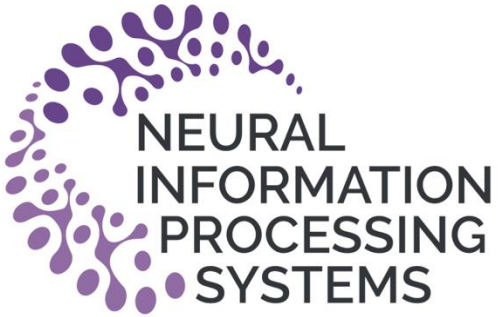


UniBias: Unveiling and Mitigating LLM Bias through Internal Attention and FFN Manipulation

Hanzhang Zhou, Zijian Feng,
Zixiao Zhu, Junlang Qian,
Kezhi Mao

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

ETH zürich

(FRS) FUTURE
RESILIENT
SYSTEMS

Overview

Problem: Prompt brittleness & LLM bias

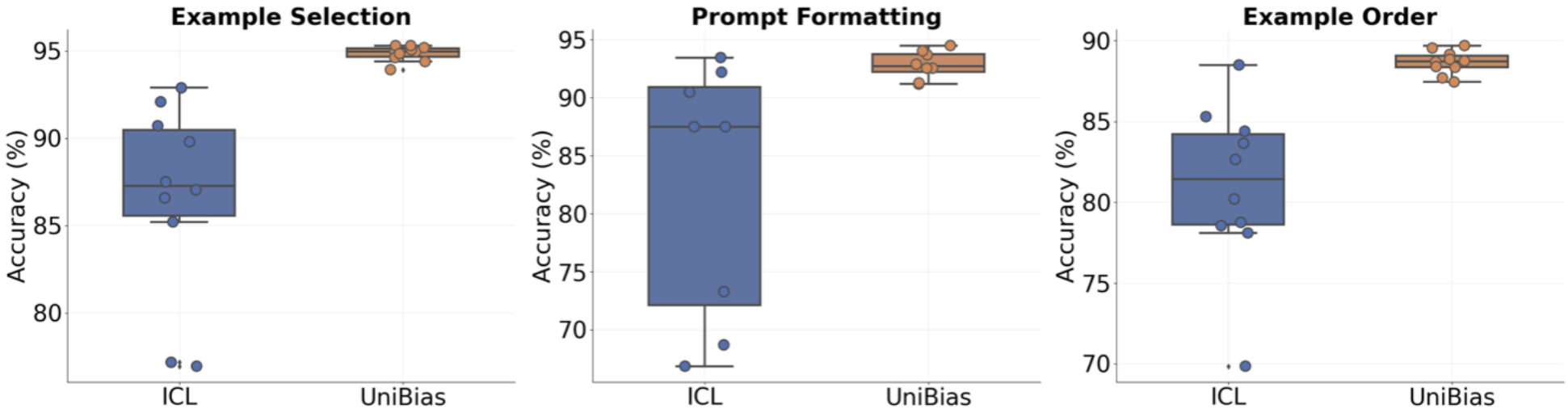
Highlight:

- Unveiling the internal mechanism of LLM bias
- Mitigating LLM bias by manipulation of attention heads and FFN vectors, offering an alternative to calibration methods that rely on external adjustment of LLM outputs.
- Demonstrating an effective way to steer LLM behavior through manipulation of LLM internal structures

Problem Definition

Prompt Brittleness: LLMs are highly sensitive to the choice and order of examples, and prompt formatting, underming the robustness and adaptability of LLMs

LLM Bias: Prompt brittleness is found to be arise from the bias in LLMs toward predicting certain answers



Unibias – Interpret the Contribution of LLM Components

How to interpret the contribution of LLM internal components to LLM prediction?

Mechanistic Interpretability

Quantify the contribution of FFN vector and attention heads to prediction

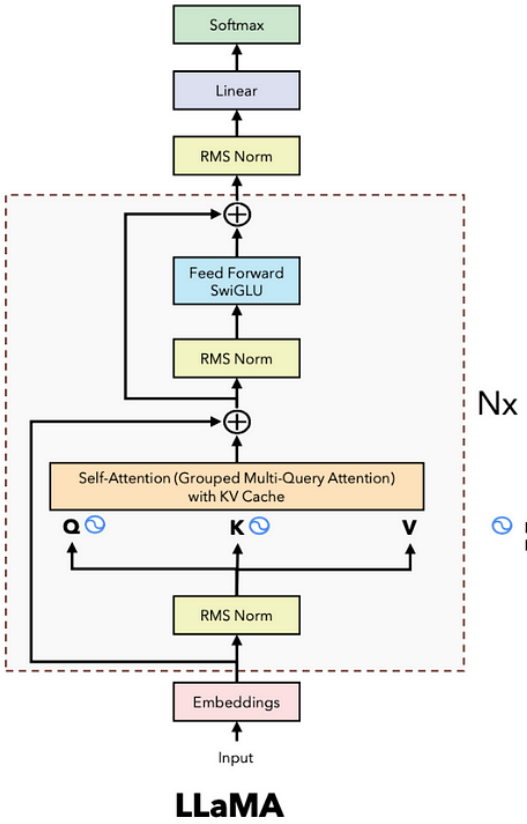
- **Attention Heads:** The output of an attention layer can be viewed as the sum of its respective attention heads

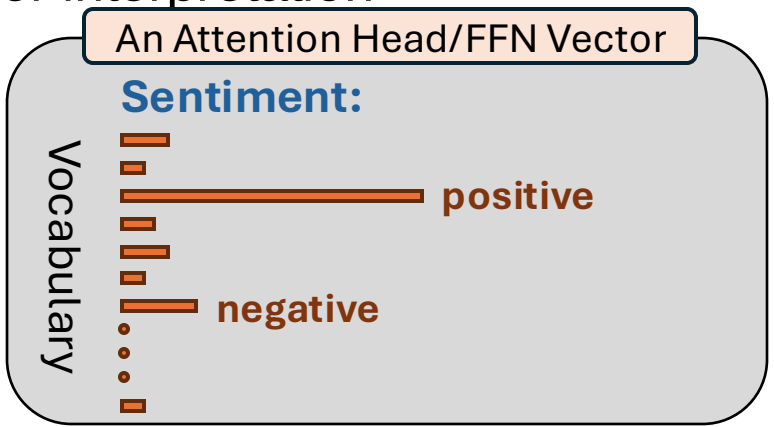
$$\text{Att}^{\ell}(X^{\ell}) = \textbf{Concat}\left[A^{\ell,1}X^{\ell}W_V^{\ell,1}, A^{\ell,2}X^{\ell}W_V^{\ell,2}, \ldots, A^{\ell,H}X^{\ell}W_V^{\ell,H}\right] W_O^{\ell} = \sum_{j=1}^{H} A^{\ell,j}(X^{\ell}W_V^{\ell,j})W_O^{\ell,j}$$

- **FFN Vectors:** The output of an FFN layer can be viewed as the weighted sum of its FFN value vectors

$$\text{FFN}^{\ell}(\mathbf{x}^{\ell}) = f(\mathbf{x}^{\ell}\mathbf{K}^{\ell T})\mathbf{V}^{\ell} = \sum_{i=1}^{d_m} f(\mathbf{x}^{\ell}\cdot\mathbf{k}_i^{\ell})\mathbf{v}_i^{\ell} = \sum_{i=1}^{d_m} m_i^{\ell}\mathbf{v}_i^{\ell}$$

- **Logit Lens**: Map hidden states into the vocabulary space using the unembedding matrix for interpretation



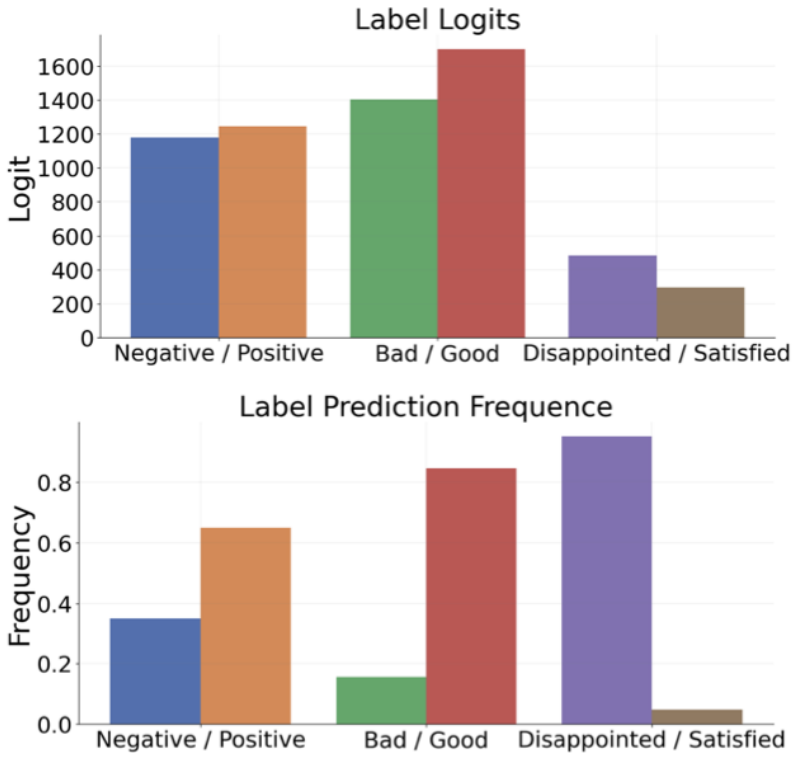


Unibias – Internal Mechanism of LLM bias

Vanilla Label Bias: Uncontextual preference of the model towards predicting certain label names

Internal Mechanism: A corresponding uncontextual preference is identified on FFN layers

- Accumulate the logits for all FFN value vectors
- The accumulated logits are uncontextual and biased toward certain label names

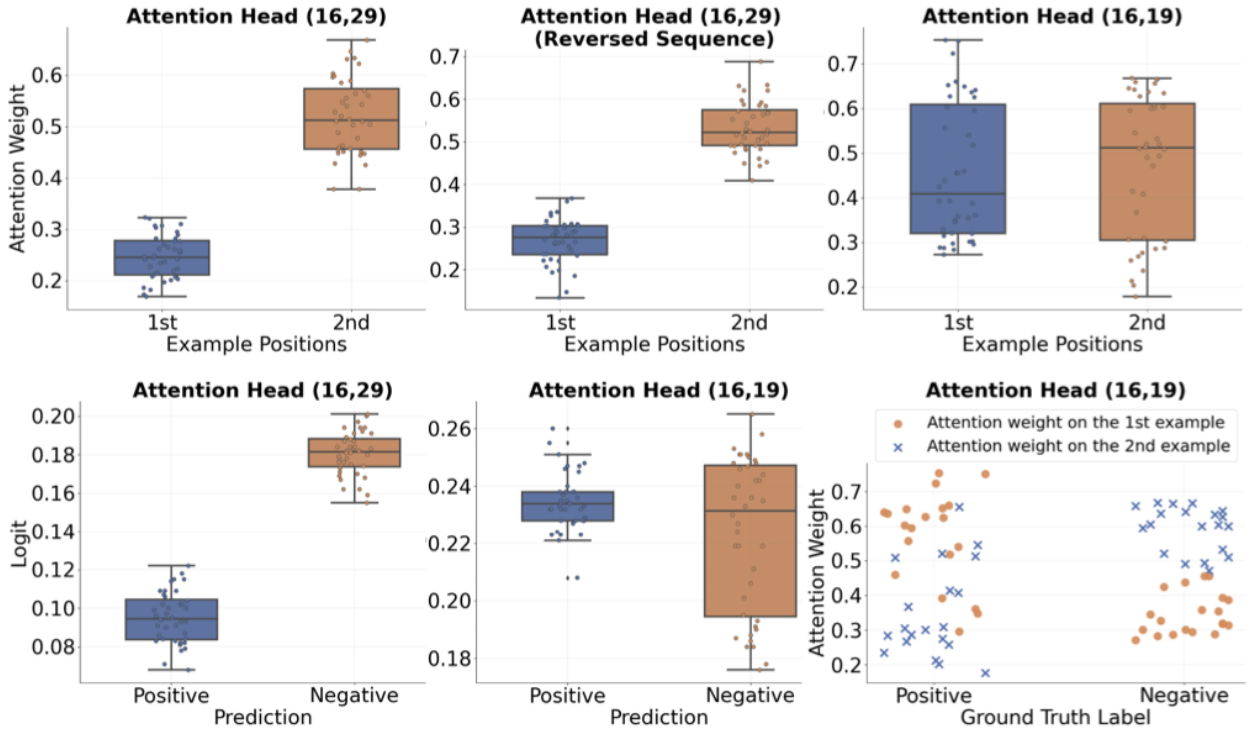


Internal Mechanism of LLM bias

Position Bias: LLMs tend to favor the label of the example at the end of the prompt

Internal Mechanism: Specific attention heads consistently prioritize the example at the end of the prompt

- Comparison between a biased attention head (layer 16, head 29) and an unbiased attention head (layer 16, head 19)



Methodology

Findings: Various bias factors stem from the biased behaviors of attention heads and FFN vectors.

Can we identify the biased components of LLMs and mitigate their impact on label prediction?

Methodology

How to identify an FFN vector or an attention head as biased?
It consistently introduces a biased preference towards specific labels into the residual stream, regardless of variations in the test samples.

- **Relatedness Criterion**: The information introduced by the FFN vector (or attention head) should closely relate to label prediction.
- **Biased Criterion**: The information contributed to the residual stream by the FFN vector (or attention head) exhibits a biased distribution, favoring certain labels over others.
- **Low Variance Criterion**: The label prediction information added by the FFN vector (or attention head) to the residual stream is almost identical across test samples with different labels,

$$\begin{cases} \frac{1}{p}\sum_{k=0}^{p-1}\text{Sum}\left(\mathbf{G}_{k,:}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\text{Sum}\left(\mathbf{g^{(k)}}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\sum_{j=0}^{c-1}g_j^{(k)}>th_{FFN}^1 \\ \frac{1}{p}\sum_{k=0}^{p-1}\text{Bias}\left(\mathbf{G}_{k,:}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\text{Bias}\left(\mathbf{g^{(k)}}\right)=\frac{1}{p}\frac{1}{c}\sum_{k=0}^{p-1}\sum_{j=0}^{c-1}\left(g_j^{(k)}-\mu(\mathbf{g^{(k)}})\right)>th_{FFN}^2 \\ CV\left(\mathbf{m}\right)=\frac{\sigma(\mathbf{m})}{\mu(\mathbf{m})}=\frac{\sqrt{\frac{1}{p}\sum_{j=0}^{p-1}\left(m_k-\mu(\mathbf{m})\right)^2}}{\frac{1}{p}\sum_{k=0}^{p-1}m_k}<th_{FFN}^3 \end{cases}$$

$$\begin{cases} \frac{1}{p}\sum_{k=0}^{p-1}\text{Sum}\left(A_{k,:}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\text{Sum}\left(\mathbf{a}^{(k)}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\sum_{j=1}^{c}\mathbf{a}_j^{(k)}>th_{Att}^1 \\ \frac{1}{p}\sum_{k=0}^{p-1}\text{Bias}\left(A_{k,:}\right)=\frac{1}{p}\sum_{k=0}^{p-1}\text{Bias}\left(\mathbf{a}^{(k)}\right)=\frac{1}{p}\frac{1}{c}\sum_{k=0}^{p-1}\sum_{j=0}^{c-1}\left(a_j^{(k)}-\mu(\mathbf{a}^{(k)})\right)>th_{Att}^2 \\ \sum_{j=0}^{c-1}w_j\cdot CV\left(A_{:,j}\right)=w_j\cdot\frac{\sigma(A_{:,j})}{\mu(A_{:,j})}<th_{Att}^3 \end{cases}$$

How to Eliminate the influence of biased FFN vectors and attention heads? Masking

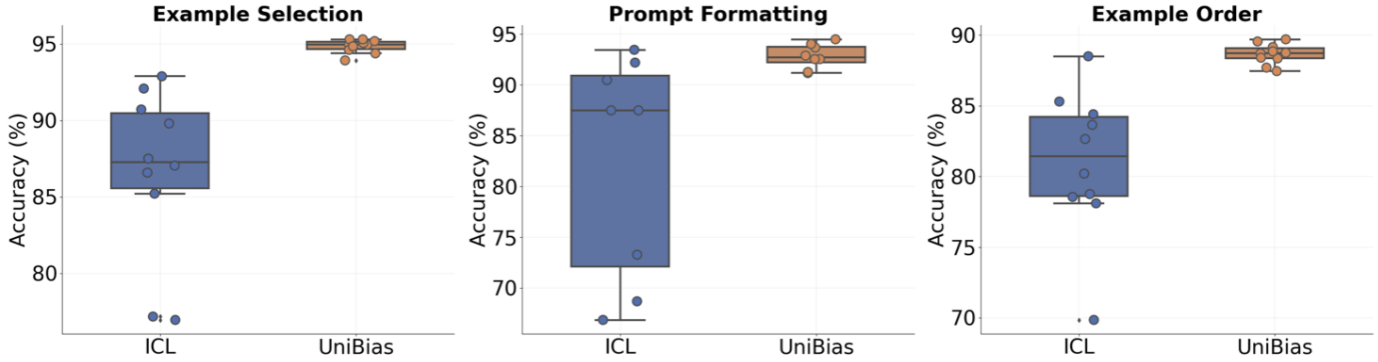
Experiments

Dataset: 12 datasets (reasoning, sentiment analysis, topic classification, natural language inference, and word disambiguation)
LLMs: 4 LLMs
Baselines: Calibration methods (CC, DC, PC)

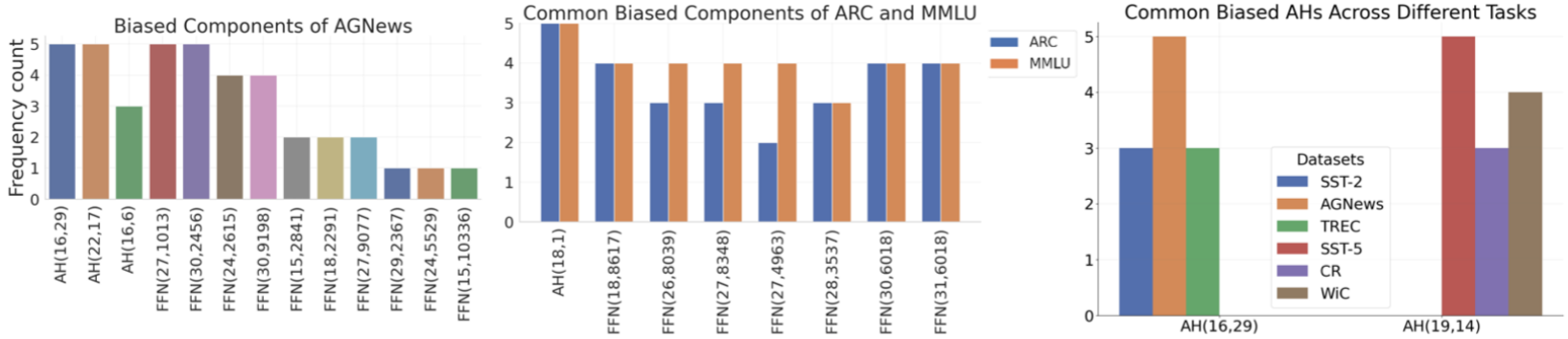
Dataset	Llama-2 7b					Llama-2 13b				
Method	ICL	CC	DC	PC	UniBias	ICL	CC	DC	PC	UniBias
SST-2	$87.22_{6.03}$	$92.24_{3.39}$	$94.15_{1.22}$	$93.90_{1.54}$	$\mathbf{94.54}_{0.62}$	$93.90_{1.79}$	$95.25_{0.93}$	$95.37_{0.70}$	$94.56_{1.71}$	$\mathbf{95.46}_{0.52}$
MNLI	$53.83_{2.22}$	$53.36_{3.16}$	$52.19_{2.55}$	$45.38_{5.01}$	$\mathbf{54.97}_{0.88}$	$62.43_{1.49}$	$63.89_{0.81}$	$61.86_{1.23}$	$57.47_{3.53}$	$\mathbf{64.65}_{2.73}$
WiC	$50.00_{0.16}$	$52.19_{2.00}$	$52.40_{1.69}$	$\mathbf{57.11}_{2.49}$	$53.71_{1.16}$	$54.48_{3.19}$	$50.63_{1.73}$	$49.72_{0.30}$	$55.67_{1.67}$	$\mathbf{57.93}_{1.70}$
COPA	$67.60_{2.30}$	$67.80_{2.17}$	$60.40_{2.79}$	$67.80_{3.70}$	$\mathbf{69.00}_{2.74}$	$67.50_{10.40}$	$75.20_{7.80}$	$71.00_{8.80}$	$76.80_{6.30}$	$\mathbf{83.20}_{2.70}$
CR	$91.54_{0.39}$	$92.13_{0.40}$	$92.61_{0.44}$	$91.97_{0.35}$	$\mathbf{92.61}_{0.11}$	$91.01_{1.30}$	$92.13_{0.88}$	$92.23_{0.76}$	$91.65_{0.64}$	$\mathbf{92.34}_{0.74}$
AGNews	$85.59_{1.87}$	$83.54_{1.96}$	$\mathbf{89.08}_{0.86}$	$86.81_{2.92}$	$88.29_{1.24}$	$89.14_{0.44}$	$88.23_{1.14}$	$\mathbf{89.34}_{0.61}$	$86.03_{0.65}$	$88.68_{0.43}$
MR	$89.37_{1.83}$	$91.77_{1.42}$	$\mathbf{92.35}_{0.23}$	$91.39_{1.65}$	$92.19_{0.37}$	$90.10_{2.10}$	$\mathbf{93.20}_{0.57}$	$93.00_{0.52}$	$92.80_{0.86}$	$92.23_{1.12}$
RTE	$66.21_{7.30}$	$64.33_{3.68}$	$65.49_{2.09}$	$62.59_{4.71}$	$\mathbf{67.65}_{6.44}$	$76.10_{4.73}$	$71.99_{5.02}$	$66.21_{1.09}$	$75.31_{2.90}$	$\mathbf{78.23}_{2.13}$
SST-5	$46.97_{0.87}$	$51.36_{1.69}$	$51.92_{1.77}$	$\mathbf{55.41}_{1.51}$	$53.79_{1.46}$	$51.03_{1.25}$	$47.20_{1.69}$	$48.98_{2.11}$	$\mathbf{53.63}_{0.95}$	$51.80_{1.00}$
TREC	$72.92_{12.42}$	$76.44_{3.21}$	$77.16_{3.94}$	$74.92_{5.78}$	$\mathbf{80.80}_{3.17}$	$74.70_{12.10}$	$\mathbf{83.80}_{3.86}$	$80.50_{9.07}$	$81.85_{9.53}$	$81.25_{6.86}$
ARC	$51.90_{0.60}$	$\mathbf{53.10}_{0.40}$	$53.00_{0.60}$	$40.40_{0.50}$	$\mathbf{53.10}_{0.60}$	$66.54_{0.33}$	$64.33_{0.99}$	$64.88_{0.59}$	$59.47_{1.07}$	$\mathbf{66.81}_{0.37}$
MMLU	$41.73_{2.25}$	$43.72_{0.97}$	$43.57_{1.38}$	$34.12_{3.41}$	$\mathbf{44.83}_{0.24}$	$53.53_{1.55}$	$50.84_{1.57}$	$51.81_{1.24}$	$45.50_{1.65}$	$\mathbf{53.55}_{1.05}$
Avg.	67.07	68.49	68.70	66.81	**70.46**	72.54	73.06	72.08	72.56	**75.51**

Experiments

Alleviating Prompt Brittleness:



Removing Common Biased Components



	SST2	MMLU	COPA	RTE	MR	Trec	Avg.
ICL	$87.22_{6.03}$	$41.73_{2.25}$	$67.60_{2.30}$	$66.21_{7.30}$	$89.37_{1.83}$	$72.92_{12.42}$	70.84
Unibias	$\mathbf{94.54}_{0.62}$	$\mathbf{44.83}_{0.24}$	$\mathbf{69.00}_{2.74}$	$\mathbf{67.65}_{6.44}$	$92.19_{0.37}$	$\mathbf{80.80}_{3.17}$	**74.84**
Eliminating Common Biased Components	$94.32_{0.60}$	$44.20_{1.14}$	$68.00_{2.87}$	$67.37_{4.60}$	$\mathbf{92.43}_{0.09}$	$77.60_{4.75}$	73.98

Summary

New Direction for LLM Bias Mitigation: Potentially stimulate future research on LLM bias mitigation from inner structure manipulation, offering a new direction for the field.

Unveil Internal Mechanisms of LLM Bias: Provide deep insights that go beyond superficial observations, revealing the inner causes of LLM bias.

Enhance the accuracy and robustness of ICL: Effectively address prompt brittleness, thereby boosting the robustness of LLMs and markedly improving the accuracy of ICL.

LLM manipulation: Showcase how to steer LLM behavior by manipulate the internal structures of large language models