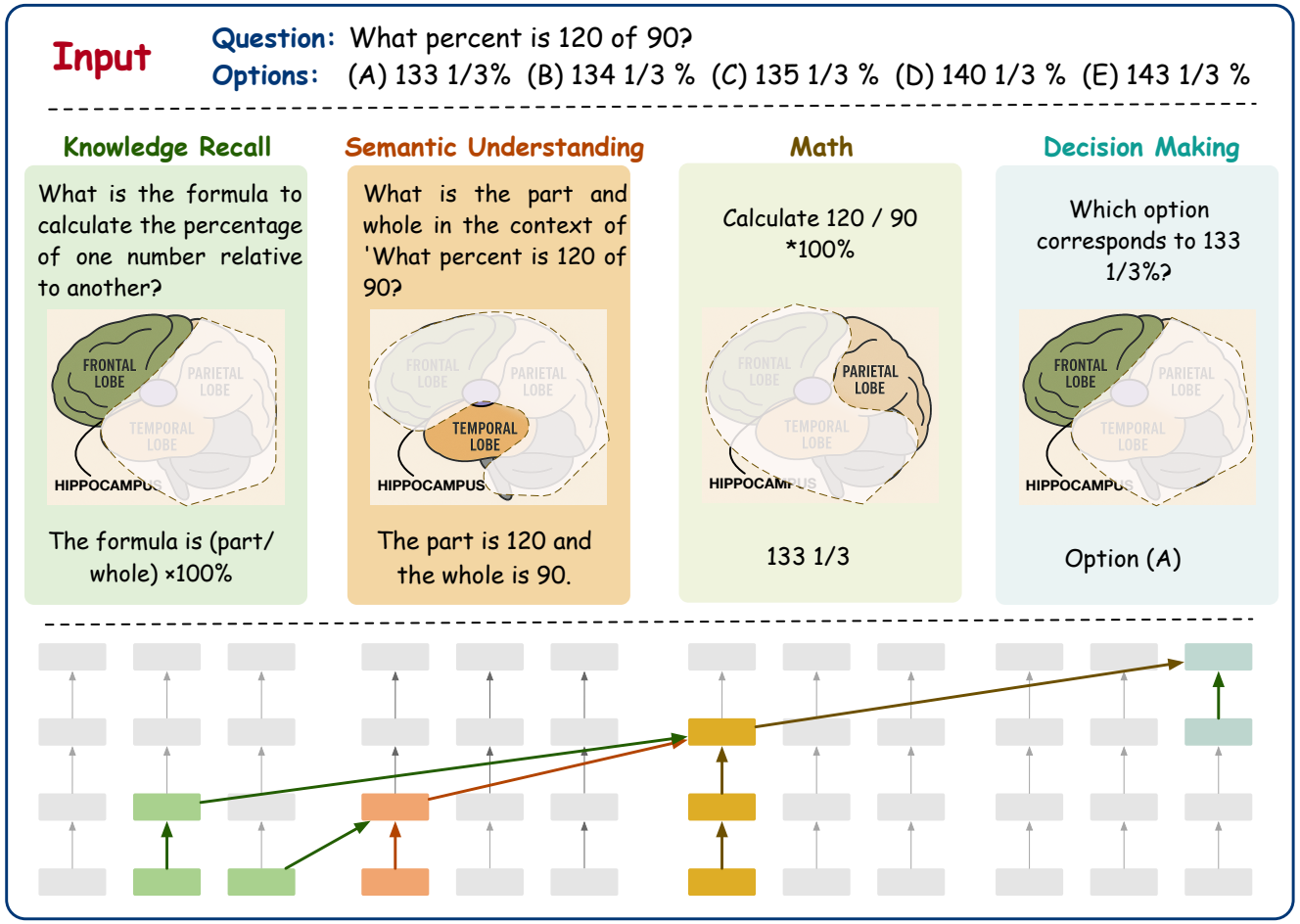
Cognitive Mirrors: Exploring the Diverse Functional Roles of Attention Heads in LLM Reasoning

Xueqi Ma [#], Jun Wang [#^], Yanbei Jiang [#], Sarah Erfani [#],
Tongliang Liu [*], James Bailey [#]

[#] The University of Melbourne
*The University of Sydney
^Amazon

Background and Motivation

To solve a complex question, the human brain engages multiple regions to perform distinct cognitive functions necessary for generating a response.



Background and Motivation

Question:

Whether there are specific attention heads in LLM play functional roles in producing answers.

We present a novel interpretability framework to systematically analyze the cognitive roles of attention heads during complex reasoning.

CogQA (3,402 subQAC)

- ➢ The low-level cognitive functions include:
- Retrieval
- Knowledge Recall
- Semantic Understanding
- Syntactic Understanding
- ➢ The high-order cognitive functions include:
- Mathematical Calculation
- Logical Reasoning
- Inference
- Decision-Making

Example 1:		
Main Question	A one-year subscription to a newspaper is offered with a 45% discount. How much does the discounted subscription cost if a subscription normally costs $80?	
Answer	We calculate first the discount: 80 × 45 / 100 = $36. So, the discounted subscription amounts to 80 – 36 = $44.	
Subquestion	**Answer**	**Cognitive Label**
1. What is the normal cost of a one-year subscription to the newspaper?	$80	Retrieval
2. What is the discount percentage offered on the subscription?	45%	Retrieval
3. How much is the discount amount in dollars for the subscription?	$36	Math Calculation
4. What is the cost of the subscription after applying the discount?	$44	Math Calculation
Example 2:		
Main Question	What does every person talk out of? Options: - name - hide - mother and father - mouth - heart	
Answer	By mouth, talking is done. Every person talk out of mouth.	
Subquestion	**Answer**	**Cognitive Label**
1. What is the primary function of talking?	To communicate verbally.	Knowledge Recall
2. Which part of the human body is primarily used for verbal communication?	Mouth	Knowledge Recall
3. Based on the options provided, which option corresponds to the part used for verbal communication?	Mouth	Decision-making

Probing Method

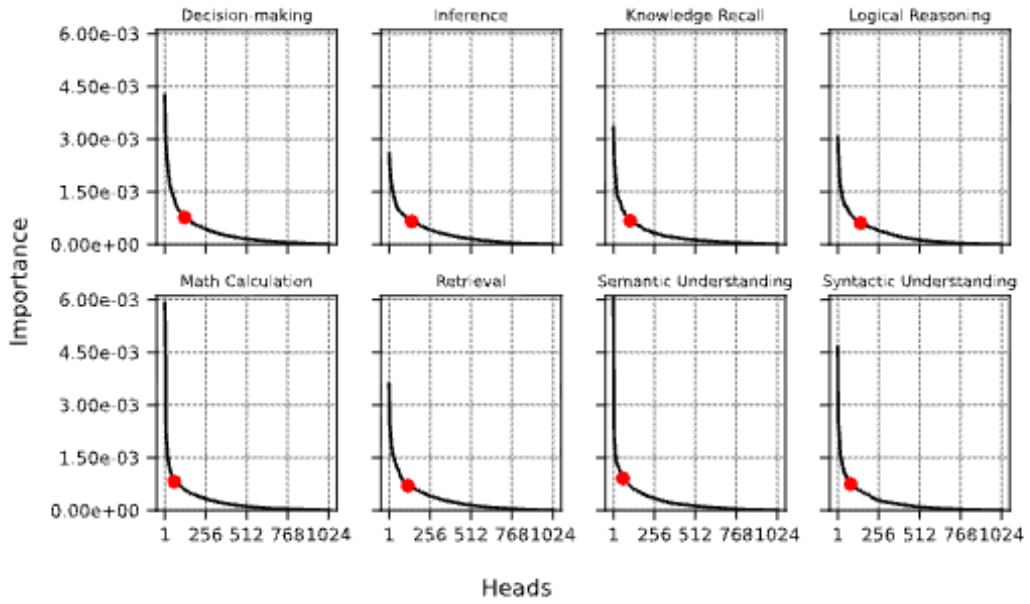
- **Probing data with head activations**

$$\mathcal{D}_{\text{probe}} = \left\{ (\bar{x}_l^{m'},\ c)_i \right\}_{i=1}^{N}, l \in \{1, \ldots, L\},\ m \in \{1, \ldots, M\}$$

- **Two-Layer MLP for classification**
- **Importance score for each head**

$$I_j^{(c)} = \mathbb{E}_{(\bar{x},c)\sim\mathcal{D}_{\text{probe}}} \left[\frac{\partial \hat{y}_c}{\partial \bar{x}_j} \cdot \bar{x}_j \right]$$

- **Heads with high importance score identified as cognitive heads**



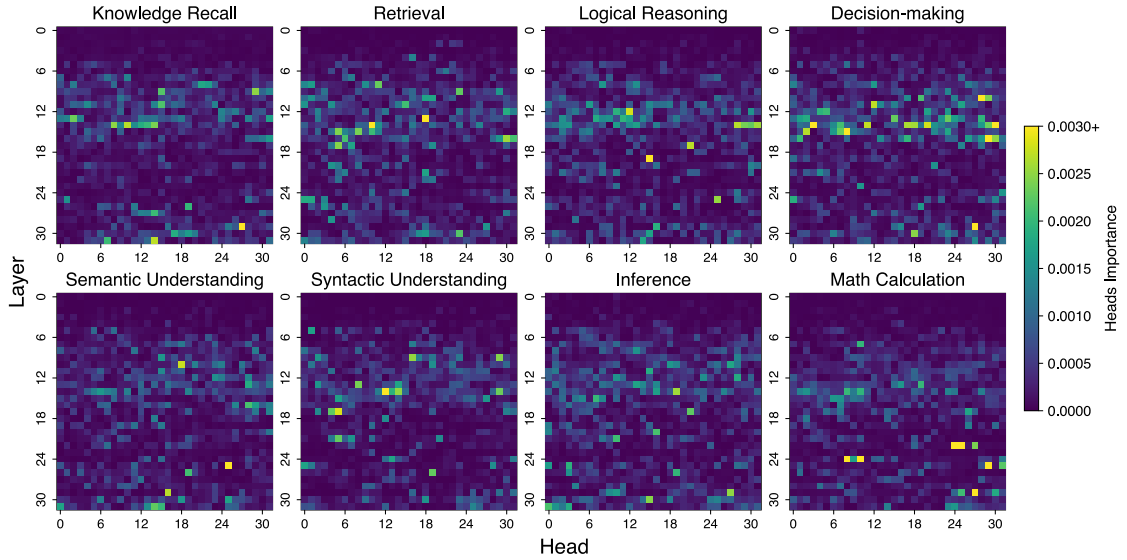
Experiments

THE UNIVERSITY OF MELBOURNE

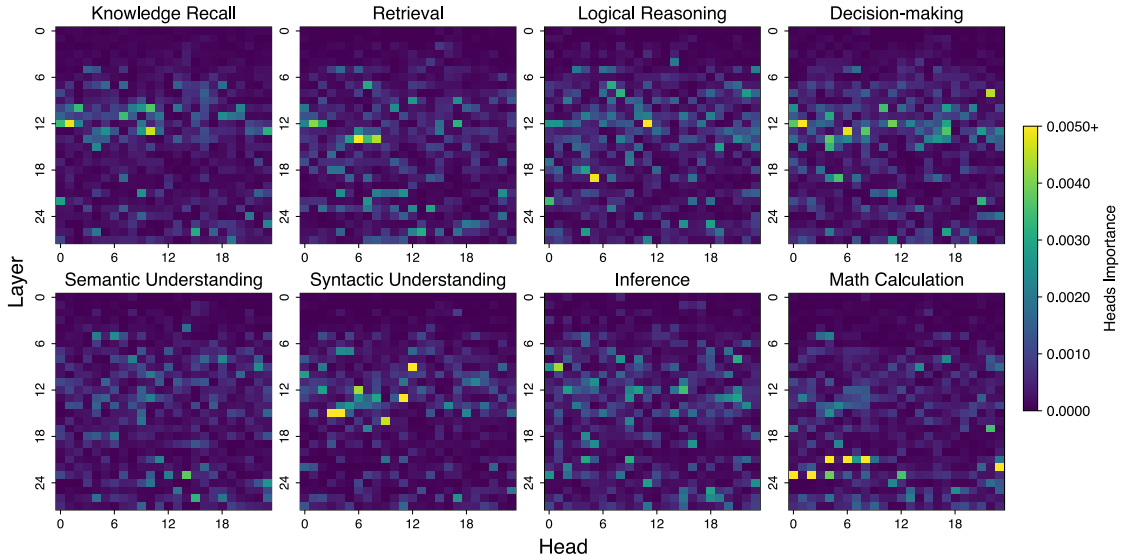
❖ Three LLM Families

- Llama3.1-8b and Llama3.2-3b
- Qwen3-8b and Qwen3-4b
- Yi-1.5-9b and Yi-1.5-6b

❖ Properties of Cognitive Heads (Sparsity, universally, layered functional organization)



Llama3.1-8b



Llama3.2-3b

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

THE UNIVERSITY OF MELBOURNE

❖ Functional Contributions of Cognitive Heads

Model	Inter_Head	Information Extraction and Analysis Functions								Higher-Order Processing Functions							
		Retrieval		Recall		Semantic		Syntactic		Math		Inference		Logic		Decision	
		comet	acc	comet	acc	comet	acc	comet	acc	comet	acc	comet	acc	comet	acc	comet	acc
Llama3.1-8B	random	90.83	84.71	87.85	83.84	91.44	97.50	87.81	66.17	94.25	83.08	91.90	70.18	91.39	54.69	97.64	90.91
	cognitive	**44.96**	**8.24**	**56.93**	**38.38**	**81.98**	**75.00**	**69.20**	**40.00**	**87.81**	**66.17**	**76.65**	**52.63**	**52.07**	**4.69**	**56.02**	**4.55**
Llama3.2-3B	random	87.89	86.47	76.35	68.69	90.54	90.00	75.82	40.00	94.98	69.65	95.66	85.96	92.75	76.56	93.30	81.82
	cognitive	**49.47**	**17.06**	**49.69**	**13.13**	**52.29**	**10.00**	**43.62**	**0.00**	**92.01**	80.10	**53.60**	**7.02**	**46.69**	**0.00**	**49.25**	**0.00**
Qwen3-8B	random	92.81	75.29	89.90	53.54	92.73	42.50	88.60	80.00	92.69	60.20	94.45	24.56	94.15	20.31	96.52	31.82
	cognitive	**59.19**	**38.24**	**64.81**	**30.30**	**85.95**	47.50	**46.26**	**0.00**	**89.29**	**53.23**	**72.77**	35.09	**87.61**	21.88	**83.17**	54.55
Qwen3-4B	random	94.17	84.71	84.61	77.78	86.91	77.50	98.15	80.00	87.15	44.78	96.89	87.72	92.00	75.00	94.79	72.73
	cognitive	**80.13**	**64.71**	**63.10**	**35.35**	**65.95**	**60.00**	**46.25**	**0.00**	**82.40**	46.27	**84.88**	**64.91**	**82.79**	**39.06**	**45.49**	**13.64**
Yi-1.5-9B	random	86.83	79.41	82.02	54.55	77.40	35.00	81.53	60.00	76.04	36.32	89.83	36.84	87.53	42.19	86.27	63.64
	cognitive	**52.76**	**21.76**	**45.99**	**9.09**	**47.25**	**2.50**	**48.10**	**40.00**	**54.22**	**16.92**	**52.41**	**15.79**	**82.75**	**26.56**	**62.85**	**18.18**
Yi-1.5-6B	random	80.64	69.41	68.82	38.38	77.83	55.00	69.61	60.00	73.33	43.78	77.71	22.81	81.65	29.69	88.54	72.73
	cognitive	**49.90**	**15.29**	**68.23**	41.41	**49.54**	**2.50**	**42.92**	**0.00**	76.64	43.78	**68.53**	**14.04**	**44.94**	**0.00**	**86.28**	**50.00**

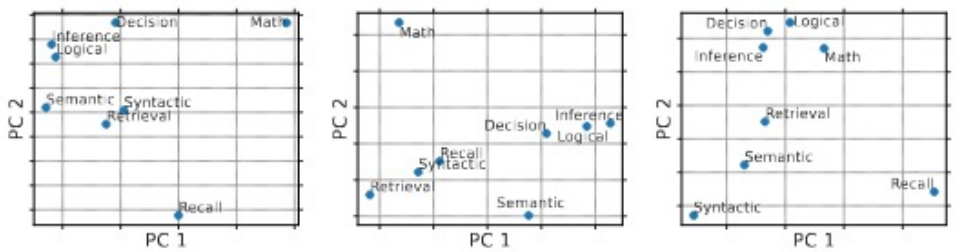
Masking cognitive heads leads to a significant decline in performance, whereas masking an equal number of random heads results in only marginal degradation across all LLMs.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

❖ Relationship Among Cognitive Heads

➢ Functional Clustering



(a) Llama3.1-8B (b) Qwen3-4B (c) Yi-1.5-6B

Related cognitive functions localize in overlapping brain regions

Heads linked to reasoning, inference, and decision-making group closely

➢ Hierarchical Structure

Retrieval	Knowledge	Semantic	Syntactic	Math	Inference	Logical	Decision
✗	✓	✓	✓	0.00 ↓ 100	0.00 ↓ 100	0.00 ↓ 100	0.00 ↓ 100
✓	✗	✓	✓	0.00 ↓ 100	0.00 ↓ 100	0.00 ↓ 100	0.00 ↓ 100
✓	✓	✗	✓	66.67 ↓ 33.33	88.24 ↓ 11.76	93.10 ↓ 8.90	57.14 ↓ 42.86
✓	✓	✓	✗	-	76.92 ↓ 23.08	100 0.00	100 0.00

Lower-level functions such as retrieval and comprehension support higher-level inference and decision-making

Masking attention heads tied to early-stage functions effect on later tasks

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

THE UNIVERSITY OF MELBOURNE

❖ Influence of Cognitive Heads on Downstream Tasks

➢ Negative Intervention

Negative Intervention Case

Query: Question: Solve -42*r + 27*c = -1167 and 130*r + 4*c = 372 for r.
Original Output: Okay, so I need to solve this system of equations: -42r + 27c = -1167 and 130r + 4c = 372 ... First equation: -42*4 + 27*(-37) = -168 - 999 = -1167. Correct. Second equation: 130*4 + 4*(-37) = 520 - 148 = 372. Correct. Yes, that's right. So everything checks out. So I think that's the correct answer. Therefore, r is 4. **Negative Intervention (Math) Output:** ... To eliminate c, I need the coefficients of c in both equations to be the same (or negatives). The coefficients are 27 and 4. The least common multiple of 27 and 4 is 100? Wait, 27 is 3^3, and 4 is 2^2, so LCM is $2^2 * 3^3 = 4 * 27 = 100$. So if I can make the coefficients of c to be 100 and 100, or 27 and 4. Wait, but 27 and 4. So if I multiply the first equation by 4 and the second equation by 27, then the coefficients of c would be 27*4 = 100 and 4*27 = 100. Then, if I subtract the equations, the c terms would be 100 and 100, but since they are both positive, subtracting the equations would eliminate c. ...

Masking the Math Calculation heads leads to errors in arithmetic tasks, while retrieval and language functions remain largely unaffected

➢ Positive Intervention

Positive Intervention Case

Query: On a school trip to the seashore, Alan and his friends collected shells. Alan collected four times as many shells as Ben did. Ben got a late start and only collected a third of what Laurie did. If Laurie collected 36 shells how many did Alan collect?
Original Output: Alan collected four times as many shells as Ben did, and Ben collected a third of what Laurie did. Since Laurie collected 36 shells, Ben collected 36 / 3 = 12 shells. Therefore, Alan collected 4 * 12 = 48 shells. However, the question asks for the number of shells Alan collected, and it is given that Alan collected four times as many shells as Ben did. Therefore, the correct answer is 4 * 12 = 48 * 4 / 3 = 72.
Positive Intervention (Math) Output : ...Since Ben collected a third of what Laurie did, he collected 36 / 3 = 12 shells. Alan collected four times as many shells as Ben did, so he collected 4 * 12 = 48 shells...The correct answer is 48

Adjusting cognitive head activations along their functional directions reliably enhances performance on complex reasoning tasks

NEURAL INFORMATION PROCESSING SYSTEMS

Conclusion

- We propose an interpretability framework that connects attention heads in large language models (LLMs) to human cognitive functions involved in reasoning.
- We introduce CogQA, a cognitively grounded dataset that decomposes complex questions into step-by-step subquestions with a chain-of-thought design.
- Leveraging CogQA, we develop a multi-class probing method to identify and characterize attention heads responsible for distinct cognitive operations within the transformer architecture.
- Our analysis across multiple LLM families and scales demonstrates that attention heads exhibit universality, sparsity, intrinsic roles, and dynamic, hierarchical organization.

THE UNIVERSITY OF MELBOURNE

Thank You!

Xueqi Ma

xueqim@student.unimelb.edu.au