

# Hierarchical Reasoning Model: разбор разбора

Фуфаев В. В.

11 сентября 2025 г.

## О себе

Фуфаев Владимир Владимирович

Мехмат МГУ, Кандидат наук (2006-2018)

<https://istina.msu.ru/profile/FufaevVV>

Постдок ФКН ВШЭ (2019-2021)

ВТБ, УПАМО (с 2021)



# Проблема

Удивительная неэффективность LLM в решении алгоритмических задач<sup>1</sup>

Model	Percent of puzzles solved fully	Percent of cells answered correctly
GPT-4o	0%	9.5%
Gemini-1.5 Pro	0%	10.2%

Causal Language Modeling Can Elicit Search and Reasoning Capabilities on Logic Puzzles, 2024

---

<sup>1</sup>Непостижимая эффективность математики в естественных науках. Ю.Вигнер

# Проблема

Удивительная неэффективность LLM в решении алгоритмических задач<sup>1</sup>

Model	Percent of puzzles solved fully	Percent of cells answered correctly
GPT-4o	0%	9.5%
Gemini-1.5 Pro	0%	10.2%

Causal Language Modeling Can Elicit Search and Reasoning Capabilities on Logic Puzzles, 2024

Chain-of-Thought?

Обучение на паззлах?

---

<sup>1</sup>Непостижимая эффективность математики в естественных науках. Ю.Вигнер

## Авторы статьи

Hierarchical Reasoning Model, 2025

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu,  
Sen Song, Yasin Abbasi Yadkori



Sapient Intelligence, Singapore

## Авторы статьи

Hierarchical Reasoning Model, 2025

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu,  
Sen Song, Yasin Abbasi Yadkori



Sapient Intelligence, Singapore

Sen Song



Tsinghua University, China

## Авторы статьи

Hierarchical Reasoning Model, 2025

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu,  
Sen Song, Yasin Abbasi Yadkori



Sapient Intelligence, Singapore

Sen Song



Tsinghua University, China

<https://discord.gg/sapient>



<https://github.com/sapientinc/HRM>



## Автор разбора

Николенко Сергей Игоревич

Чемпион мира ЧГК (2015, 2017,  
<https://www.wikipedia.org/>)

Кандидат физико-математических наук (2009)

ПОМИ РАН

СПбАУ

ФКН ВШЭ





## Автор разбора

Николенко Сергей Игоревич

Чемпион мира ЧГК (2015, 2017,  
<https://www.wikipedia.org/>)

Кандидат физико-математических наук (2009)

ПОМИ РАН

СПбАУ

ФКН ВШЭ

Data Fusion 2023: Современное положение дел в мультимодальном информационном поиске



<https://t.me/sinecor/>



## Анализ проблемы

Формулировка проблемы:

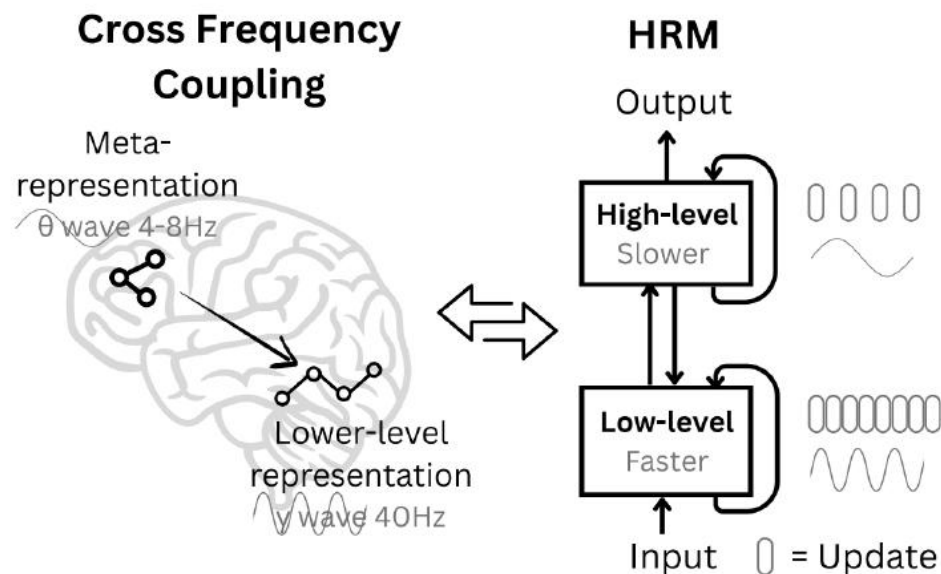
1. Есть задачи класса  $P$  (решаемые за полиномиальное время), а есть задачи, решаемые схемами постоянной глубины ( $AC^0$ ).
2. LLM “думают” на естественном языке.

## Анализ проблемы

Формулировка проблемы:

1. Есть задачи класса  $P$  (решаемые за полиномиальное время), а есть задачи, решаемые схемами постоянной глубины ( $AC^0$ ).
2. LLM “думают” на естественном языке.

Bio-inspired решение

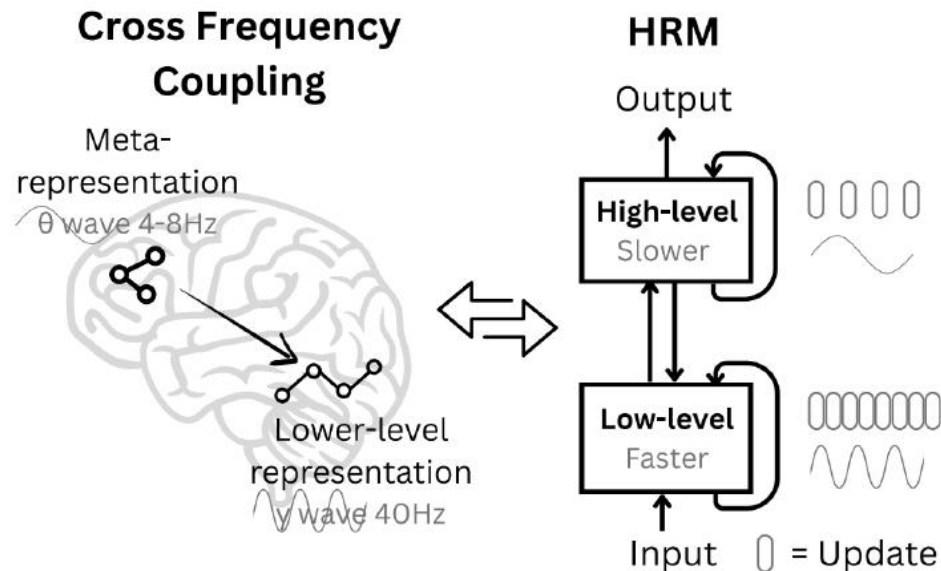


## Анализ проблемы

Формулировка проблемы:

1. Есть задачи класса  $P$  (решаемые за полиномиальное время), а есть задачи, решаемые схемами постоянной глубины ( $AC^0$ ).
2. LLM “думают” на естественном языке.

Bio-inspired решение



The HRM model consists of four learnable components: an input network  $f_I(\cdot; \theta_I)$ , a low-level recurrent module  $f_L(\cdot; \theta_L)$ , a high-level recurrent module  $f_H(\cdot; \theta_H)$ , and an output network  $f_O(\cdot; \theta_O)$ .

## Под капотом

```
98 class Attention(nn.Module):
109     self.qkv_proj = CastedLinear(self.hidden_size, (self.num_heads + 2 * self.num_key_value_head
110     self.o_proj = CastedLinear(self.output_size, self.hidden_size, bias=False)
111
112     def forward(self, cos_sin: CosSin, hidden_states: torch.Tensor) -> torch.Tensor:
113         batch_size, seq_len, _ = hidden_states.shape
114
115         # hidden_states: [bs, seq_len, num_heads, head_dim]
116         qkv = self.qkv_proj(hidden_states)
117
118         # Split head
119         qkv = qkv.view(batch_size, seq_len, self.num_heads + 2 * self.num_key_value_heads, self.head
120         query = qkv[:, :, :self.num_heads]
121
122         # key and value
123         key = qkv[:, :, self.num_heads : self.num_heads + self.num_key_value_heads]
124         value = qkv[:, :, self.num_heads + self.num_key_value_heads : self.num_heads + self.num_key_value_heads + self.num_value_value_heads]
125
126         # flash attn
127         attn_output = flash_attn_func(q=query, k=key, v=value, causal=self.causal)
128         if isinstance(attn_output, tuple): # fa2 and fa3 compatibility
129             attn_output = attn_output[0]
130
131         # attn_output: [batch_size, num_heads, seq_len, head_dim]
132         attn_output = attn_output.view(batch_size, seq_len, self.output_size) # type: ignore
133         return self.o_proj(attn_output)
```

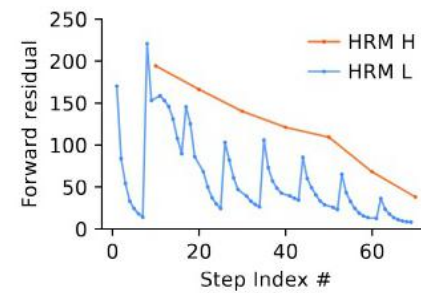


## Важные трюки

### Bio-inspired

Трюк №1: Hierarchical convergence

Медленная модель обучается  $N$  шагов,  
быстрая - TN (Решает  $N$  разных задач)



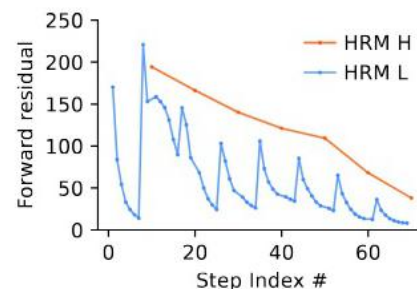


## Важные трюки

## Bio-inspired

## Трюк №1: Hierarchical convergence

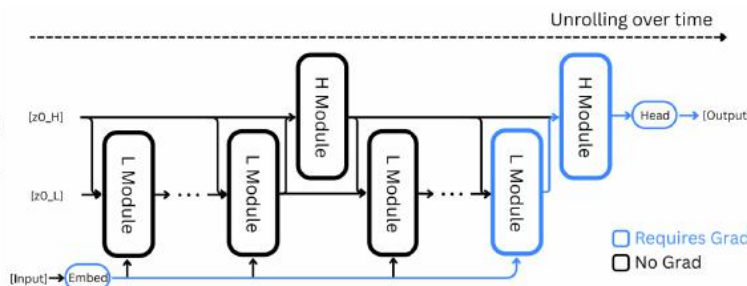
Медленная модель обучается  $N$  шагов, быстрая - TN (Решает  $N$  разных задач)



## Трюк №2: Approximate gradient

Теорема о неявной функции в окрестности точки экстремума, первые приближения.

$O(1)$  вместо  $O(T)$  по памяти.

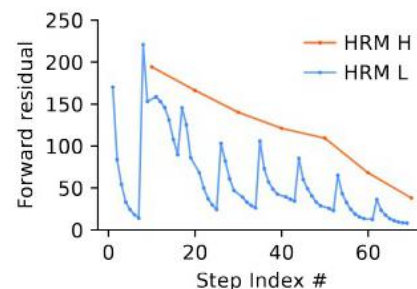


## Важные трюки

## Bio-inspired

## Трюк №1: Hierarchical convergence

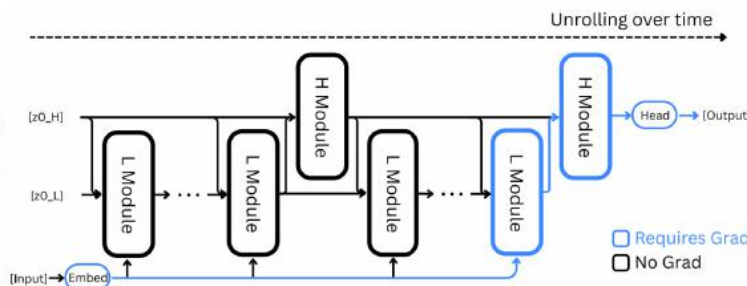
Медленная модель обучается  $N$  шагов, быстрая - TN (Решает  $N$  разных задач)



## Трюк №2: Approximate gradient

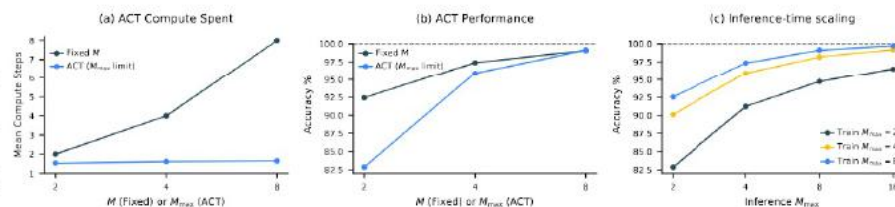
Теорема о неявной функции в окрестности точки экстремума, первые приближения.

$O(1)$  вместо  $O(T)$  по памяти.



## Трюк №3: Adaptive computational time

Подбирает число прогонов модели  $M$ . Q-learning algorithm из RL (Марковский процесс: забираем награду или движемся дальше).





Еще больше трюков!

Дополнение к трюку 3:

Deep supervision: Дополнительные лоссы на каждом шаге (регуляризация).

## Еще больше трюков!

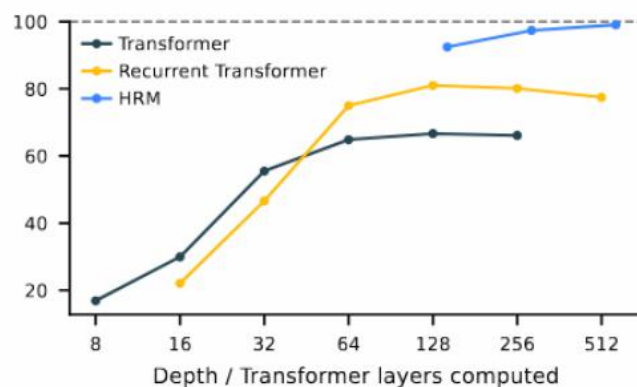
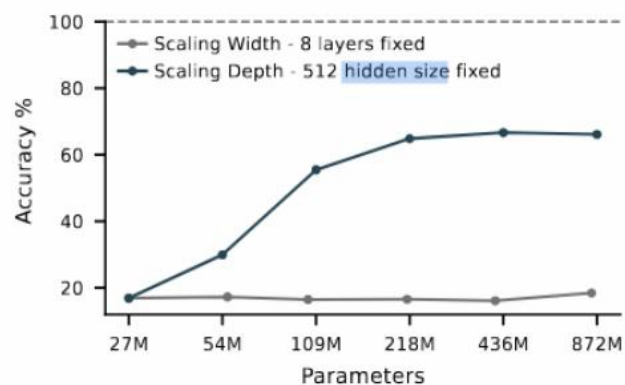
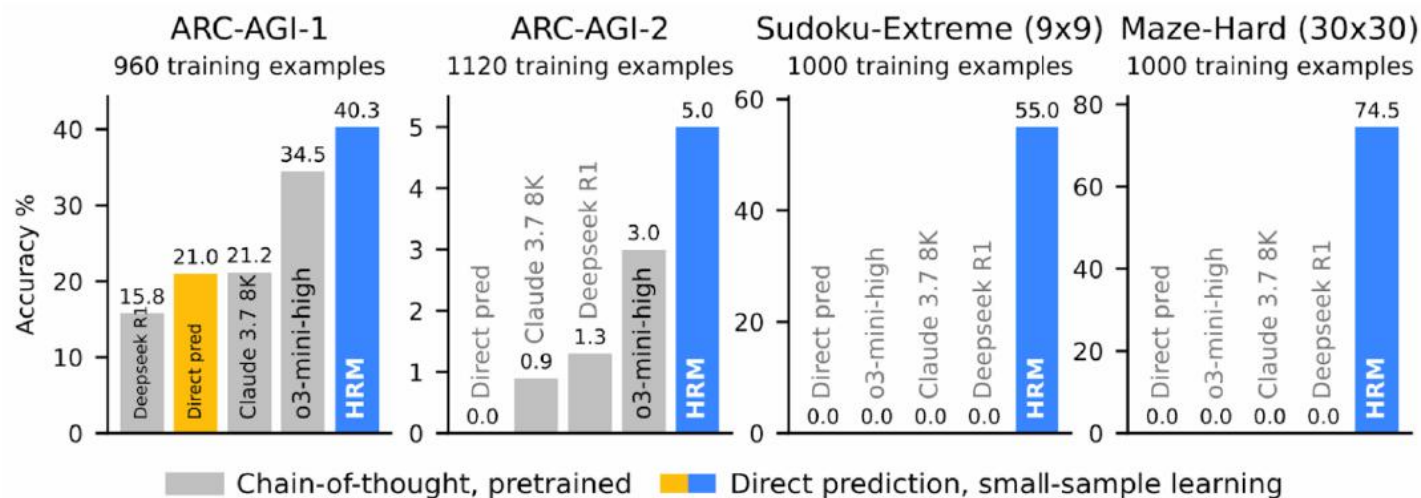
Дополнение к трюку 3:

Deep supervision: Дополнительные лоссы на каждом шаге (регуляризация).

For all Transformer blocks in this work—including those in the baseline models—we incorporate the enhancements found in modern LLMs (based on Llama<sup>53</sup> architectures). These improvements include Rotary Positional Encoding<sup>54</sup>, Gated Linear Units<sup>55</sup>, RMSNorm<sup>56</sup>, and the removal of bias terms from linear layers. Furthermore, both HRM and recurrent Transformer models implement a Post-Norm architecture. with weights initialized via truncated LeCun Normal initialization, while the scale and bias parameters are excluded from RMSNorm.

# Результаты

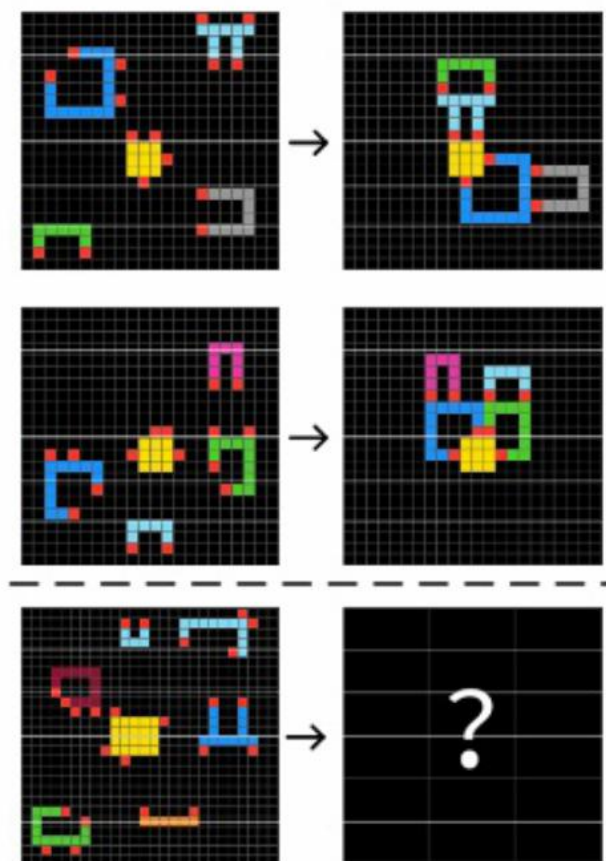
Прогресс в решении задач (27M параметров)



4M training examples

# Результаты

## Интерпретируемость

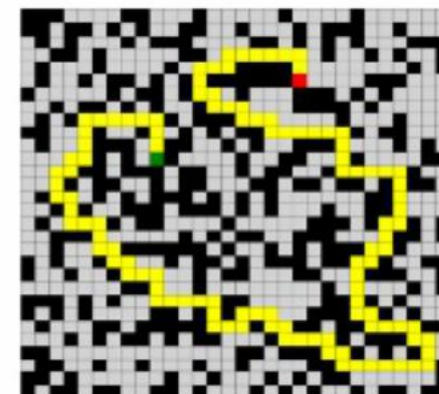
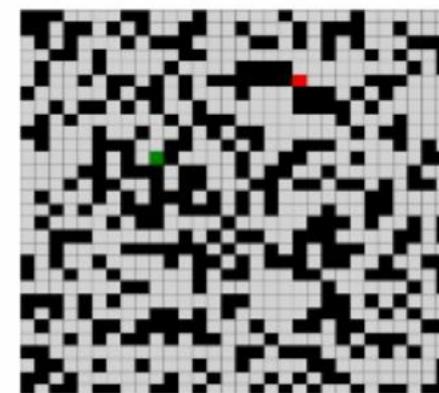


(a) ARC-AGI

	8	4			5		6
				8		7	
3				4			
	3	8	4				2
		6			3		8
9							6
			5				
				2			1
2	5		3			8	

7	8	4	1	2	5	9	6	3
2	6	1	3	8	9	7	4	5
3	5	9	6	4	7	8	1	2
5	3	8	4	9	6	1	2	7
4	1	6	2	7	3	5	9	8
9	7	2	8	5	1	4	3	6
6	9	3	5	1	8	2	7	4
8	4	7	9	6	2	3	5	1
1	2	5	7	3	4	6	8	9

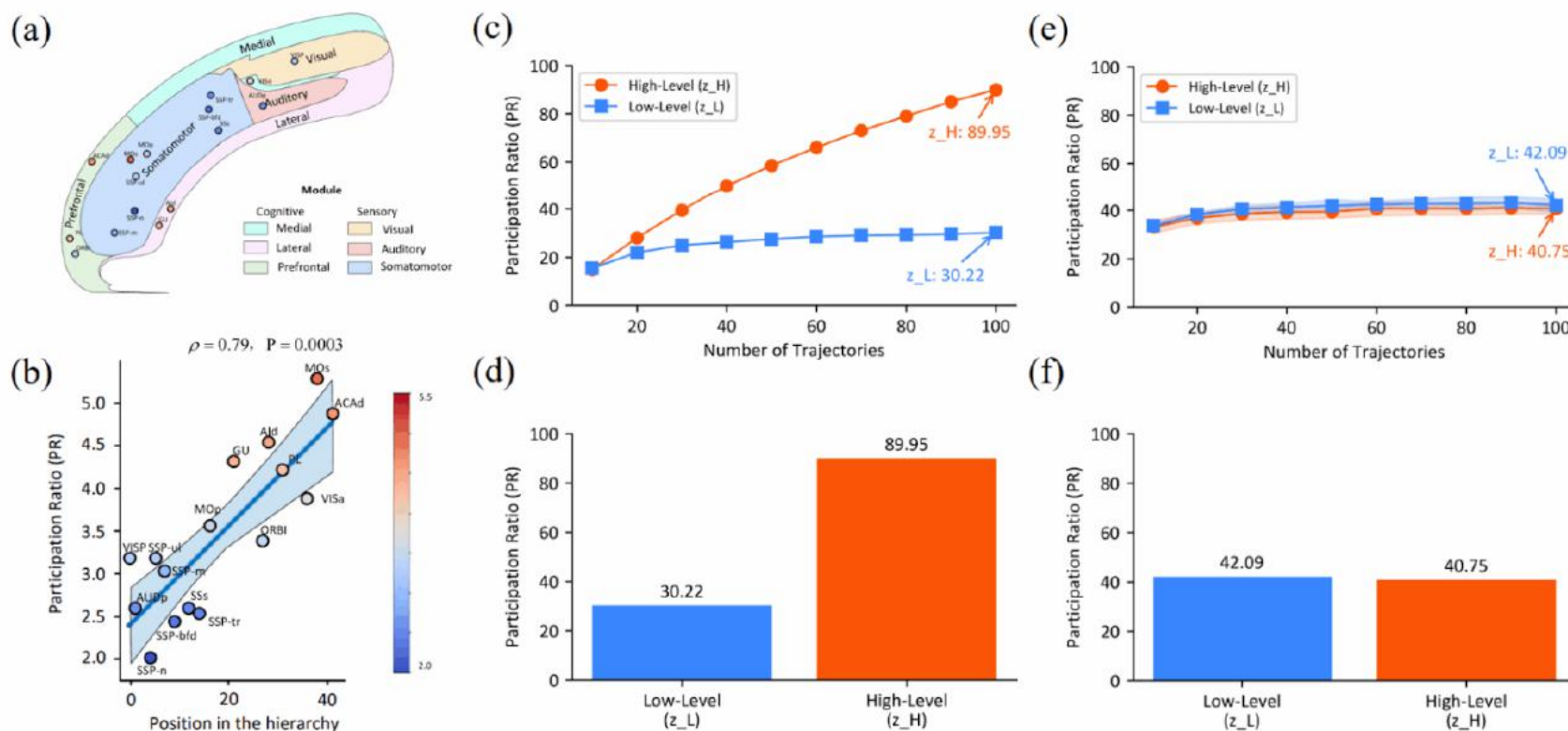
(b) Sudoku-Hard



(c) Maze navigation

# Результаты

Аналогии с биологией (Спонтанная иерархия размерностей)



Bio: Participation Ratio (PR), a measure of effective neural dimensionality

$$PR = \frac{(\sum \lambda_i)^2}{\sum \lambda_i^2}, \text{ eigenvalues of the covariance matrix of neural trajectories.}$$

HRM: trajectories = tasks, the covariance matrix derived from neural states gathered across multiple Sudoku-solving trajectories.



## Итоги

С. И. Николенко:

- 1) Что если масштабировать это до размеров современных LLM?
- 2) Это новый алгоритмический AI? (компактный, не требующий больших датасетов и специализированный на конкретных задачах) Его можно использовать как дополнение к LLM.

## Итоги

С. И. Николенко:

- 1) Что если масштабировать это до размеров современных LLM?
- 2) Это новый алгоритмический AI? (компактный, не требующий больших датасетов и специализированный на конкретных задачах) Его можно использовать как дополнение к LLM.

Разбор разбора:

- а) Статья 10/10, разбор 9/10 (-1 из-за bio-inspiration -0 за мат. подробности).
- б) Не единственный возможный способ моделировать такую работу мозга.  
(Energy-based Transformers)

## Контакты

Благодарю за внимание!



## Контакты

Благодарю за внимание!

t.me/tlg\_vld\_f



ArtAI: Theory and Practice  
t.me/Art\_AI\_T\_P



<https://github.com/fufaevv1v1>



## Контакты

Благодарю за внимание!

t.me/tlg\_vld\_f



ArtAI: Theory and Practice  
t.me/Art\_AI\_T\_P



<https://github.com/fufaevv1v1>



2025\_08\_Transformers\_Week  
Flash Attention  
Energy-based Transformers