their evening would begin at Chez Loui, an elegant French restaurant. Chez Loui was a very formal place and Tim wanted to look his best. After he proposed a toast, Tim would ask for her hand in marriage.

CONSISTENT SITUATION

He was going to tar his roof and then lay down shingles. Tim knew tarring was messy and sticky work. On hot days, the tar seemed to get everywhere. He knew by the end of the day he would be covered with the stuff.

ENDING

Tim went about getting ready. He had a hard time choosing what to wear. **At last he grabbed some old faded jeans.** He searched his drawers for his socks. Tim finished getting ready and grabbed his keys and wallet. Tim locked the door behind him and was on his way.

de Langis et al (2025)

PROBES REVEAL DISTINCTION ONLY AT EVENT

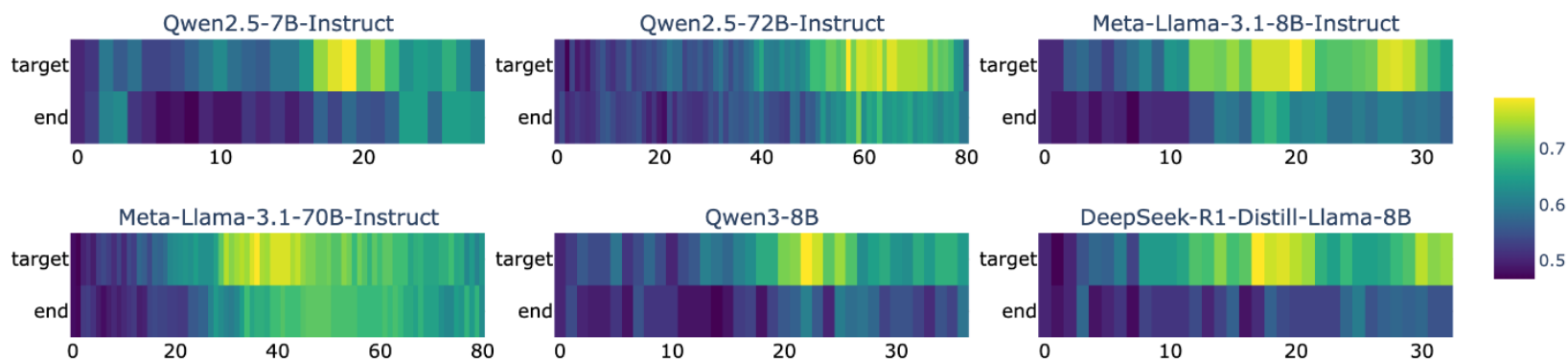


Figure 4: Mean accuracies across 10-fold cross-validation for probing hidden state representations to identify incoherent narratives. We probe both at the end of the target sentence that contains the incoherence in the incoherent story version (target), and at the conclusion of the story (end). The x 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

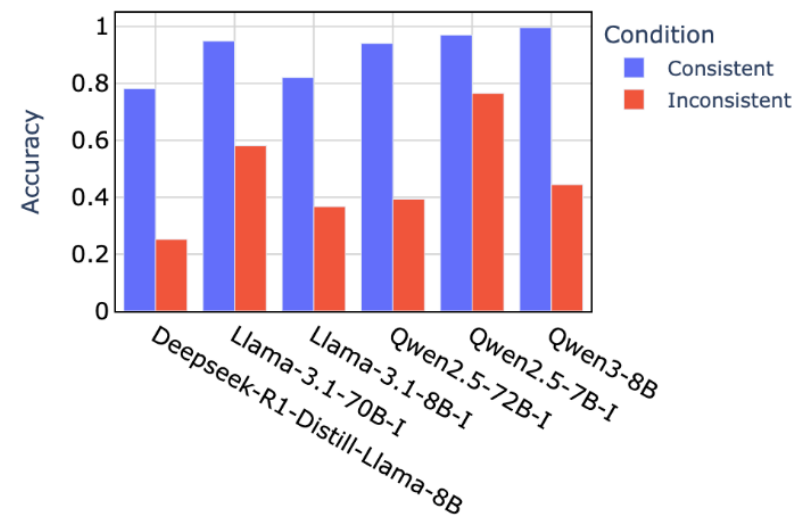


Figure 6: Model accuracies by condition when responding True or False to whether a story is coherent.

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

Preliminaries

Cognitive Processes

Comprehension

Annotation

Preliminaries

Cognitive Processes

Comprehension

Annotation

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!!