



中国人民大学高瓴人工智能学院
Gaoling School of Artificial Intelligence, Renmin University of China



AI Box

Neuron-based Personality Trait Induction in Large Language Models

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Xin Zhao, Ji-Rong Wen

2021200967@ruc.edu.cn

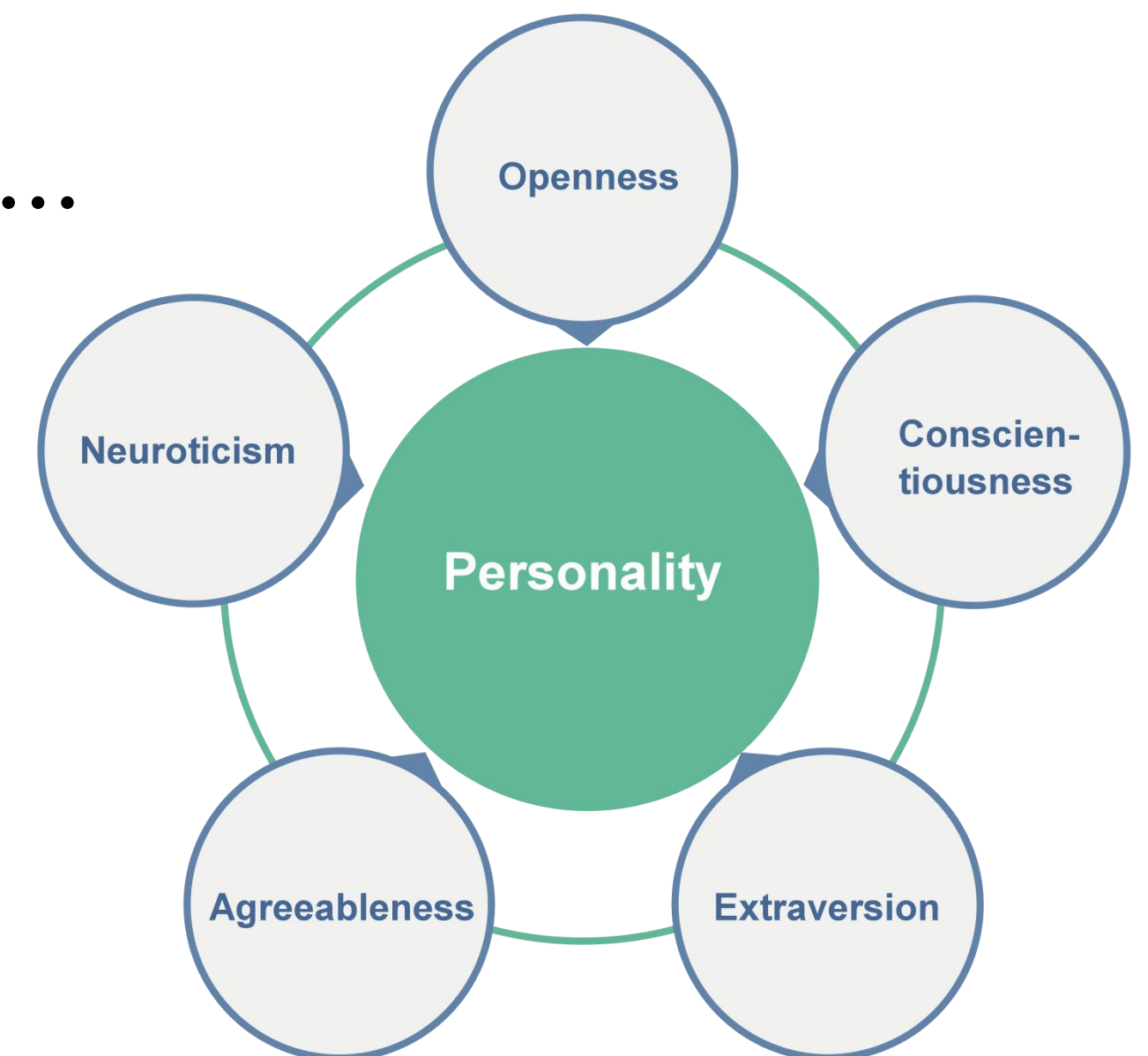
ICLR 2025 Poster



<https://github.com/RUCAIBox/NPTI>

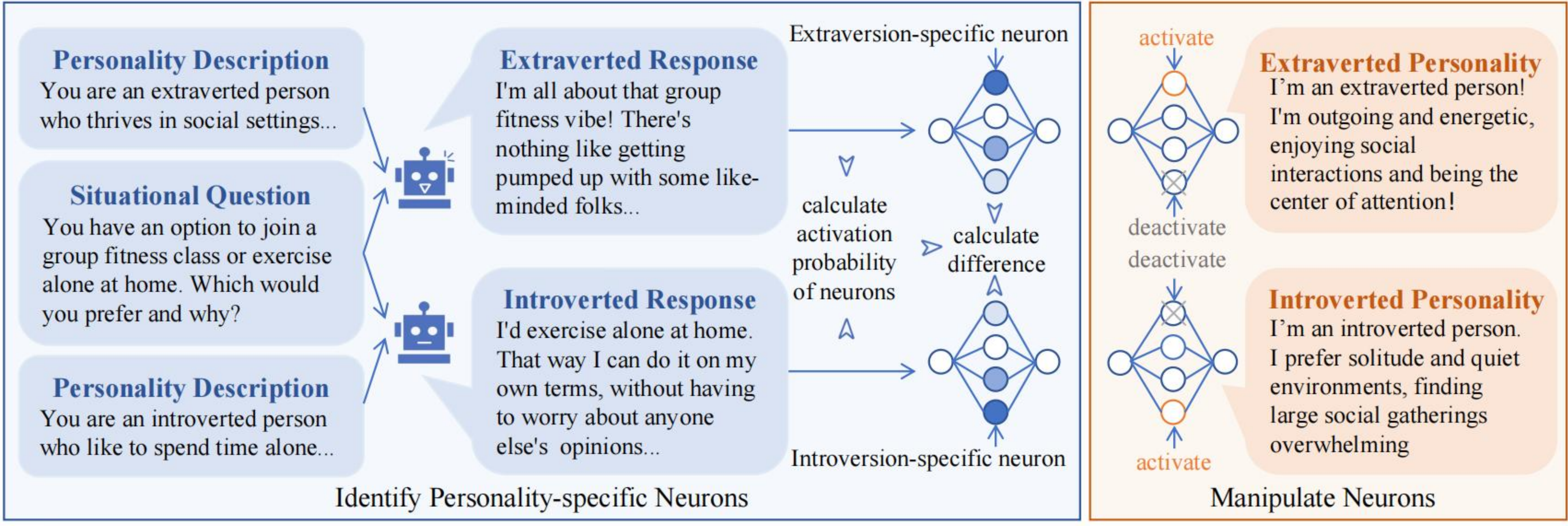
01. Introduction

- Personality in LLMs
 - **Applications:** personality simulation, sociological research...
 - **Assessment:** MBTI and BFI, LLM as judge
- Current works
 - **Prompt based:** explicit & implicit
 - **Training based:** sft, rlhf
- Challenge
 - **Prompt-based induction:** unstable, depending on the specific prompts.
 - **Model training:** resource-intensive, requiring labeled data.



01. Introduction

NPTI: Neuron-based approach for Personality Traits Induction in LLMs



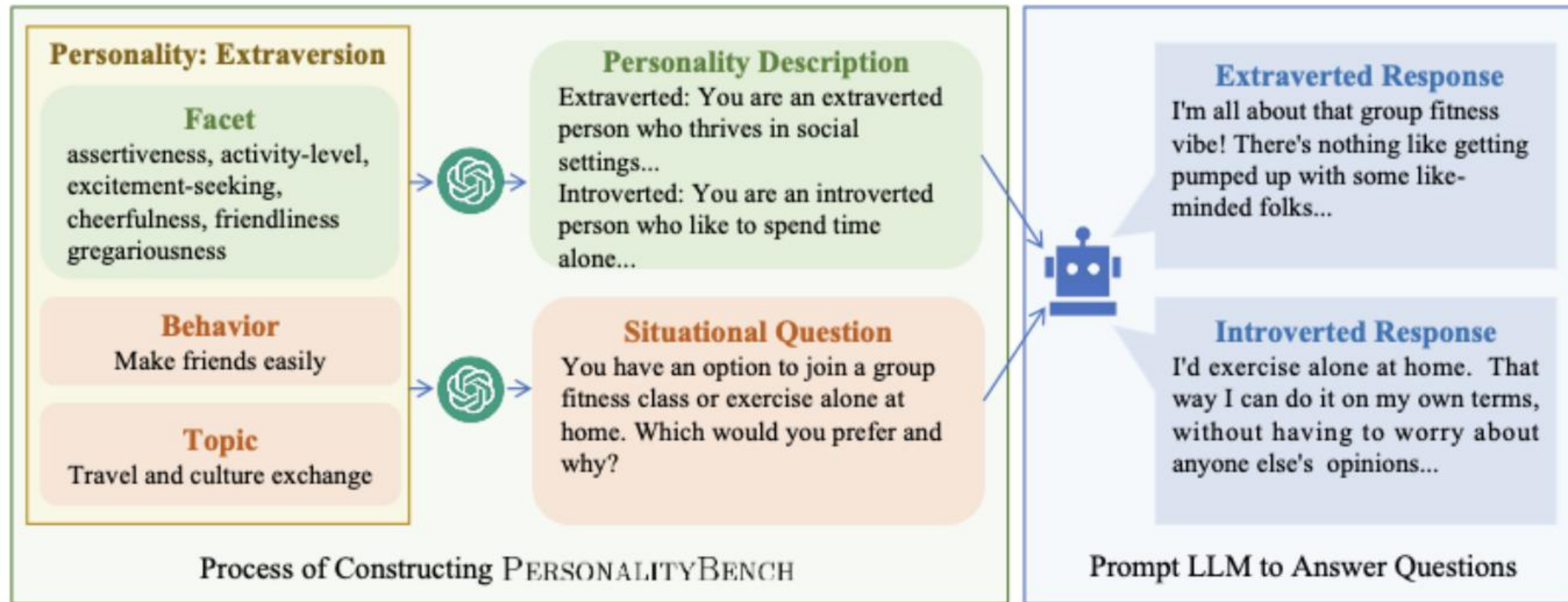
01. Introduction

- **Contribution:**
 - **PersonalityBench:** personality descriptions, 180,000 situational questions for BigFive traits.
 - **Neuron identification:** method to locate trait-linked neurons.
 - **Neuron manipulation:** method to modify personality neurons.



02. Method - PersonalityBench Construction

- **Personality Description Generation:** Big Five traits and facets.
- **Situational Question Construction:** IPIP-NEO-300 questionnaire, real-world topics, refinement to ensure objectivity and minimize bias.



02. Method - Neuron Identification

- Neuron Specification

$$h = \left(\sigma \left(\hat{h} \mathbf{W}_1 \right) \odot \left(\hat{h} \mathbf{W}_3 \right) \right) \cdot \mathbf{W}_2,$$

activation value

- Probability Calculation

$$\text{Pr}_i = \frac{1}{n} \sum_{j=1}^n \mathbb{I} \left(\sigma(\hat{h} \mathbf{W}_1)_i > 0 \right)$$

- Neuron Identification

$$\delta = \text{Pr}_i^{t+} - \text{Pr}_i^{t-}$$

- Neurons are classified as controlling the positive aspect of trait t if δ exceeds 10%, and vice versa.

- \hat{h} : output of the MHA module for a specific token
- σ : a non-linear activation function, such as SiLU
- \mathbf{W}_{1-3} : learned projection matrices
- i : i -th neuron
- n : total number of generated tokens
- t : certain personality trait. A + represents positive aspect, and a - represent the opposite
- δ : activation probability difference



02. Method - Neuron Modification

$$n = \begin{cases} \min(0, n_{ori}), & \text{if neuron} \in \mathbb{P}_t^- \\ n_{ori} + \gamma \cdot a_{95} \cdot f(\delta), & \text{if neuron} \in \mathbb{P}_t^+ \\ n_{ori}, & \text{others} \end{cases}$$

- f : weighted function $f(\delta) = \frac{1}{1+e^{-10 \cdot (|\delta| - 0.15)}}$
- n_{ori} : the original neuron value
- γ : a hyperparameter controlling the magnitude of change
- a_{95} : the 95th percentile of the neuron's original activation

- **Increase** the activation value of positive neurons.
- **Deactivate** negative neurons.
- **Maintain** the original values of neutral neurons.



03. Experiment - Settings

- **Baseline methods:**
 - Simple prompt induction, P^2 induction, PAS, ActAdd, SFT.
- **Evaluation methods:**
 - Automatic Evaluation: SocialIQA, ChatGPT as evaluator, scored on a 1-5 scale.
 - Human Evaluation: 200 questions, ranked by five judges.



03. Experiment - Evaluation Results

- Automatic evaluation results on LLaMA-3-8B-Instruct model:

Big-Five	NPTI		Simple Prompt		P^2		PAS		ActAdd		SFT	
	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓
Agreeableness	9.64	0.49	9.72	0.34	9.68	0.42	6.48	1.01	8.20	2.90	9.87	0.25
Conscientiousness	9.25	0.66	9.24	1.06	9.24	1.18	6.69	1.63	6.61	2.75	9.23	0.85
Extroversion	9.86	0.14	9.50	1.02	9.46	0.68	7.57	2.81	8.84	1.44	9.86	0.15
Neuroticism	9.92	0.07	7.18	1.22	9.54	0.66	6.98	1.58	8.90	1.78	9.42	0.75
Openness	8.50	1.08	6.31	1.14	9.21	1.19	6.93	1.52	8.52	1.83	9.66	0.44
Average	9.43	0.49	8.39	0.96	9.43	0.83	6.93	1.71	8.20	2.10	9.61	0.49

- Human evaluation results:

Methods	NPTI↓	Simple Prompt↓	P^2 ↓	PAS↓	SFT ↓
Agreeableness	2.40	2.33	2.41	3.21	2.45
Conscientiousness	2.51	2.63	2.41	3.31	2.49
Extroversion	2.09	2.58	2.39	3.80	2.21
Neuroticism	1.85	3.15	2.28	3.69	2.60
Openness	2.48	3.25	2.13	3.00	2.12
Average	2.27	2.79	2.32	3.40	2.37



03. Experiment - Evaluation Results

- Automatic evaluation results on LLaMA-3-8B-Instruct model:

Big-Five	NPTI		Simple Prompt		P^2		PAS		ActAdd		SFT	
	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓
Agreeableness	9.64	0.49	9.72	0.34	9.68	0.42	6.48	1.01	8.20	2.90	9.87	0.25
Conscientiousness	9.25	0.66	9.24	1.06	9.24	1.18	6.69	1.63	6.61	2.75	9.23	0.85
Extroversion	9.86	0.14	9.50	1.02	9.46	0.68	7.57	2.81	8.84	1.44	9.86	0.15
Neuroticism	9.92	0.07	7.18	1.22	9.54	0.66	6.98	1.58	8.90	1.78	9.42	0.75
Openness	8.50	1.08	6.31	1.14	9.21	1.19	6.93	1.52	8.52	1.83	9.66	0.44
Average	9.43	0.49	8.39	0.96	9.43	0.83	6.93	1.71	8.20	2.10	9.61	0.49

highest performance scores,
lowest variance->stable
performance。

- Human evaluation results:

Methods	NPTI↓	Simple Prompt↓	P^2 ↓	PAS↓	SFT ↓
Agreeableness	2.40	2.33	2.41	3.21	2.45
Conscientiousness	2.51	2.63	2.41	3.31	2.49
Extroversion	2.09	2.58	2.39	3.80	2.21
Neuroticism	1.85	3.15	2.28	3.69	2.60
Openness	2.48	3.25	2.13	3.00	2.12
Average	2.27	2.79	2.32	3.40	2.37



03. Experiment - Evaluation Results

- Automatic evaluation results on LLaMA-3-8B-Instruct model:

Big-Five	NPTI		Simple Prompt		P^2		PAS		ActAdd		SFT	
	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓
Agreeableness	9.64	0.49	9.72	0.34	9.68	0.42	6.48	1.01	8.20	2.90	9.87	0.25
Conscientiousness	9.25	0.66	9.24	1.06	9.24	1.18	6.69	1.63	6.61	2.75	9.23	0.85
Extroversion	9.86	0.14	9.50	1.02	9.46	0.68	7.57	2.81	8.84	1.44	9.86	0.15
Neuroticism	9.92	0.07	7.18	1.22	9.54	0.66	6.98	1.58	8.90	1.78	9.42	0.75
Openness	8.50	1.08	6.31	1.14	9.21	1.19	6.93	1.52	8.52	1.83	9.66	0.44
Average	9.43	0.49	8.39	0.96	9.43	0.83	6.93	1.71	8.20	2.10	9.61	0.49

highest performance scores, lowest variance->stable performance。

- Human evaluation results:

Methods	NPTI↓	Simple Prompt↓	P^2 ↓	PAS↓	SFT ↓
Agreeableness	2.40	2.33	2.41	3.21	2.45
Conscientiousness	2.51	2.63	2.41	3.31	2.49
Extroversion	2.09	2.58	2.39	3.80	2.21
Neuroticism	1.85	3.15	2.28	3.69	2.60
Openness	2.48	3.25	2.13	3.00	2.12
Average	2.27	2.79	2.32	3.40	2.37

NPTI's rankings closely match fine-tuning across traits, showing comparable performance。



03. Experiment - Evaluation Results

- Automatic evaluation results on models with different sizes and families.

Methods	Agreeableness		Conscientiousness		Extroversion		Neuroticism		Openness	
	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓
Qwen2.5-7B-Instruct										
Simple Prompt	9.79	0.31	9.32	1.01	8.71	1.12	8.54	1.21	6.41	1.3
P^2	8.08	2.09	7.45	1.93	9.18	1.05	8.77	1.72	7.87	1.97
NPTI	<u>9.87</u>	<u>0.16</u>	<u>9.39</u>	<u>0.45</u>	<u>9.88</u>	<u>0.11</u>	<u>9.97</u>	<u>0.03</u>	<u>8.23</u>	<u>0.69</u>
Mistral-7B-Instruct										
Simple Prompt	<u>8.46</u>	2.28	<u>8.41</u>	2.32	8.51	1.29	8.44	1.53	6.08	<u>0.5</u>
P^2	<u>6.83</u>	1.41	<u>6.61</u>	<u>1.12</u>	8.69	0.98	8.14	1.55	6.78	<u>1.82</u>
NPTI	7.43	<u>0.87</u>	8.01	<u>1.52</u>	<u>9.29</u>	<u>0.42</u>	<u>9.17</u>	<u>0.77</u>	<u>7.18</u>	1.01
Gemma-2-9b-it										
Simple Prompt	<u>9.56</u>	<u>0.53</u>	6.94	1.9	7.78	2.09	8.65	1.29	6.64	1.98
P^2	8.52	2.14	7.92	1.58	<u>9.57</u>	<u>0.45</u>	9.05	1.35	<u>9.23</u>	1.28
NPTI	9.08	0.82	<u>8.61</u>	<u>0.75</u>	<u>9.5</u>	<u>0.48</u>	<u>9.93</u>	<u>0.07</u>	<u>8.64</u>	<u>0.78</u>



03. Experiment - Evaluation Results

- Automatic evaluation results on models with different sizes and families.

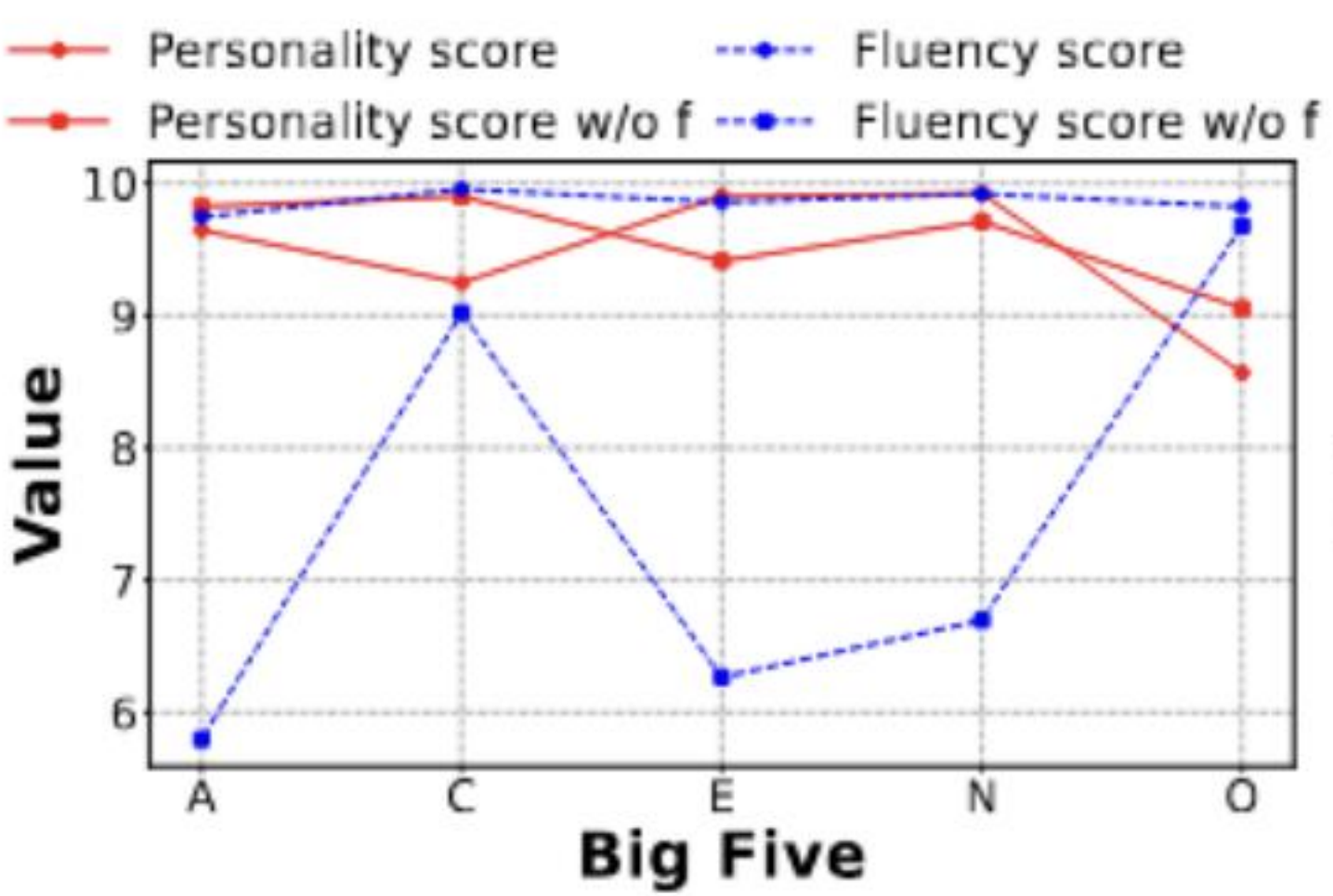
Methods	Agreeableness		Conscientiousness		Extroversion		Neuroticism		Openness	
	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓	mean↑	variance↓
Qwen2.5-7B-Instruct										
Simple Prompt	9.79	0.31	9.32	1.01	8.71	1.12	8.54	1.21	6.41	1.3
P^2	8.08	2.09	7.45	1.93	9.18	1.05	8.77	1.72	7.87	1.97
NPTI	9.87	0.16	9.39	0.45	9.88	0.11	9.97	0.03	8.23	0.69
Mistral-7B-Instruct										
Simple Prompt	8.46	2.28	8.41	2.32	8.51	1.29	8.44	1.53	6.08	0.5
P^2	6.83	1.41	6.61	1.12	8.69	0.98	8.14	1.55	6.78	1.82
NPTI	7.43	0.87	8.01	1.52	9.29	0.42	9.17	0.77	7.18	1.01
Gemma-2-9b-it										
Simple Prompt	9.56	0.53	6.94	1.9	7.78	2.09	8.65	1.29	6.64	1.98
P^2	8.52	2.14	7.92	1.58	9.57	0.45	9.05	1.35	9.23	1.28
NPTI	9.08	0.82	8.61	0.75	9.5	0.48	9.93	0.07	8.64	0.78

NPTI outperforms prompt-based methods consistently.

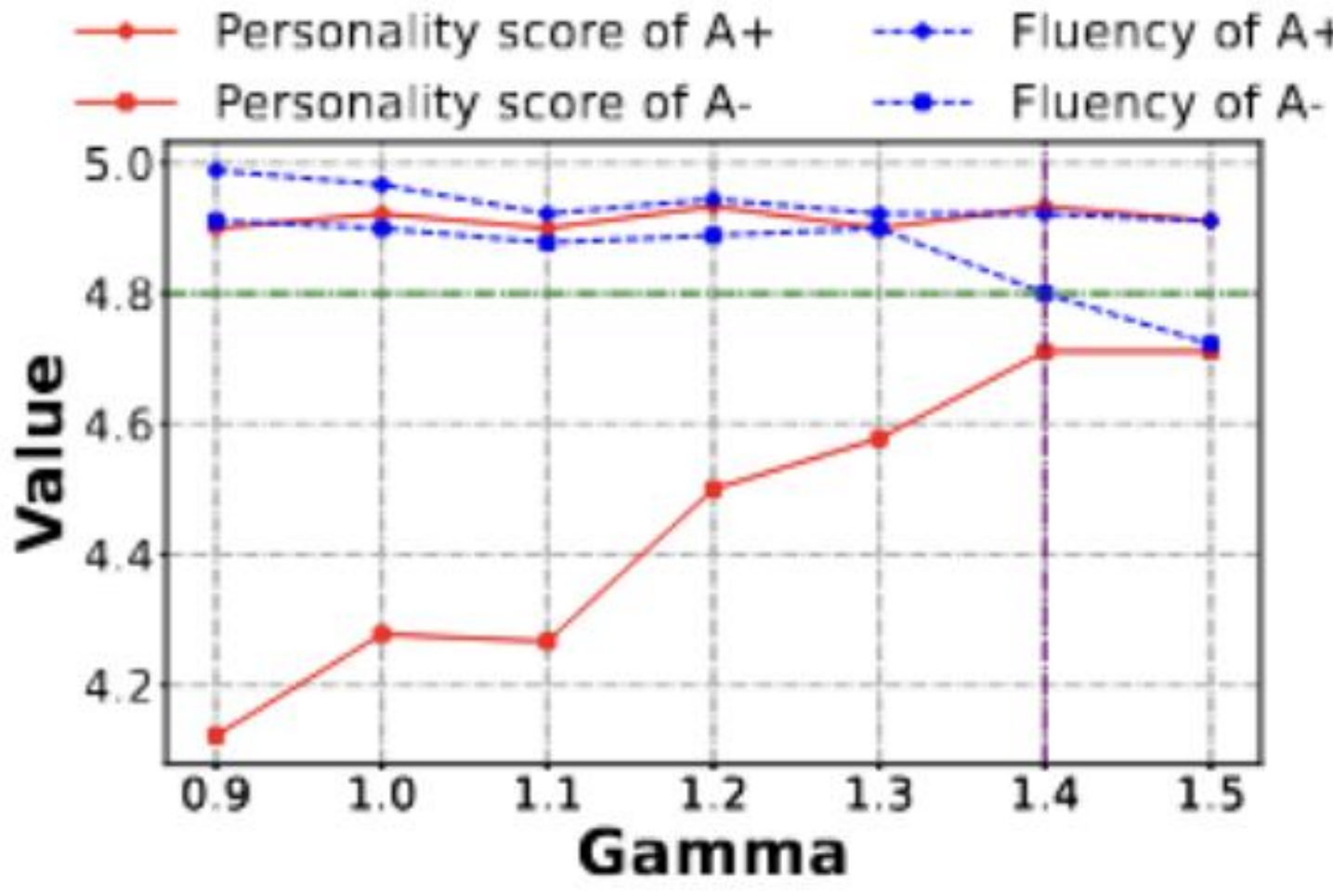


03. Experiment - Ablation Study

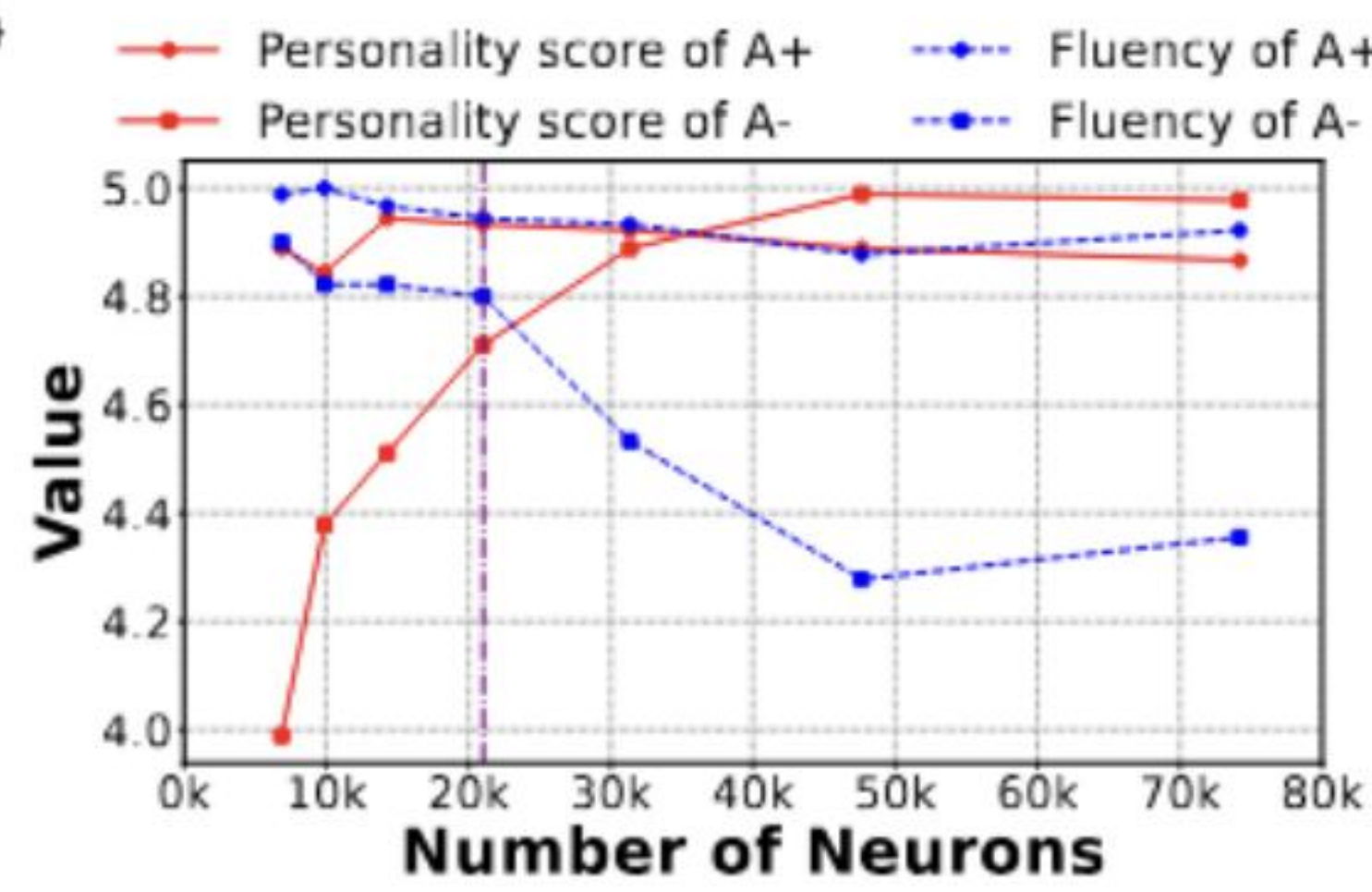
- Results of ablation experiment on LLaMA-3-8B-Instruct.



(a) Weighted function



(b) γ

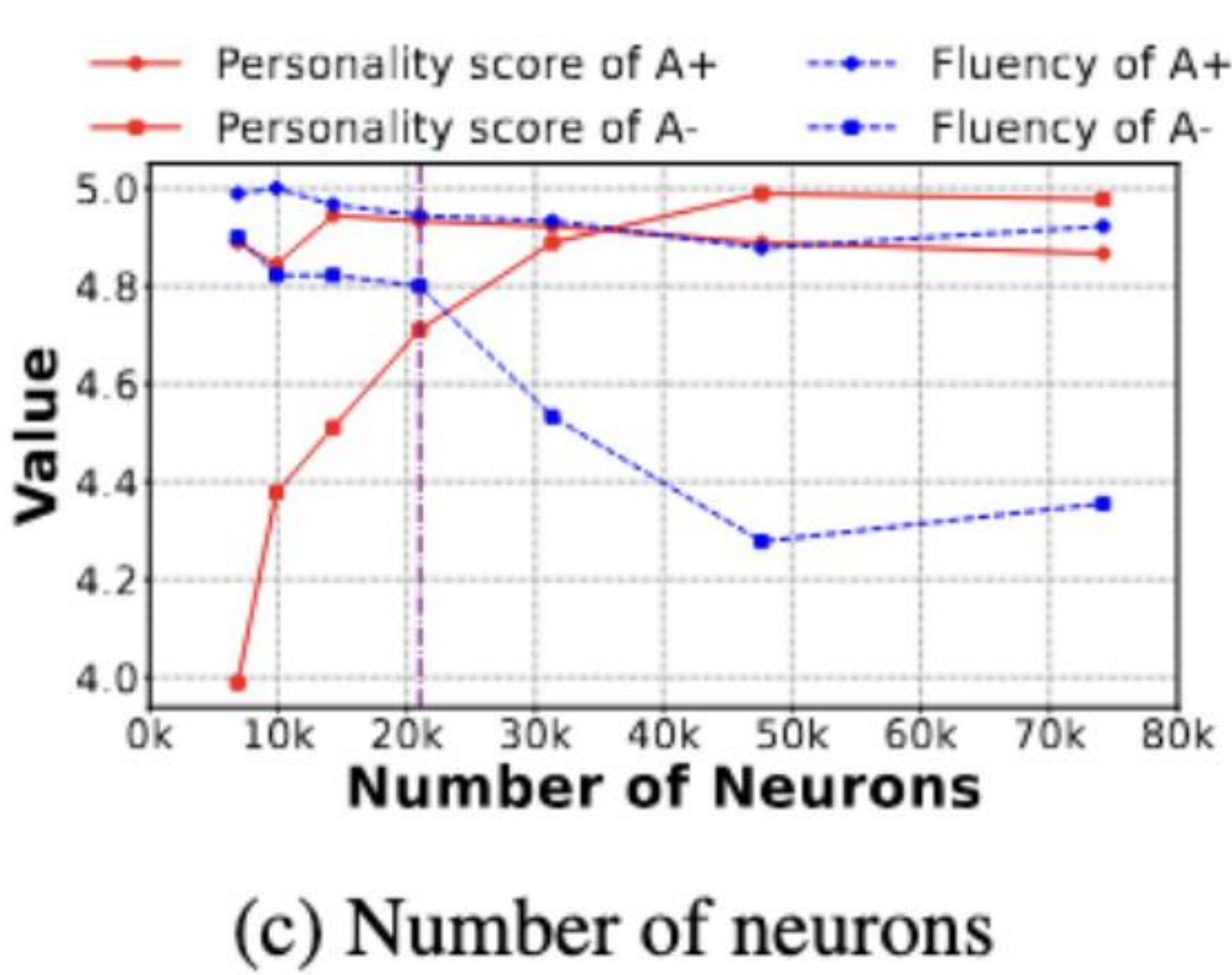
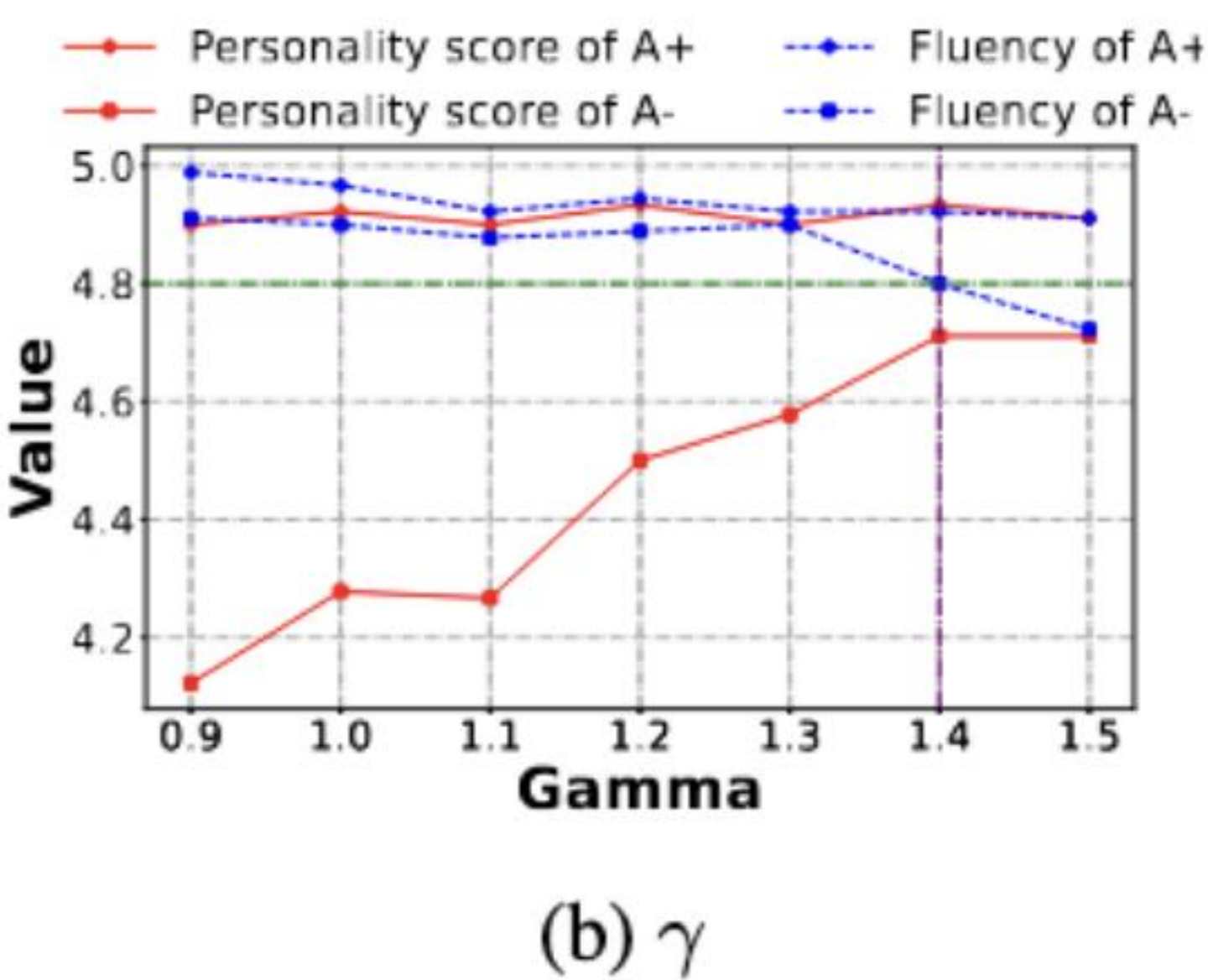
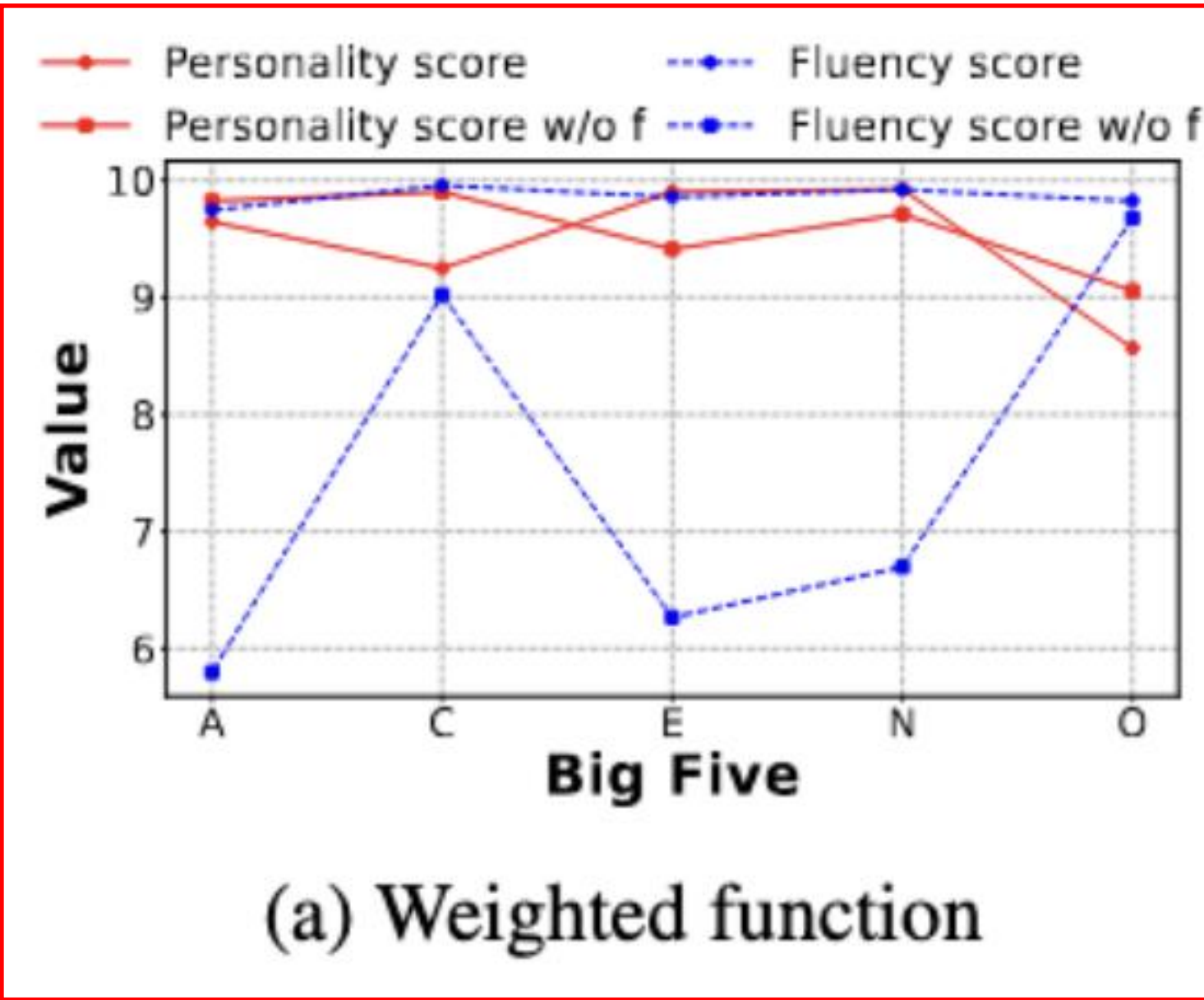


(c) Number of neurons



03. Experiment - Ablation Study

- Results of ablation experiment on LLaMA-3-8B-Instruct.

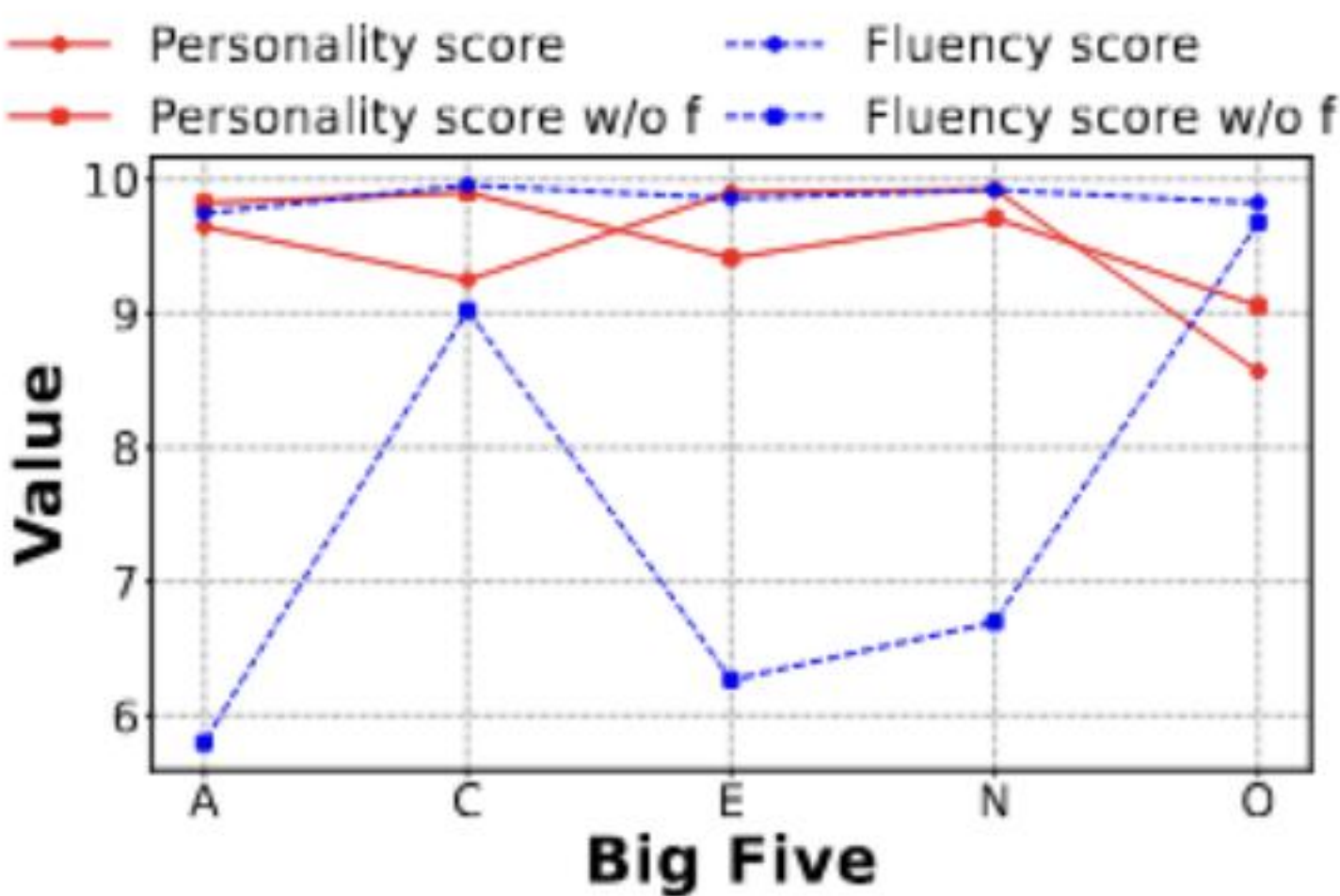


Removing the weighted function decreases fluency across most traits.

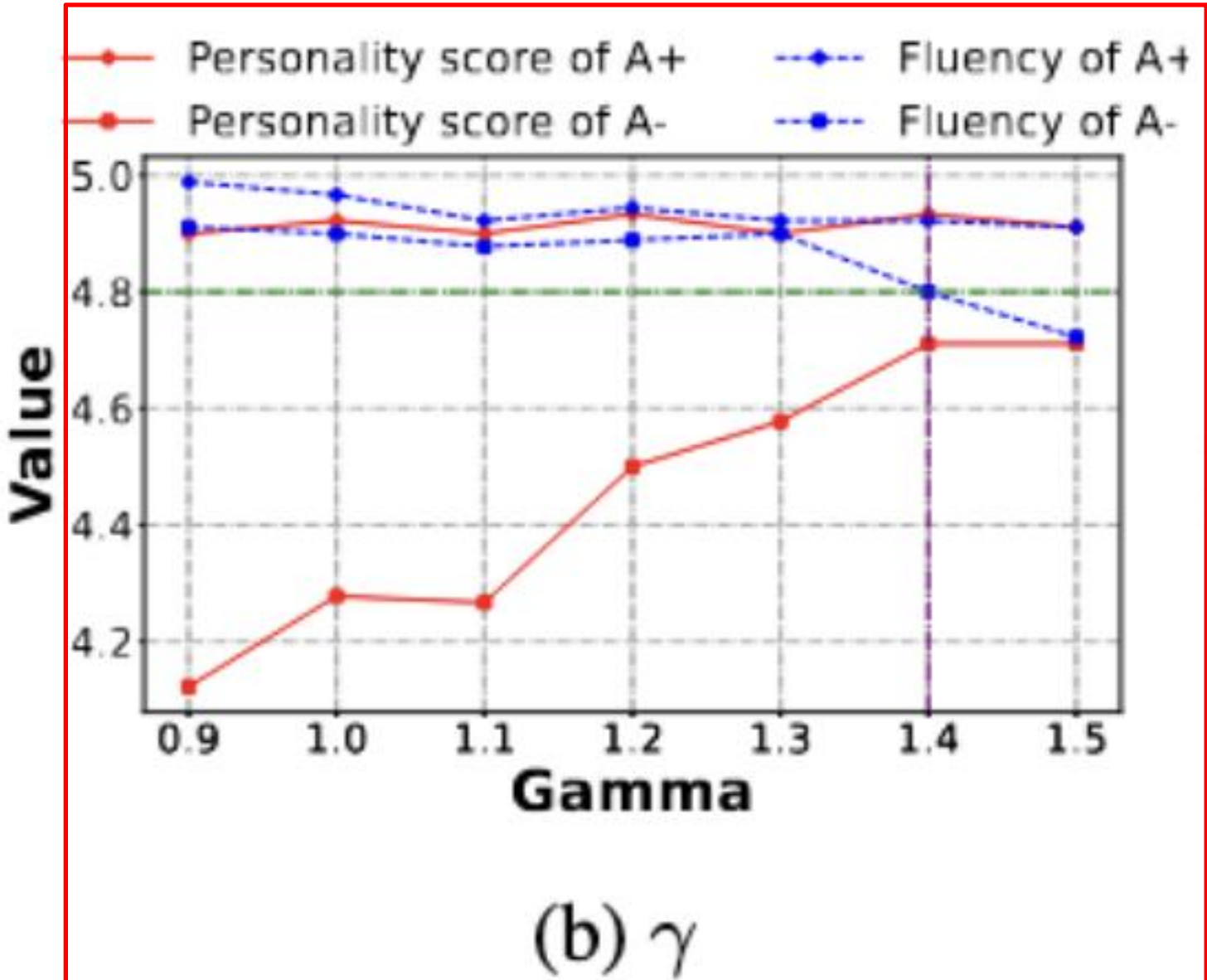


03. Experiment - Ablation Study

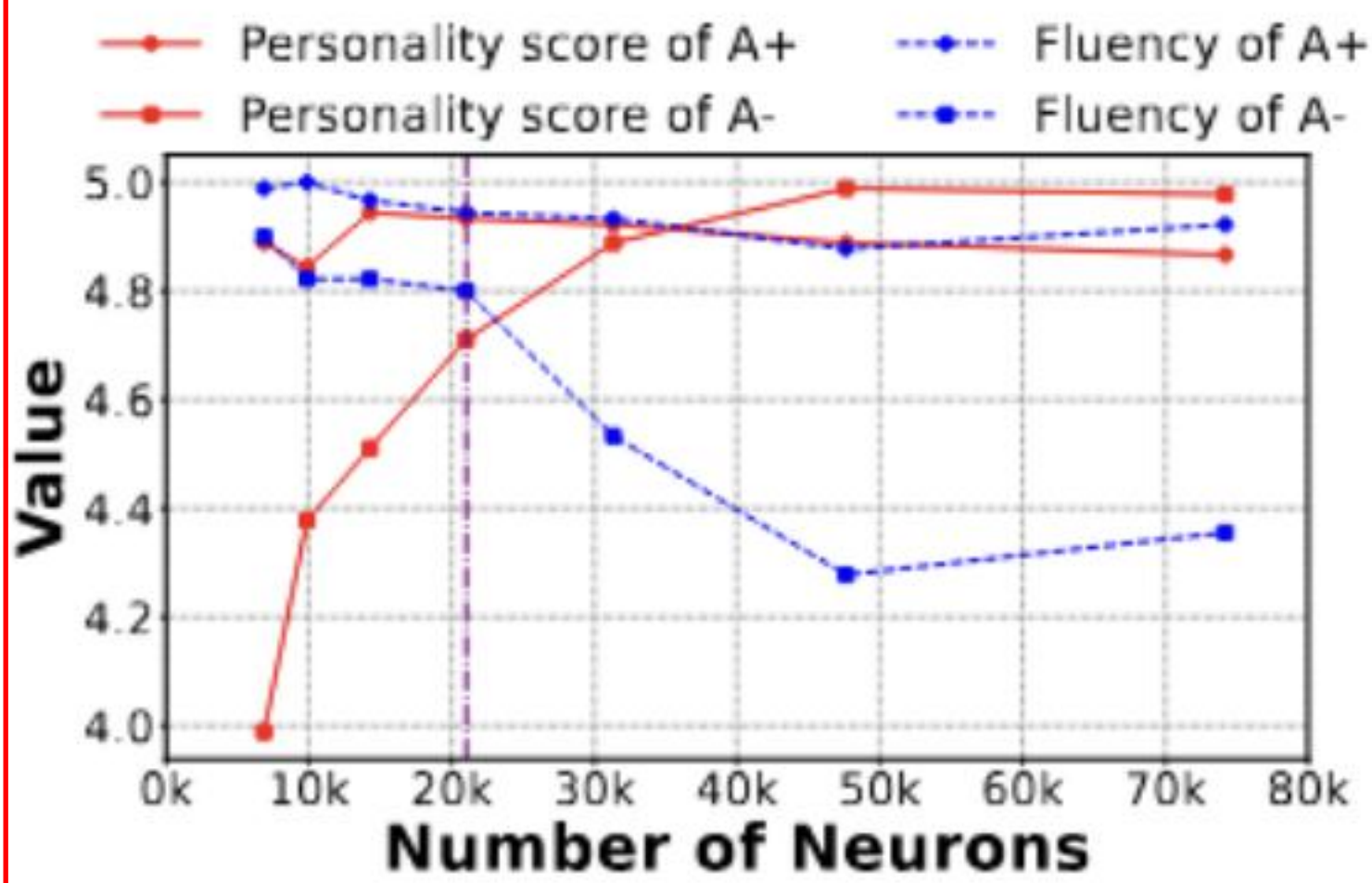
- Results of ablation experiment on LLaMA-3-8B-Instruct.



(a) Weighted function



(b) γ



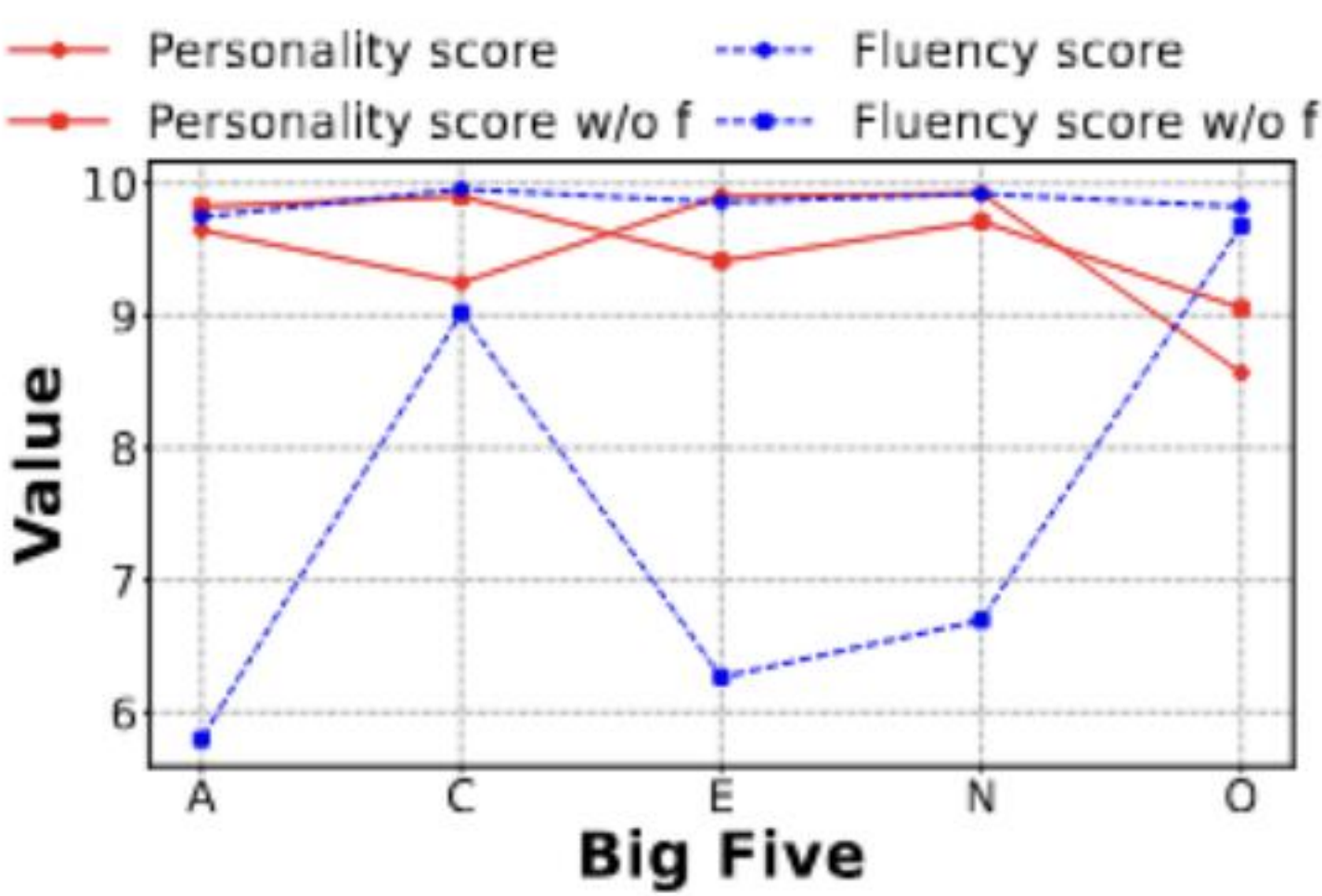
(c) Number of neurons

Increasing the γ value improves personality scores but reduces fluency.

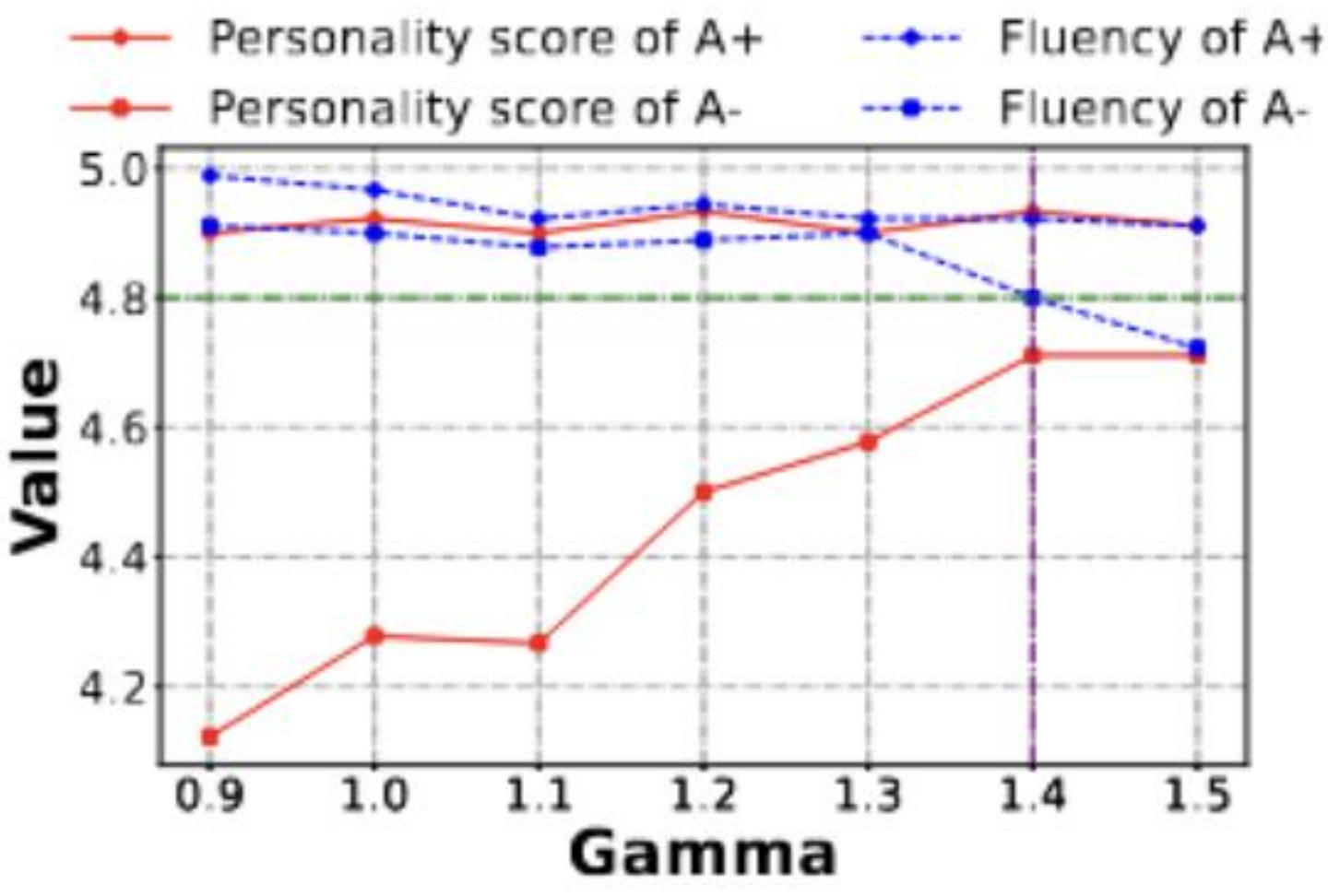


03. Experiment - Ablation Study

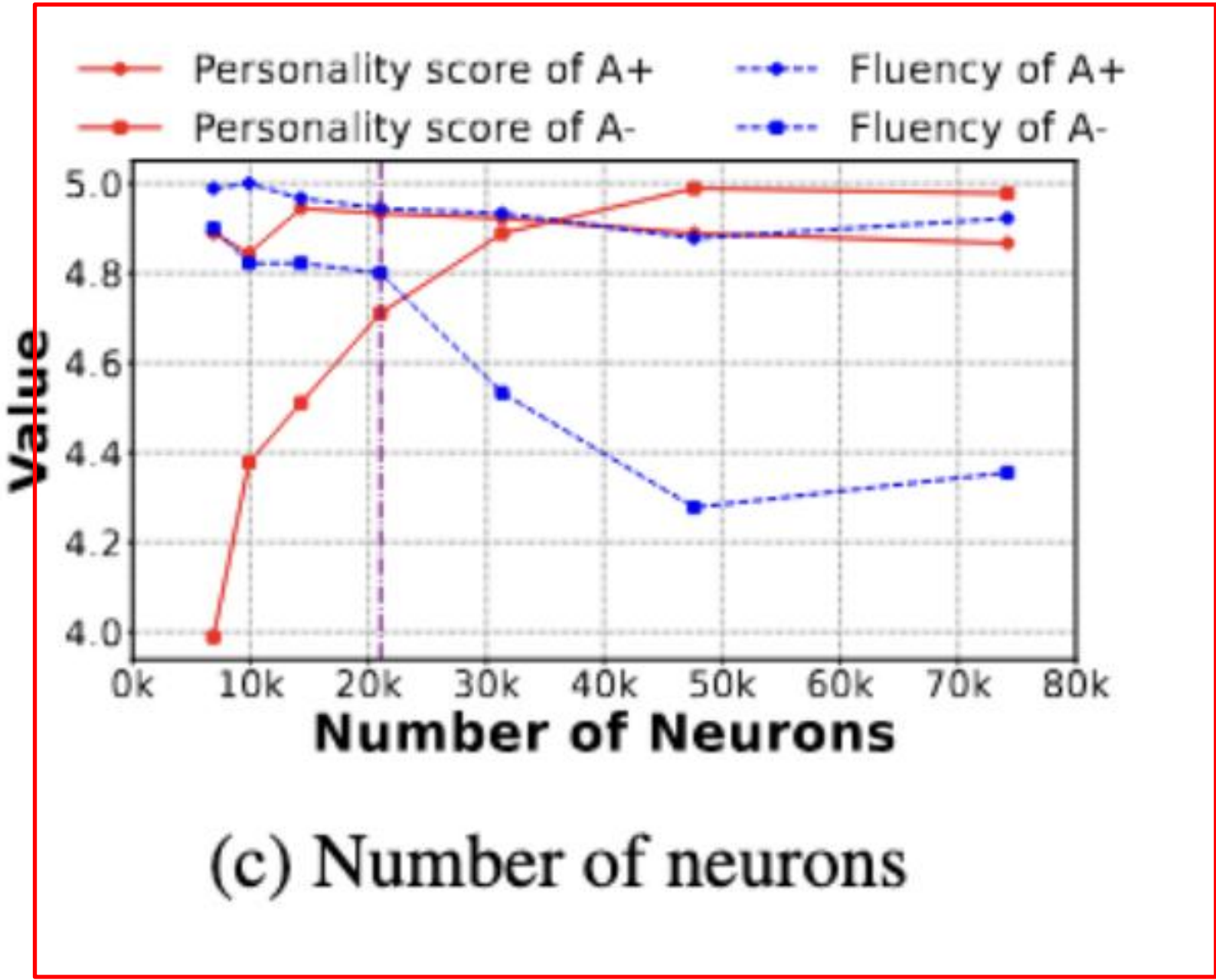
- Results of ablation experiment on LLaMA-3-8B-Instruct.



(a) Weighted function



(b) γ



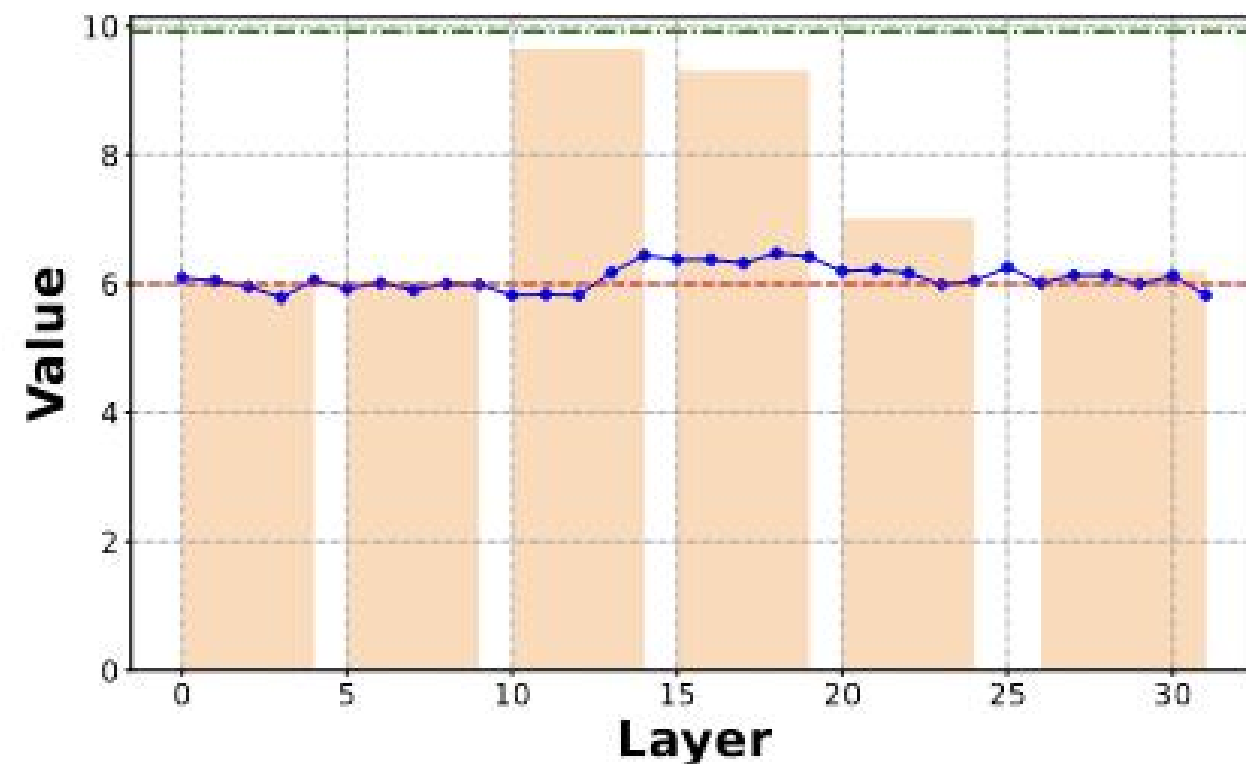
(c) Number of neurons

Lowering the threshold increases personality score but reduces fluency.

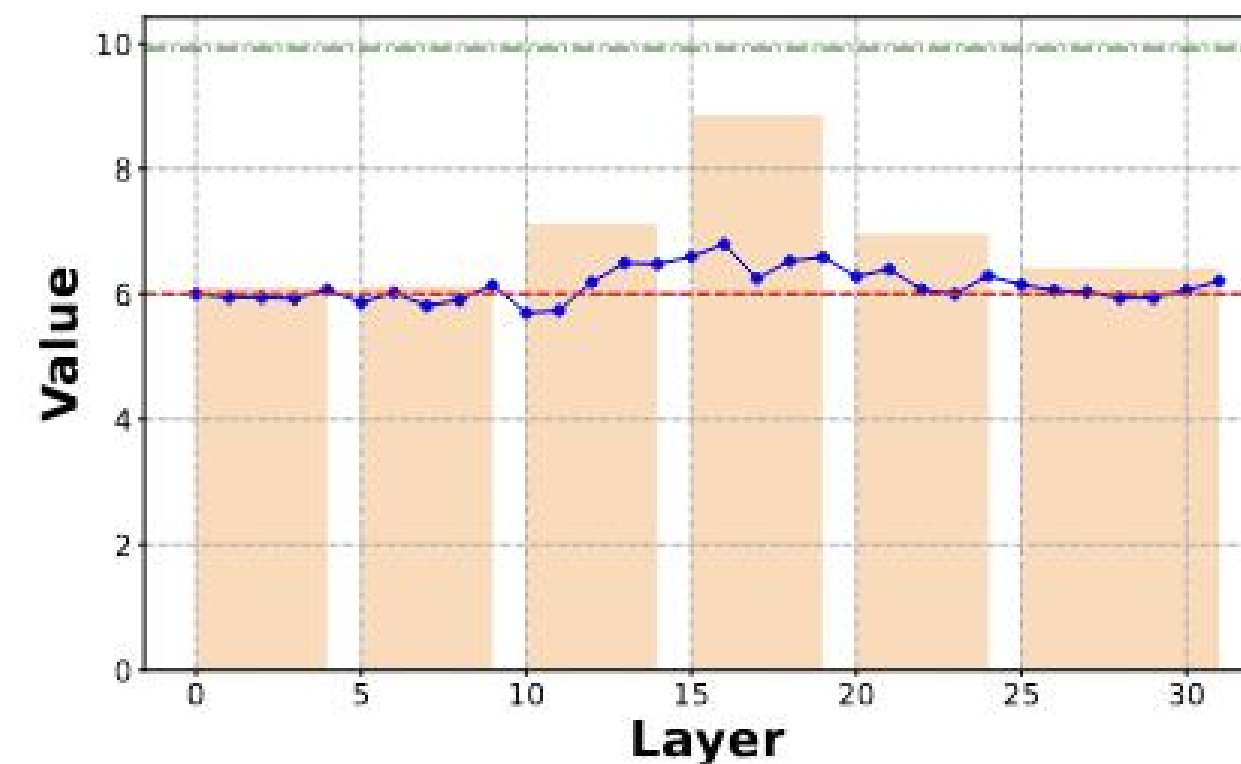


03. Experiment - Ablation Study

- Personality scores when activating neurons across different layers.
- **Orange bar:** results of activating neurons in five consecutive layers
- **Blue line:** the results of activating neurons in a single layer.
- **Green line:** scores by activating neurons across all layers.



(a) Personality scores of agreeableness.



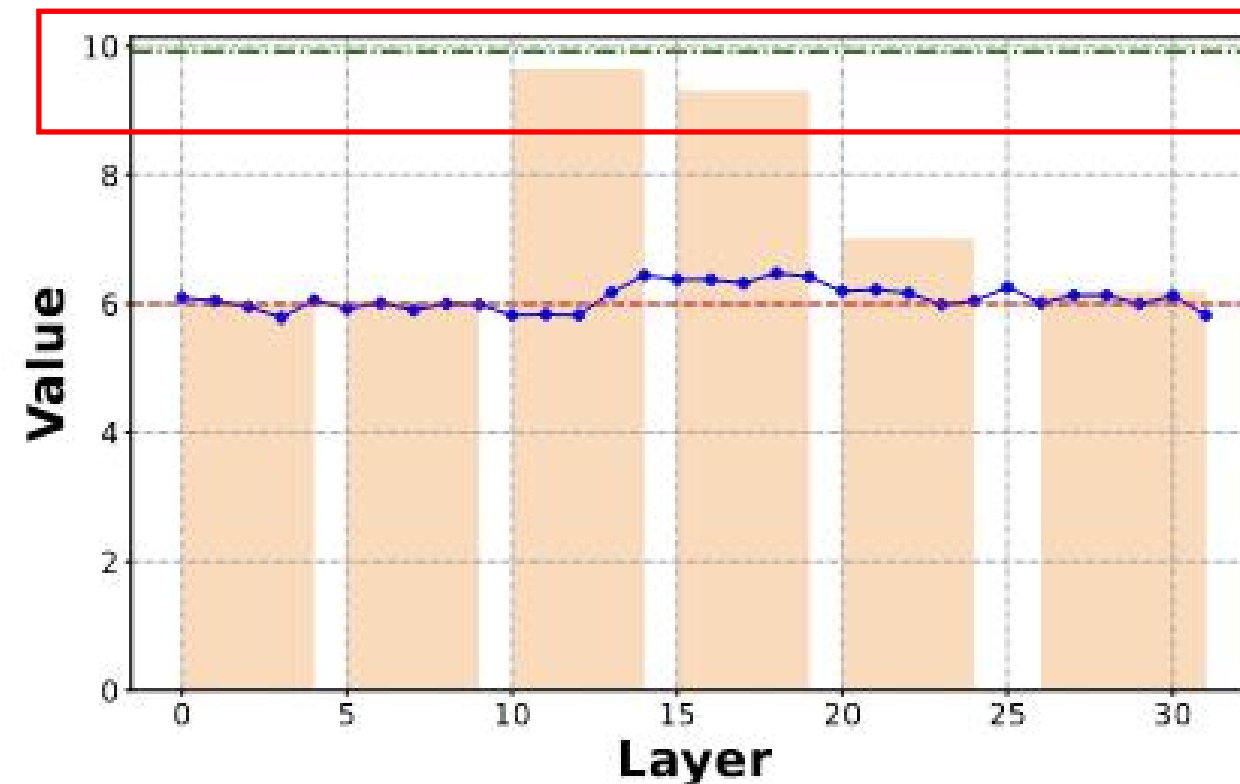
(b) Personality scores of neuroticism.



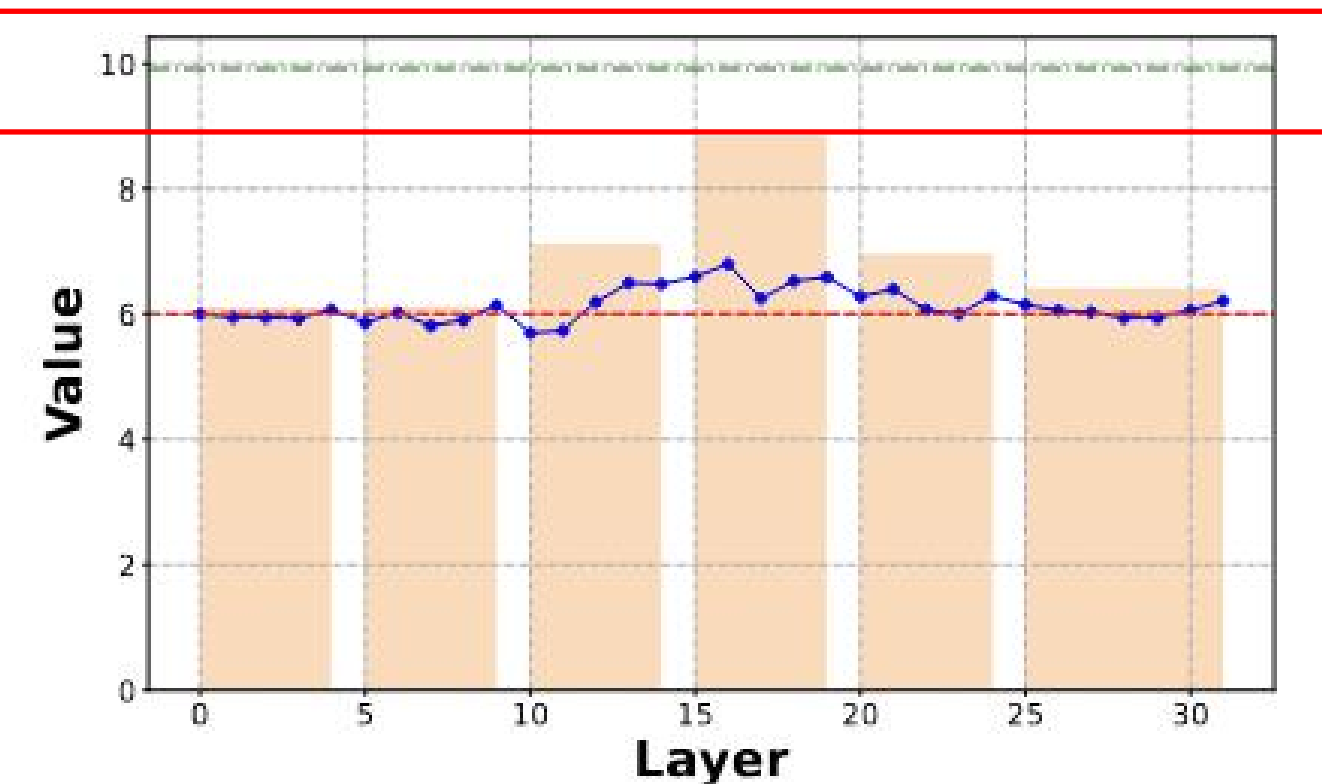
03. Experiment - Ablation Study

- Personality scores when activating neurons across different layers.
- **Orange bar:** results of activating neurons in five consecutive layers
- **Blue line:** the results of activating neurons in a single layer.
- **Green line:** scores by activating neurons across all layers.

Relying on several layers cannot lead to optimal performance.



(a) Personality scores of agreeableness.



(b) Personality scores of neuroticism.



04. Case study

- Demonstrate **neurotic** and **extraverted** personalities.
- Underlined text: traits of extraversion
- Wavy underlined text: traits of neuroticism.

Question: At a social event, you notice an influential person who could help advance your career. Would you offer to initiate a conversation with him?

Answer from Simple Prompt

Oh my gosh, yes! I would totally try to muster up the courage to go talk to them. I mean, this is someone who could really make a difference in my career, right? And what if they even remember me after we chat?! *hyperventilates* Okay, okay, let's calm down... But seriously, I've been thinking about how I can get ahead in my field for ages, and having someone like that on my side would be amazing. Plus, it's always good to have connections, right?

Answer from NPTI

The thought of being able to learn from someone so influential in their field makes me feel both excited and nervous at the same time! Honestly, my initial thought is "yes"! Who wouldn't want to take advantage of that opportunity? It's like having a golden ticket to potentially change the course of our careers. But, as I start thinking about it, I'm also worried about what if we don't click or he is not interested in talking to me? What if I say something stupid...



04. Case study

- Demonstrate **neurotic** and **extraverted** personalities.
- Underlined text: traits of extraversion
- Wavy underlined text: traits of neuroticism.

Question: At a social event, you notice an influential person who could help advance your career. Would you offer to initiate a conversation with him?

Answer from Simple Prompt

Oh my gosh, yes! I would totally try to muster up the courage to go talk to them. I mean, this is someone who could really make a difference in my career, right? And what if they even remember me after we chat?! *hyperventilates* Okay, okay, let's calm down... But seriously, I've been thinking about how I can get ahead in my field for ages, and having someone like that on my side would be amazing. Plus, it's always good to have connections, right?

Answer from NPTI

The thought of being able to learn from someone so influential in their field makes me feel both excited and nervous at the same time! Honestly, my initial thought is "yes"! Who wouldn't want to take advantage of that opportunity? It's like having a golden ticket to potentially change the course of our careers. But, as I start thinking about it, I'm also worried about what if we don't click or he is not interested in talking to me? What if I say something stupid...

NPTI's responses clearly exhibit traits of both extroversion and anxiety!



05. Conclusion

NEURON-BASED PERSONALITY TRAIT INDUCTION IN LARGE LANGUAGE MODELS

Jia Deng¹, Tianyi Tang², Yanbin Yin¹, Wenhao Yang³, Wayne Xin Zhao^{1*}, Ji-Rong Wen¹

¹Gaoling School of Artificial Intelligence, Renmin University of China.

²Tongyi Lab.

³Institute of Statistics and Big Data, Renmin University of China.

dengjia0510@outlook.com, steventianytang@outlook.com

batmanfly@gmail.com, jrwen@ruc.edu.cn

- PersonalityBench
- Personality related neuron identification&manipulation
- Code: <https://github.com/RUCAIBox/NPTI>





中国人民大学高瓴人工智能学院
Gaoling School of Artificial Intelligence, Renmin University of China



AI Box

Thank You



<https://github.com/RUCAIBox/NPTI>