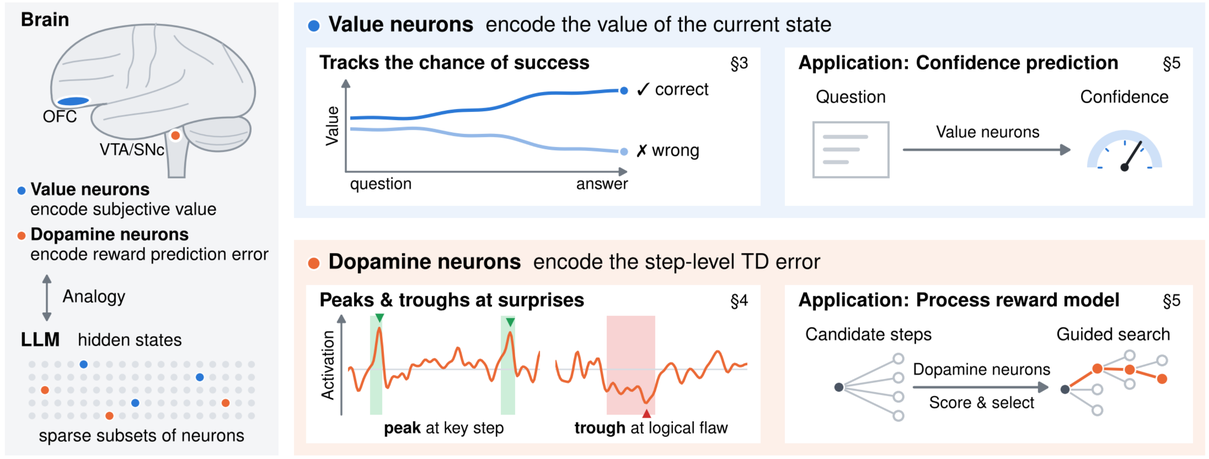
Sparse Reward Subsystem in Large Language Models

Guowei Xu[1], Mert Yuksekgonul[2], James Zou[2]

[1]Tsinghua University [2]Stanford University



1. Motivation

The human brain contains sparse, specialized neurons: **value neurons** encode the subjective reward, and **dopamine neurons** encode the difference between expected and received reward.

This is known as the brain's **reward subsystem**. We hypothesize that an analogous reward subsystem may also exist within large language models.

1.1 Why may reward subsystem exist in LLMs?

An optimal policy is tightly coupled with value functions. Since LLMs approximate strong policies, value information is likely encoded in their representations.

1.2 Why may these neurons be sparse?

(1) **Feasibility.** The terminal reward is binary, so reward information reduces to a scalar → low-dimensional encoding suffices.

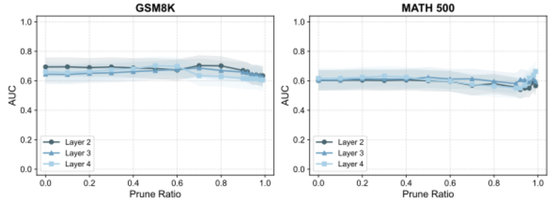
(2) **Realization.** The superposition hypothesis predicts that frequent, high-importance features receive sparse representations.

2. Value Neurons

We train a **value probe** V (2-layer MLP) on hidden states s_t using TD learning: $\mathcal{L} = \mathbb{E}\left[\sum_t (\gamma V_l(s_t) - V_l(s_{t-1}))^2\right]$. We then prune input neurons by L1 weight norm at ratio p, and measure AUC$(V_l^p(s_0), r(s_T))$.

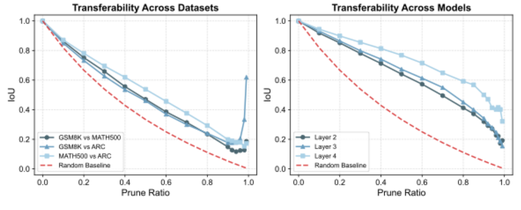
2.1 Empirical Evidence

AUC remains **stable** when neurons are pruned, confirming that value neurons are sparse.



2.2 Transferability

IoU of value neuron positions across datasets and across models consistently exceeds the random baseline.

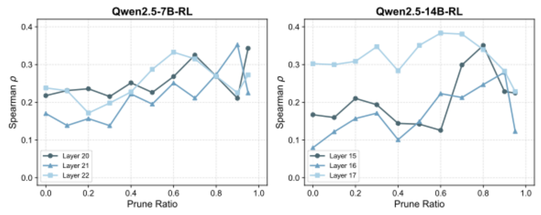


3. Dopamine Neurons

We train a **dopamine probe** D to predict step-level TD error: $\mathcal{L} = \mathbb{E}\left[\sum_t (D_l(s_t) - \delta_t)^2\right]$, where $\delta_t = \gamma \hat{V}(s_t) - \hat{V}(s_{t-1})$. The same L1 pruning identifies dopamine neurons.

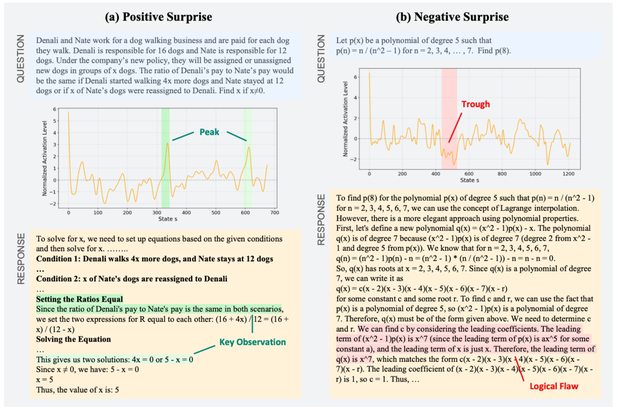
3.1 Empirical Evidence

Spearman correlation between predicted and actual TD errors remains **stable** under pruning.



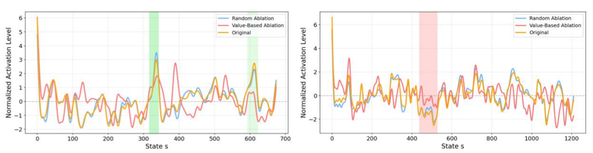
3.2 Visualization

Dopamine neurons show **activation peaks** at moments of unexpected progress (positive surprise) and **troughs** at logical flaws (negative surprise),



3.3 Functional Coherence

Disrupting value neurons in earlier layers **impairs** the reward subsystem in subsequent layers.



4. Applications

We find that value neurons can serve as predictors of **model confidence**, and dopamine neurons can function as a **process reward model**.

4.1 Predicting Model Confidence

Using value neurons to predict model confidence achieves **a higher AUC**.

Model	Value Neurons	LCD [9]	Verbal Conf.	Token Conf.	Question Len.
MATH500					
Llama-3.1-8B	**0.69**	0.64	**0.69**	0.54	0.63
Phi-3.5-mini	**0.73**	0.63	0.54	0.41	0.62
Gemma-3-4B	**0.72**	0.65	0.62	0.41	0.71
Qwen3.5-0.8B	0.71	0.70	0.41	0.56	**0.77**
ARC					
Llama-3.1-8B	**0.66**	0.56	0.55	0.49	0.55
Phi-3.5-mini	**0.63**	0.54	0.29	0.46	0.54
Gemma-3-4B	**0.59**	0.50	0.47	0.44	0.49
Qwen3.5-0.8B	**0.63**	0.54	0.55	0.50	0.65
Average	**0.67**	0.60	0.52	0.48	0.62

4.2 Process Reward Model

Using dopamine neurons as a process reward model achieves **higher accuracy**.

Method	Qwen-2.5-7B SimpleRL-Zoo	Qwen-2.5-14B SimpleRL-Zoo	Llama-3.1-8B Instruct	Average
Greedy	69.6	67.0	47.3	61.3
Random	77.3	74.8	46.3	66.1
Majority of N	64.8	76.0	54.2	65.0
Implicit PRM	69.8	74.8	45.8	63.5
Mean logprob	72.2	78.7	41.7	64.2
Min logprob	74.1	76.0	48.6	66.2
P(true)	78.7	76.0	47.2	67.3
Dopamine Neurons	**79.6**	**82.7**	**54.6**	**72.3**

5. Summary

We identify a sparse reward subsystem in LLMs: **value neurons** encode expected reward, and **dopamine neurons** encode reward prediction errors.

Homepage

Code

NEURAL INFORMATION PROCESSING SYSTEMS