ICLR
International Conference On Learning Representations

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

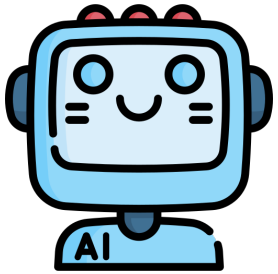
Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension Ability

Presenter: Yujin Han

Department of Computer Science, The University of Hong Kong

Motivation

- Large language models (LLMs) have demonstrated unprecedented capability in various natural language tasks.
- Debate on:



Truly understand the semantics of the question.

or

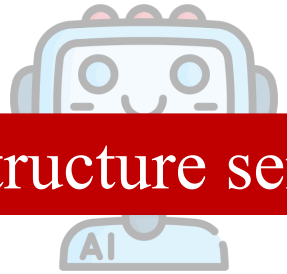
Just stochastic parrot

1. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners (ACL'24)
2. Do large code models understand programming concepts (ICML'24)

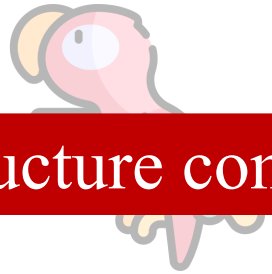
…

Motivation

- Large language models (LLMs) have demonstrated unprecedented capability in various natural language tasks.
- Debate on:



But surface structure sensitivity does not prevent deep structure comprehension



Truly understand the semantics of the question.

Just stochastic parrot

1. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners (ACL'24)
2. Do large code models understand programming concepts (ICML'24)

…

Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension Ability

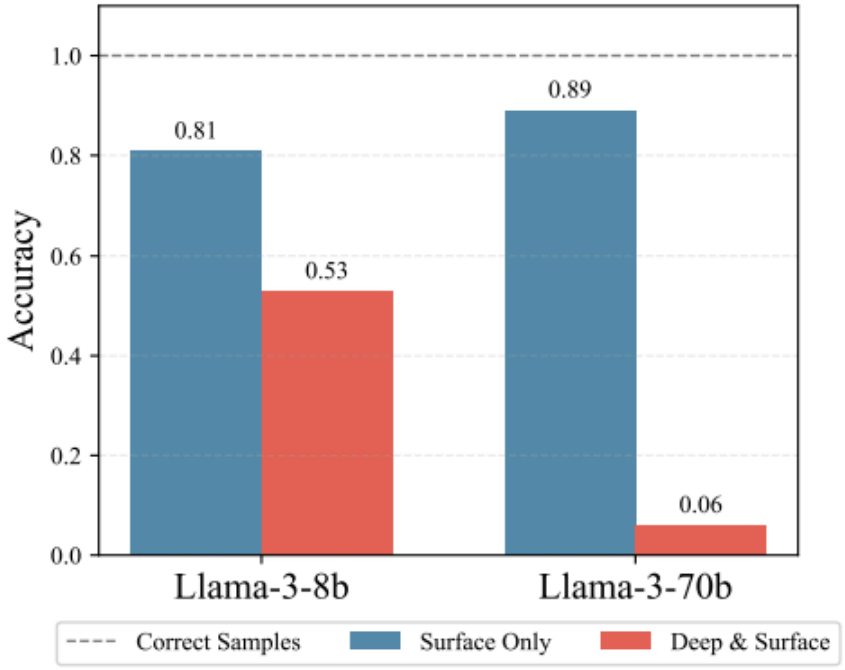
Defined in surface structure theory [Chomsky,1971]

Table 1: Examples of two-digit multiplication with interventions on deep and surface structures: deep structure embodies core semantics (e.g., numbers and operators), while surface structure encompasses linguistic forms (e.g., question format). Among given intervention strategies, changes in deep structure inherently alter surface structure. More examples on both structures in Appendix A.

Example Questions	Deep & Surface Intervention	Surface Intervention Only	Strategy
What is 50 times 20 ? A:1000	What is ⟨Mask⟩ times 20? A:None	What is 50 times 20⟨Mask⟩ A:1000	*Mask*
	How much is 10 multiplied by 50? A:500	How much is 20 multiplied by 50? A:1000	*Rephrase*
	What is * times 20? A:None	What * 50 times 20? A:1000	*Replace*
	50 is What times 20? A:2.5	is What 50 times 20? A:1000	*Swap*

Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension Ability

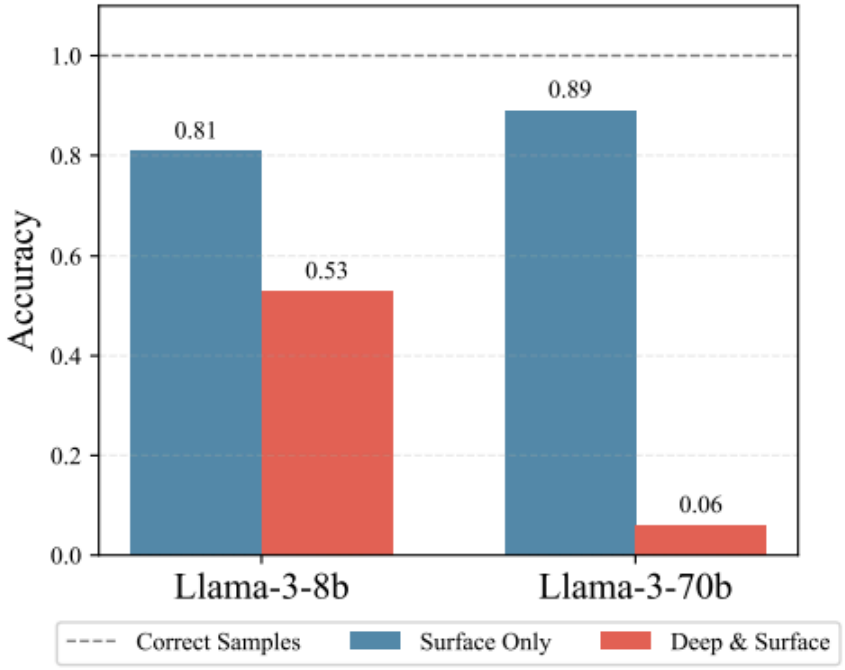
Experiments on 2-digit multiplication Data (Mask strategy)



Surface-only interventions cause slight accuracy decline, while combined surface and deep modifications result in significant performance degradation

Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension Ability

Experiments on 2-digit multiplication Data (Mask strategy)



Question: Do LLMs genuinely comprehend deep structure for problem-solving, or do they primarily rely on learning surface structure?

Metric: (1) Quantify LLMs' understanding capabilities of deep and surface structures; (2) Be widely applicable across diverse tasks and LLMs

Method

Think from causal graph with mediation

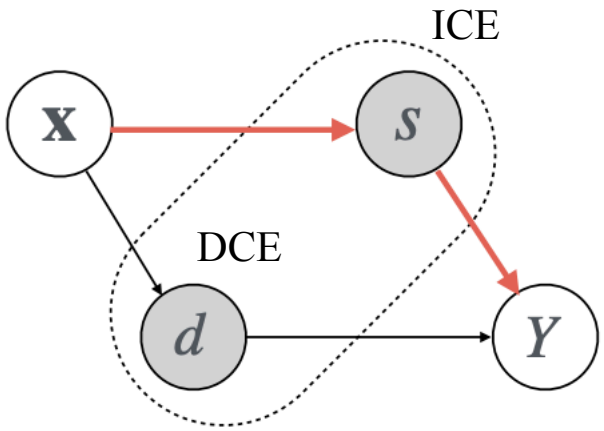


Figure 3: Causal graph with mediation: $\boldsymbol{x} \rightarrow d \rightarrow Y$ shows deep structures' direct causal effect, $\boldsymbol{x} \rightarrow s \rightarrow Y$ indicates surface structures' indirect causal effect via mediator s.

Treatment assignment variable T on input x_i

$$T = \begin{cases} 0 & \text{intervention alters } s_i\text{, preserves } d_i \\ 1 & \text{intervention alters both } s_i \text{ and } d_i \end{cases}$$

Accurate DCE

$$\underbrace{\delta_{\text{DCE}}}_{\text{DCE}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=1)) - Y_i^{\text{origin}}]}_{\text{ICE}}$$

The required assumptions for causal mediation analysis are satisfied.

Method

Think from causal graph with mediation

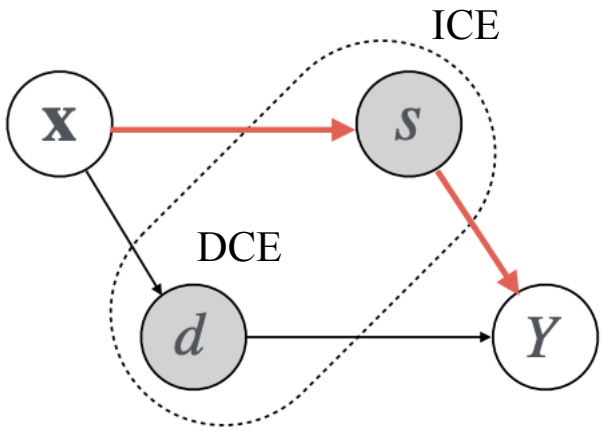


Figure 3: Causal graph with mediation: $\boldsymbol{x} \rightarrow d \rightarrow Y$ shows deep structures' direct causal effect, $\boldsymbol{x} \rightarrow s \rightarrow Y$ indicates surface structures' indirect causal effect via mediator s.

What we can:

Accurate DCE

$$\underbrace{\delta_{\text{DCE}}}_{\text{DCE}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=1)) - Y_i^{\text{origin}}]}_{\text{ICE}}$$

Method

Think from causal graph with mediation

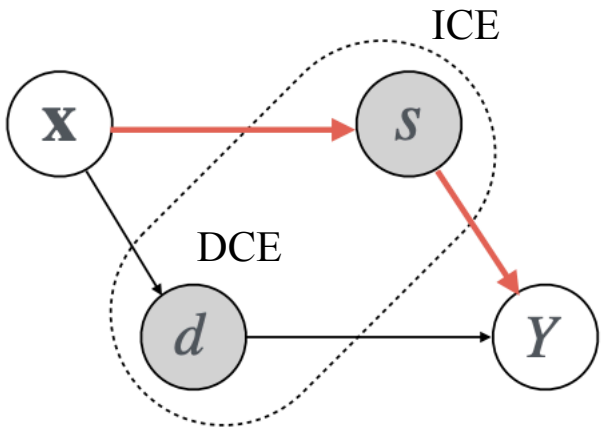


Figure 3: Causal graph with mediation: $\boldsymbol{x} \rightarrow d \rightarrow Y$ shows deep structures' direct causal effect, $\boldsymbol{x} \rightarrow s \rightarrow Y$ indicates surface structures' indirect causal effect via mediator s.

What we cannot:

Accurate DCE

$$\underbrace{\delta_{\text{DCE}}}_{\text{DCE}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=1)) - Y_i^{\text{origin}}]}_{\text{ICE}}$$

Method

Think from causal graph with mediation

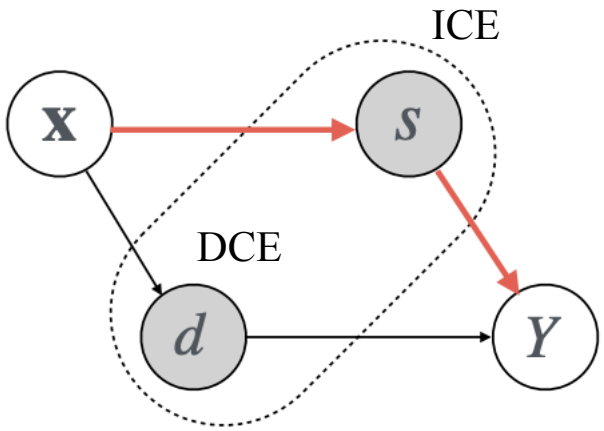


Figure 3: Causal graph with mediation: $\boldsymbol{x} \rightarrow d \rightarrow Y$ shows deep structures' direct causal effect, $\boldsymbol{x} \rightarrow s \rightarrow Y$ indicates surface structures' indirect causal effect via mediator s.

Therefore,

Accurate DCE

$$\underbrace{\delta_{\text{DCE}}}_{\text{DCE}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=1)) - Y_i^{\text{origin}}]}_{\text{ICE}}$$

Approximate DCE

$$\underbrace{\delta_{\text{ADCE}}}_{\text{nated DCE (ADCE)}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=0)) - Y_i^{\text{origin}}]}_{\text{approximated ICE (AICE)}}$$

Method

Accurate DCE

$$\underbrace{\delta_{\text{DCE}}}_{\text{DCE}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=1)) - Y_i^{\text{origin}}]}_{\text{ICE}}$$

as similar as possible

Approximate DCE

$$\underbrace{\delta_{\text{ADCE}}}_{\text{mated DCE (ADCE)}} = \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=1, s(T=1)) - Y_i^{\text{origin}}]}_{\text{TE}} - \underbrace{\mathbb{E}_{\boldsymbol{x}_i}[Y_i(T=0, s(T=0)) - Y_i^{\text{origin}}]}_{\text{approximated ICE (AICE)}}$$

Strategy

Mask: masking k non-core semantic words closest to the masked core semantic word in TE

Rephrase: minimizing word-level modifications to transform TE

Approximate DCE (ADCE)

Approximate ICE (AICE)

Property: Causality not Correlation

Theorem 1. *(ADCE as a Combination of PN and PS) Let T be the treatment variable in Equation 2 and $\hat{Y}$ the outcome of the indicator function in Equation 5. Assume $\hat{Y}$ is monotonic with respect to T, for ADCE, it holds that:*

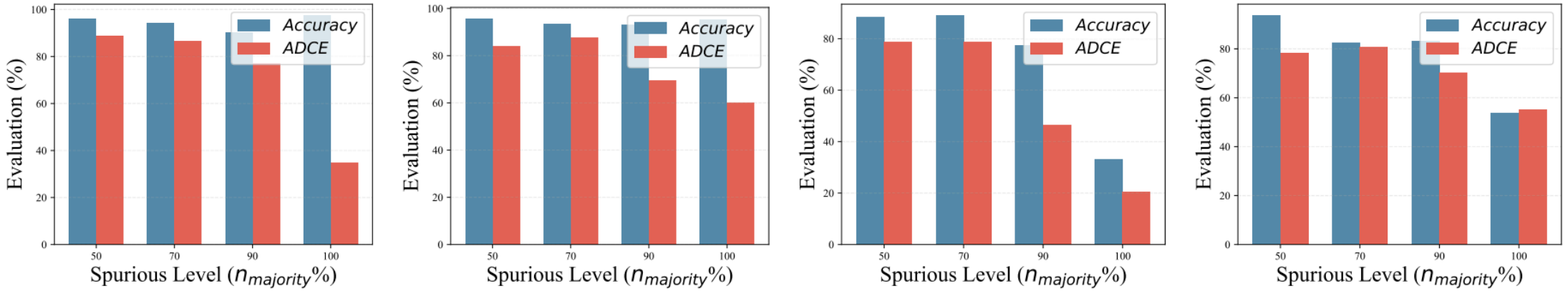
$$\delta_{\text{ADCE}} = \frac{\alpha}{2} \cdot \delta_{\text{PS}} + \frac{\beta}{2} \cdot \delta_{\text{PN}} \tag{7}$$

where $\alpha := \mathbb{P}(\hat{Y} = 1 | T = 1, s(T = 1))$, $\beta := \mathbb{P}(\hat{Y} = 0 | T = 0, s(T = 0))$.

probability of sufficiency (PS) sufficient condition $X \Rightarrow Y$

probability of necessity (PN) necessary condition $Y \Rightarrow X$

Property: ADCE is Causal, Accuracy is Correlated



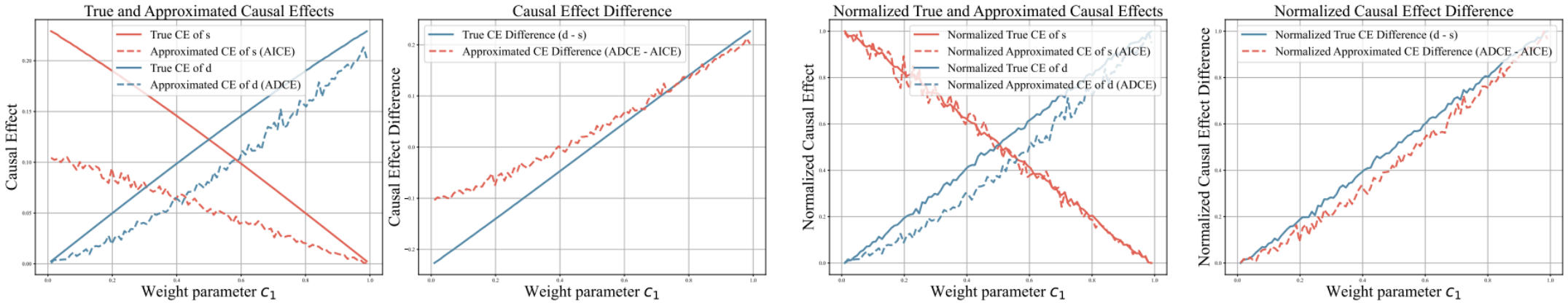
(a) Majority, 8b (b) Majority, 70b (c) Minority, 8b (d) Minority, 70b

Figure 8: Spurious correlation results in LLama-3. In majority groups with spurious correlations, increasing correlation levels lead to high accuracy but declining ADCE. In minority groups without spurious correlations, accuracy and ADCE trends align. ADCE better reflects the model's reliance on spurious attributes over core semantics in spurious conditions, compared to accuracy.

Property: The accurate of ADCE

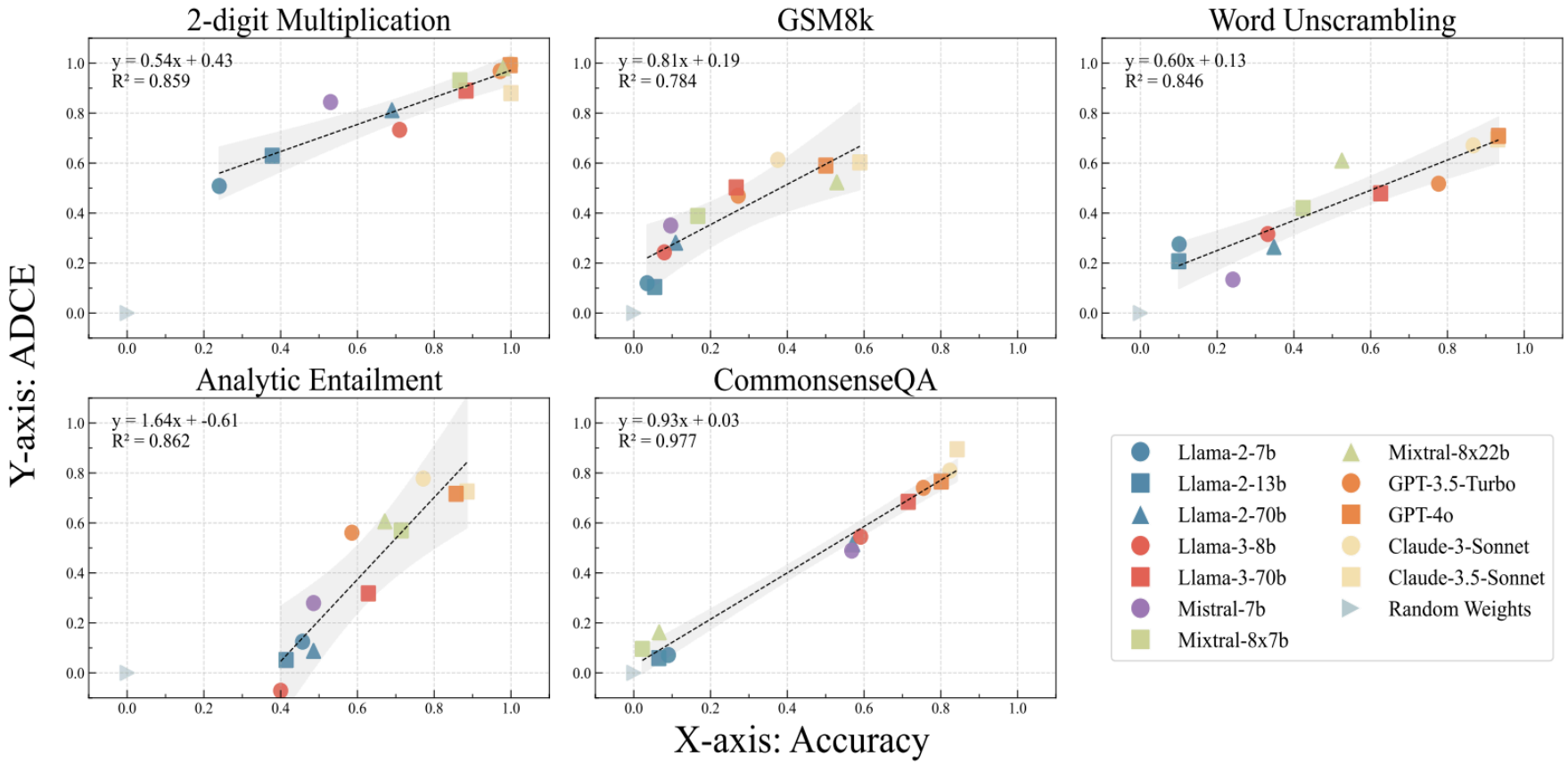
SCM $x \sim \mathcal{N}(0,1), \quad d = x + \epsilon_d, \quad s = x + \epsilon_s.$

$$y = \begin{cases} 1, & \text{if } \sigma(c_1 \cdot d + c_2 \cdot s + \epsilon_y) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

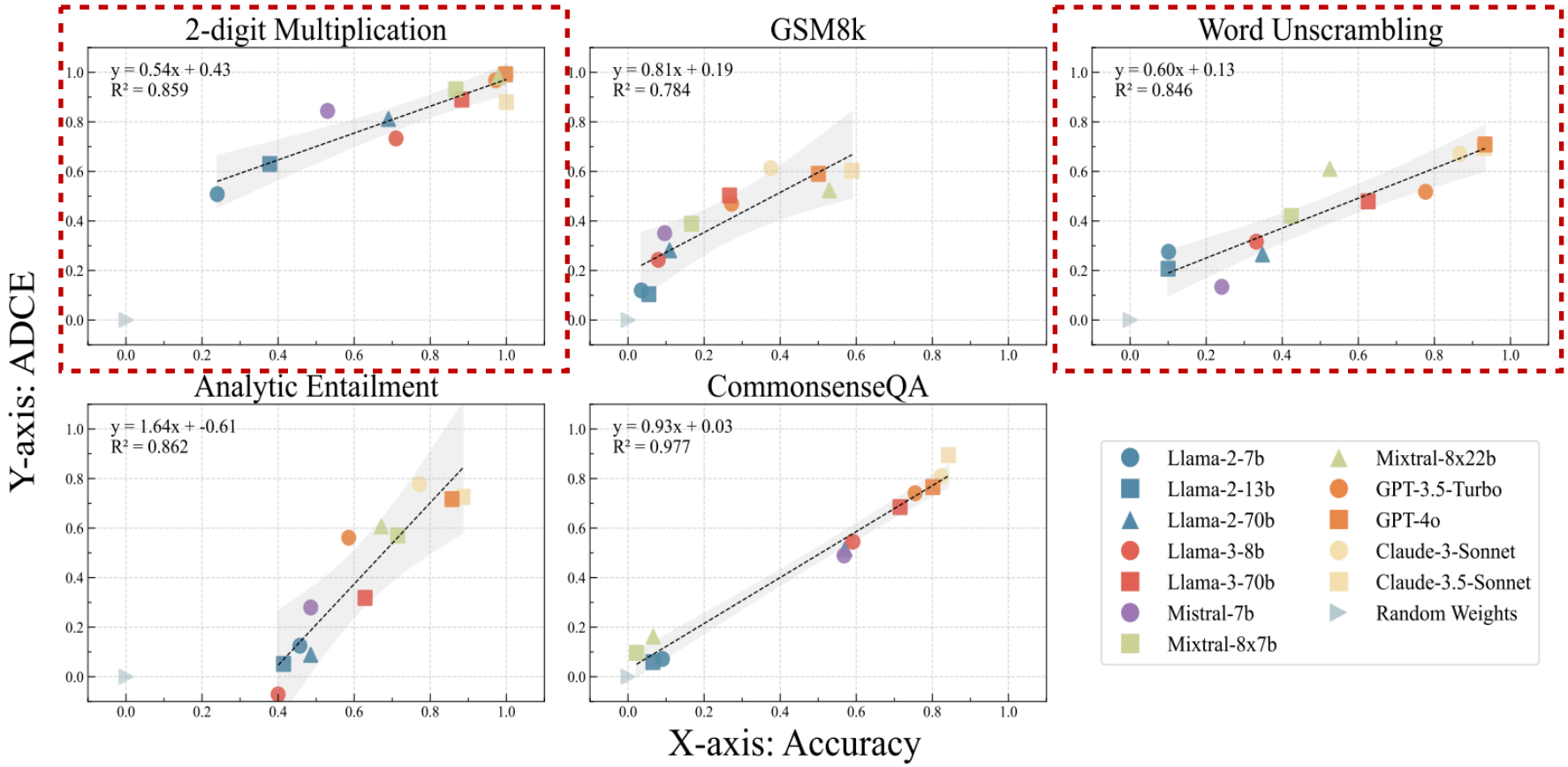


(a) Unnormalized Causal Effects (b) Normalized Causal Effects

Experiments

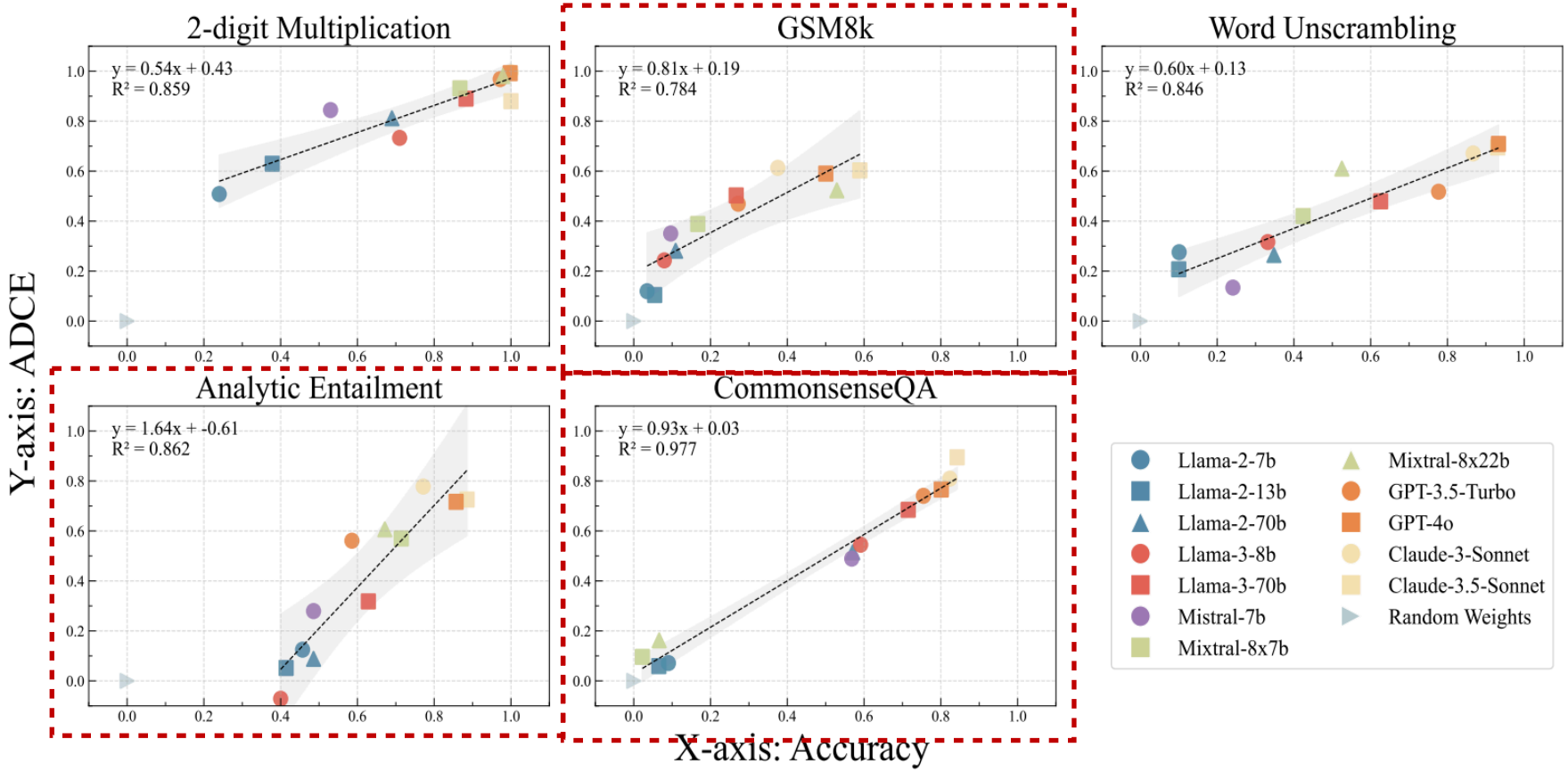


Experiments



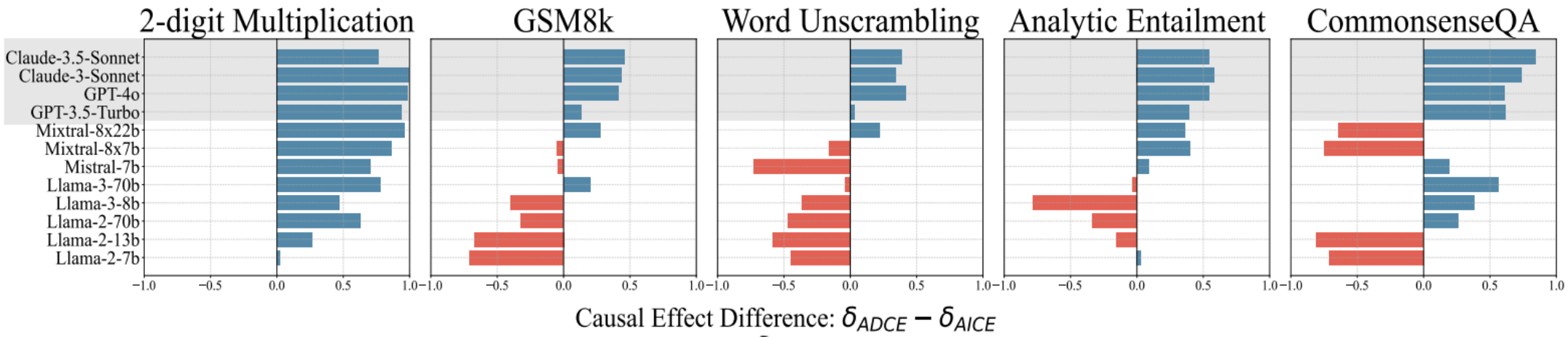
Low-β tasks (e.g., 2-Digit Multiplication, Word Unscrambling) have fixed formats and single-skill requirements, needing small deep structure understanding for improvement.

Experiments



High-β tasks (e.g., GSM8k, Analytic Entailment, CommonsenseQA) involve multi-step reasoning, diverse logical relationships and broad knowledge, demanding varied deep structure comprehension for accuracy gains.

Experiments



Open-source models (e.g. LLama-2) are more sensitive to surface structure; however, as model scale increases, this sensitivity is mitigated

More Experiments

- Post training strategy: SFT, In-context learning, Instruction Fine-Tuning, Fine-Tuning with In-Context Learning … helps!

- Noisy data: ADCE (AICE) reasonable

Thank You

Yujin Han

Department of Computer Science

The University of Hong Kong