Evaluating and Explaining Prompt Sensitivity of LLMs Using Interactions

Ruiyang Qin, Qingzhuo Wang, Tian Wang, Zhihua Wei, Wen Shen†

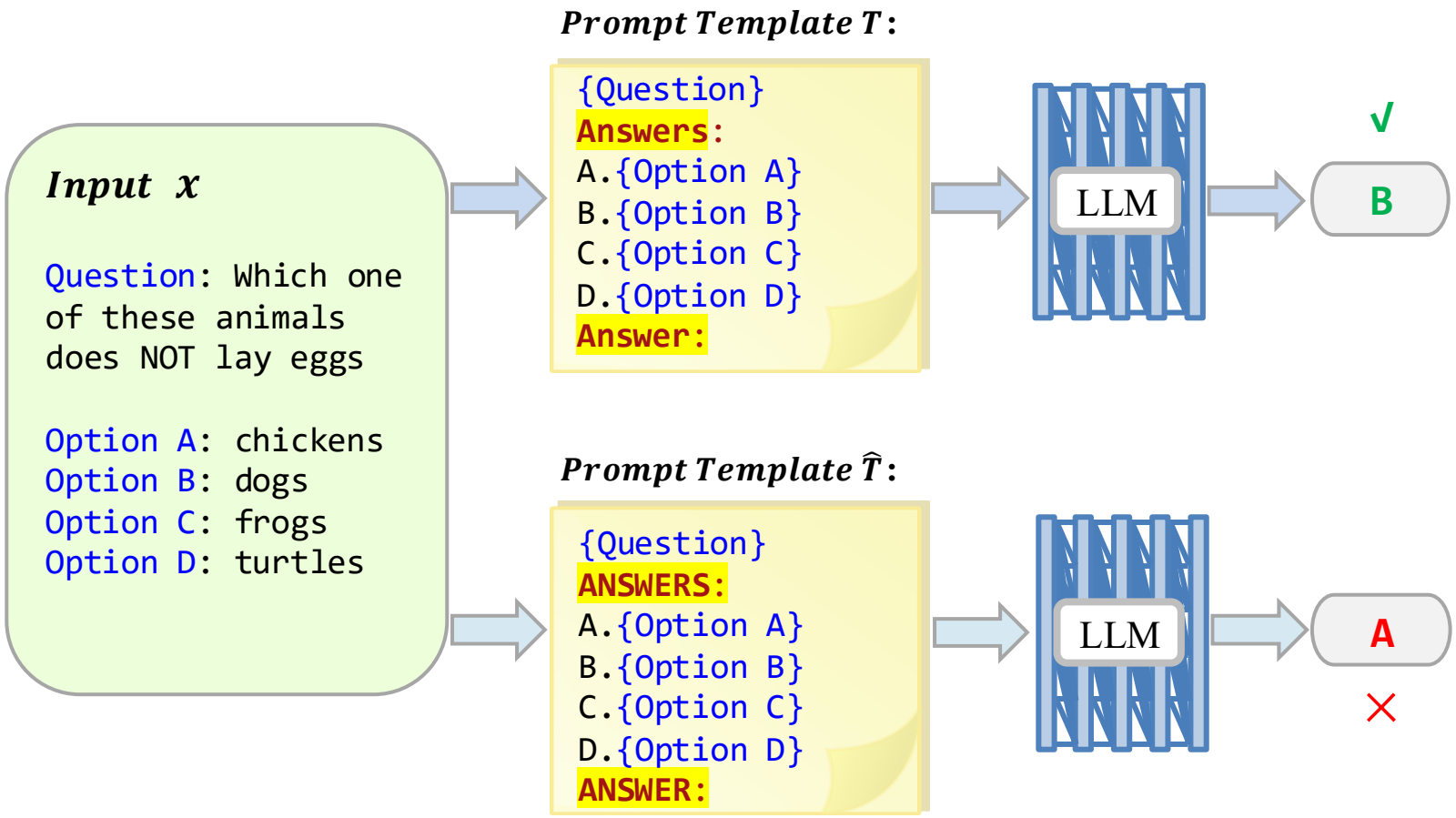
Tongji University

(† Correspondence)

Prompt Sensitivity and Traditional Metrics

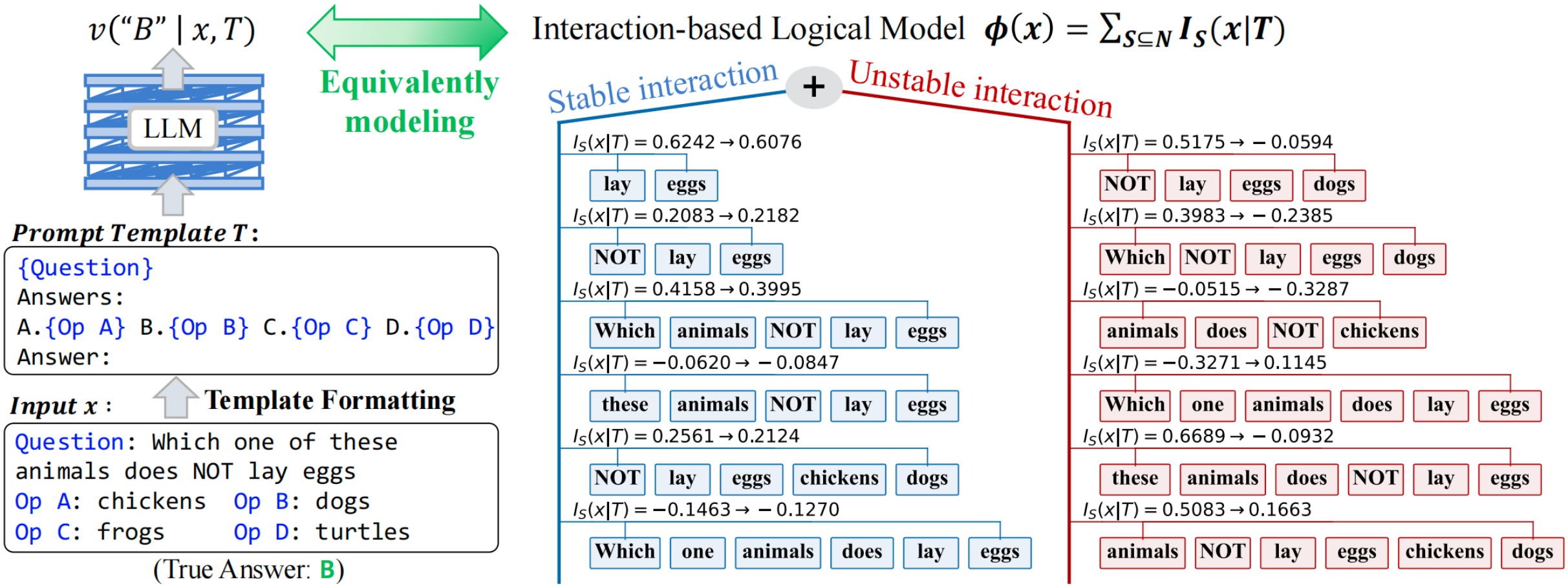
Prompt Sensitivity : Subtle changes in prompts can result in divergent outputs.

Traditional (Coarse-grained) Metrics : These output-based metrics typically measure changes in performance, such as task accuracy or output consistency.



Using Interactions for Fine-Grained Analysis of Prompt Sensitivity

$$v(x) = \phi(x) = \sum_{S \subseteq N} I_S(x|T)$$



Preliminaries: Interactions

Given **a DNN** v and **an input sentence** x with n words indexed by $N = \{1, 2, \ldots, n\}$,

let $v(x) \in R$ denote the scalar output of the DNN.

$$v(x) = \log \frac{p(y = y^* | x)}{1 - p(y = y^* | x)} \in R$$

We define a logical model $\phi(x)$ to match the output $v(x)$ of the DNN.

Given any randomly masked input x_T, $\phi(x_T)$ is defined as :

$$\phi(x_T) \triangleq \phi(x_\emptyset) + \sum_{S \subseteq N} \mathbb{1}(\mathrm{S} \mid x_T) \cdot I_S$$

Later, we'll prove :
$\forall\, T \subseteq N,$
$\phi(x_T) = v(x_T)$

The AND trigger function $\mathbb{1}(\mathrm{S} \mid x_T) \in \{0, 1\}$: an **AND relationship** between words in S.

$I_S = \sum_{S' \subseteq \mathrm{S}} (-1)^{|S| - |S'|} \cdot v(x_{S'})$: quantifies the **interaction effect** of an AND relationship.

Preliminaries: Interactions

AND Interaction *For example, given the input sentence x = "He is a green hand"*

The interaction S = {*green, hand*} contributes an effect I_S that pushes logical model $\phi(x)$'s inference towards the semantic meaning of "beginner."

x_T	$\mathbb{1}(\mathrm{S} \mid x_T)$	If triggered
x_T = {*green*}	$\mathbb{1}(\mathrm{S} \mid x_T) = 0$	✗
x_T = {*hand*}	$\mathbb{1}(\mathrm{S} \mid x_T) = 0$	✗
x_T = {*green, hand*}	$\mathbb{1}(\mathrm{S} \mid x_T) = 1$	✓

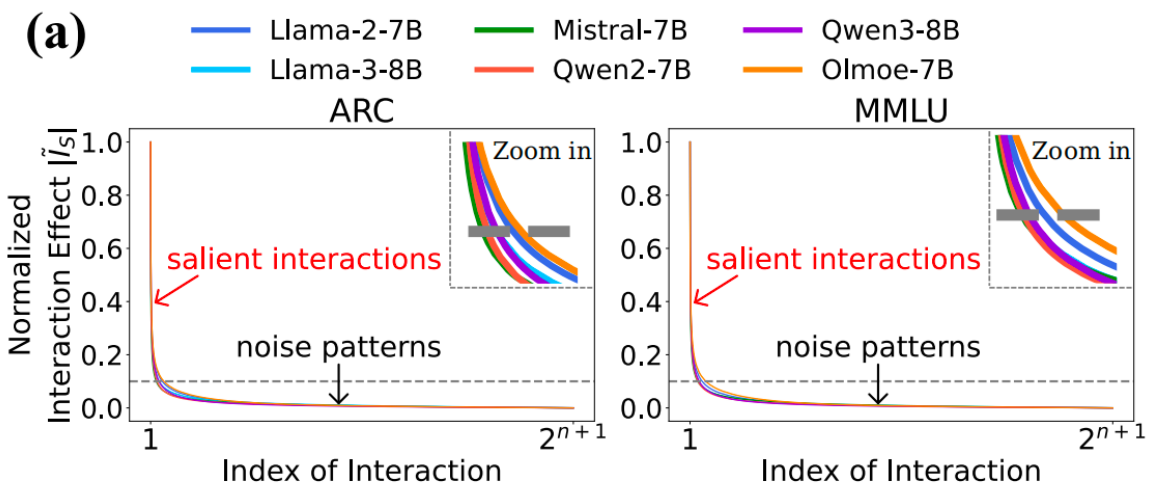
Only if the interaction is triggered (✓), the effect I_S is added to the output of $\phi(x_T)$

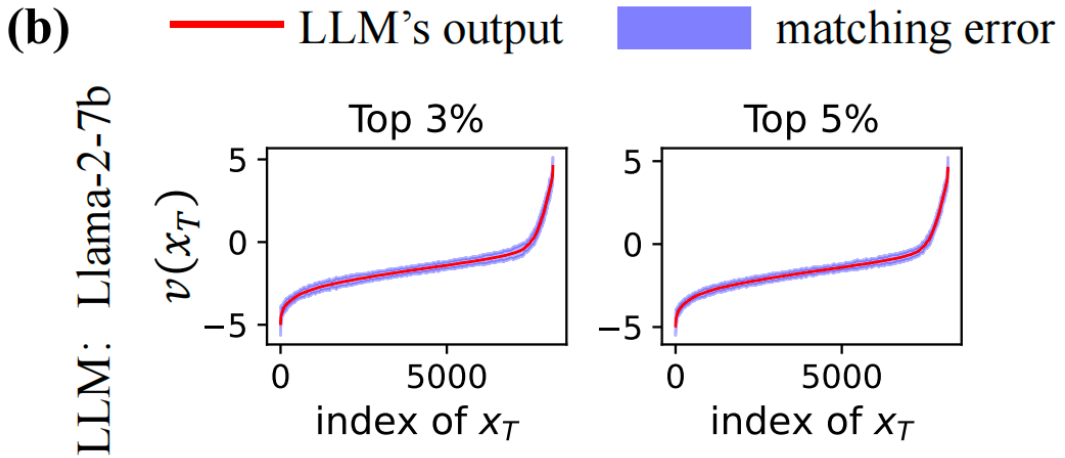
Faithfulness of Considering Interactions as Inference Patterns Used by LLMs

Theorem (*Universal matching property*, proved by [1]): For every masked input x_T, the output of the logical model $\phi(\cdot)$ can always match the DNN's output $v(\cdot)$.

$$\forall\, T \subseteq N, v(x_T) = \phi(x_T) = v(x_\emptyset) + \sum_{S\subseteq N} \mathbb{1}(S \mid x_T) \cdot I_S$$

Sparsity property: The LLM's output can be faithfully approximated by a small set of **salient interactions**, while the majority of interactions can be regarded as **noise patterns**.



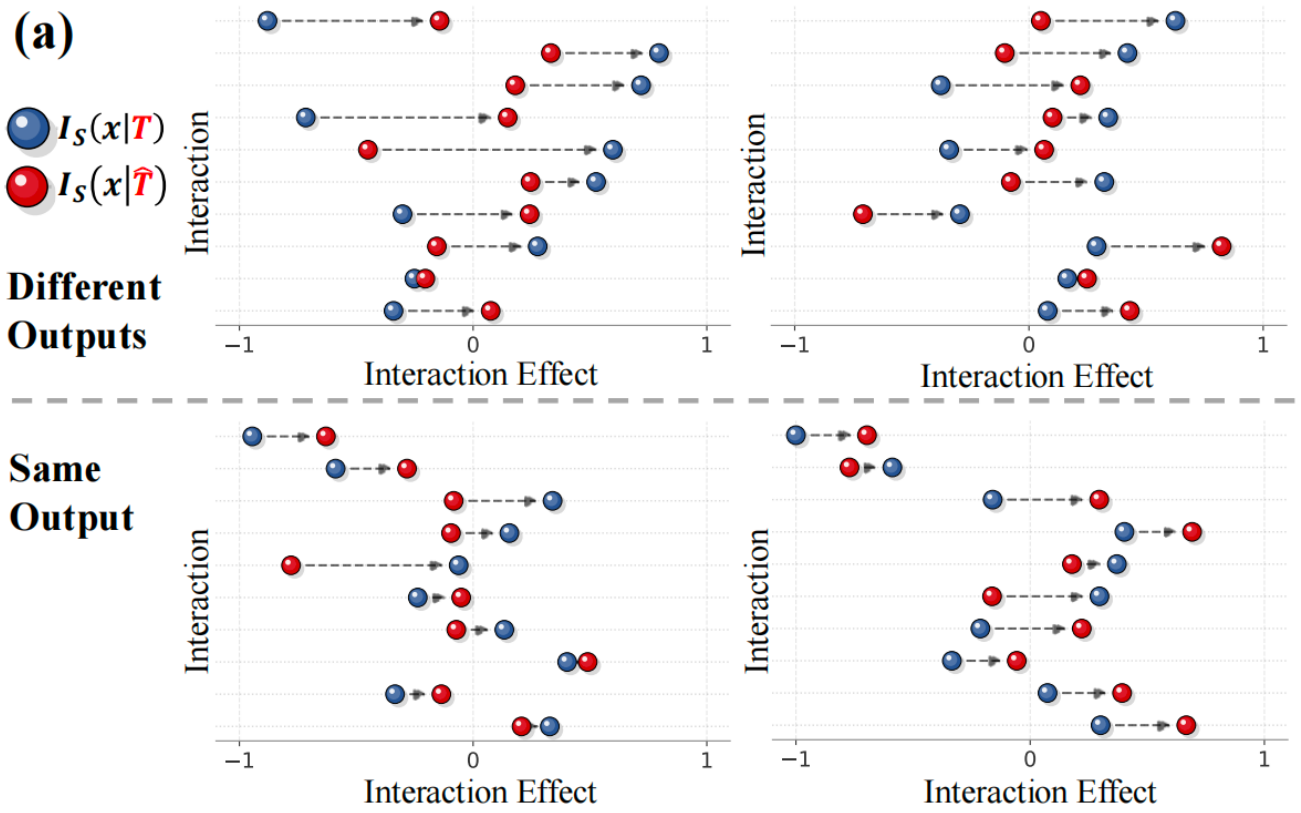


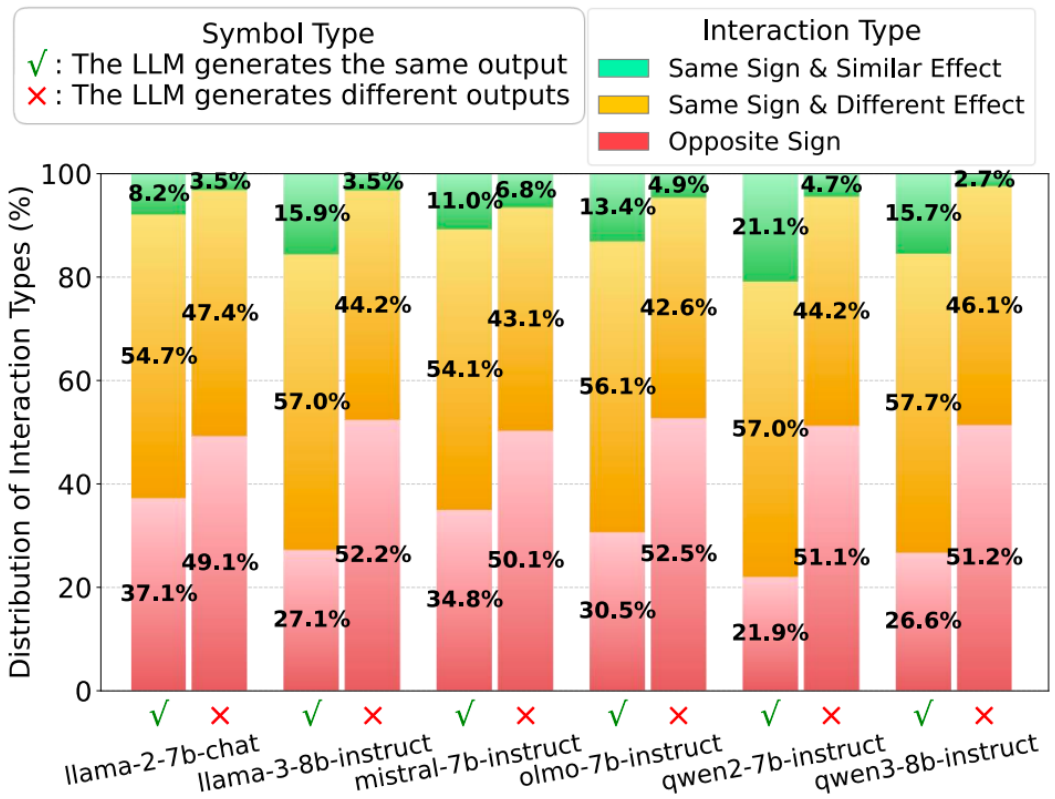
[1] Ren et al. Defining and Quantifying the Emergence of Sparse Concepts in DNNs.

Using Interactions as a Fine-Grained Tool to Analyze the Prompt Sensitivity of LLMs

The interactions are quite unstable, even when the LLM's output **remains the same**.

The output-level analysis is **insufficient** to capture the unreliable internal patterns of LLMs.





Evaluating Interaction-Based Prompt Sensitivity

We propose a new metric **I**nteraction-Based **P**rompt **S**ensitivity (**IPS**) to evaluate the prompt sensitivity of LLMs.

$$IPS \triangleq E_x\left[E_{T,\hat{T}}\left[\frac{1}{|\Omega_{\text{union}}|}\sum_{S\in\Omega_{\text{union}}}\frac{\left|\widetilde{I}_S(x|T)-\widetilde{I}_S\left(x|\hat{T}\right)\right|}{\frac{\left|\widetilde{I}_S(x|T)\right|+\left|\widetilde{I}_S\left(x|\hat{T}\right)\right|}{2}}\right]\right]$$

where,

$$\Omega_{\text{salient}}(x|T)=\{S\in\Omega(x)\mid\left|\widetilde{I}_S(x|T)\right|>\tau\}$$

$$\Omega_{\text{salient}}\left(x|\hat{T}\right)=\{S\in\Omega(x)\mid\left|\widetilde{I}_S\left(x|\hat{T}\right)\right|>\tau\}$$

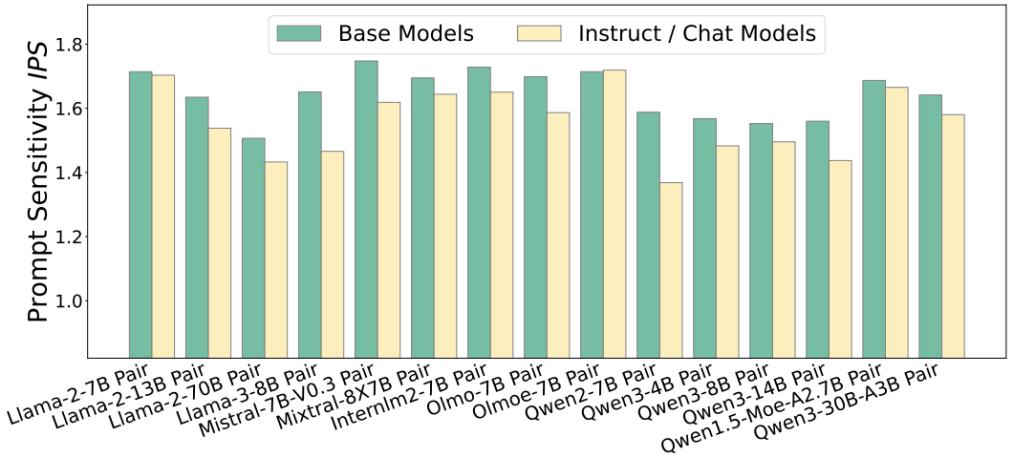
$$\Omega_{\text{union}}=\Omega_{\text{salient}}(x|T)\cup\Omega_{\text{salient}}\left(x|\hat{T}\right)$$

Analyzing the Factors Impacting Prompt Sensitivity

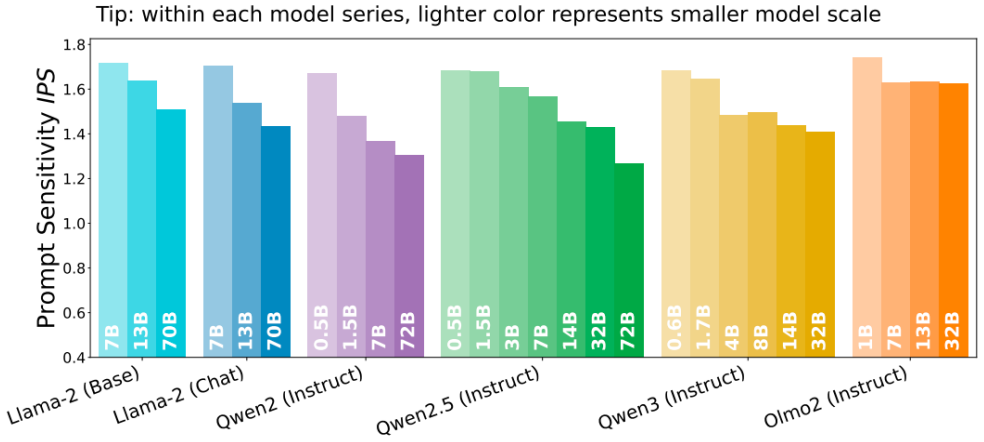
We apply the IPS metric to evaluate the prompt sensitivity of 50 open-source LLMs.

Furthermore, we discover **four factors** that can influence the prompt sensitivity of LLMs.

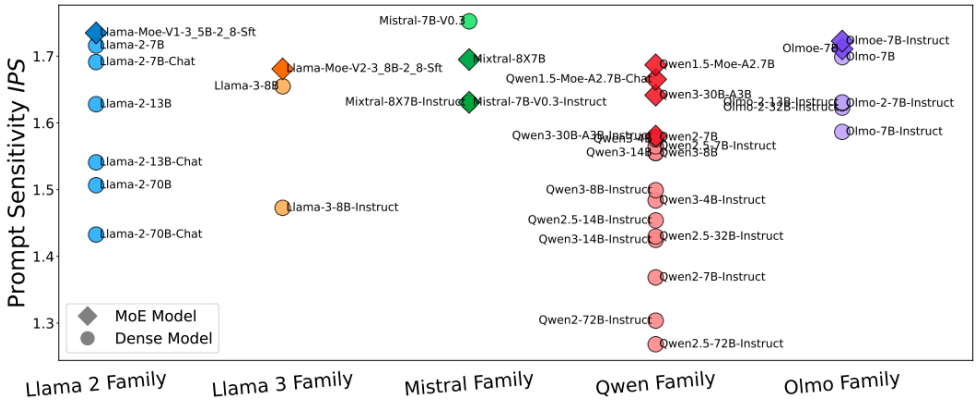
Factor 1: instruct/chat models vs. base models



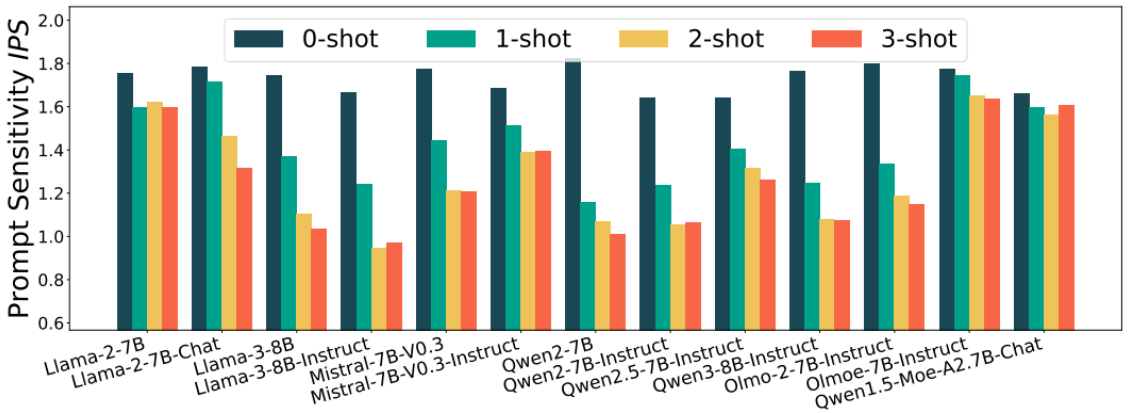
Factor 2: model scales



Factor 3: dense models vs. MoE models

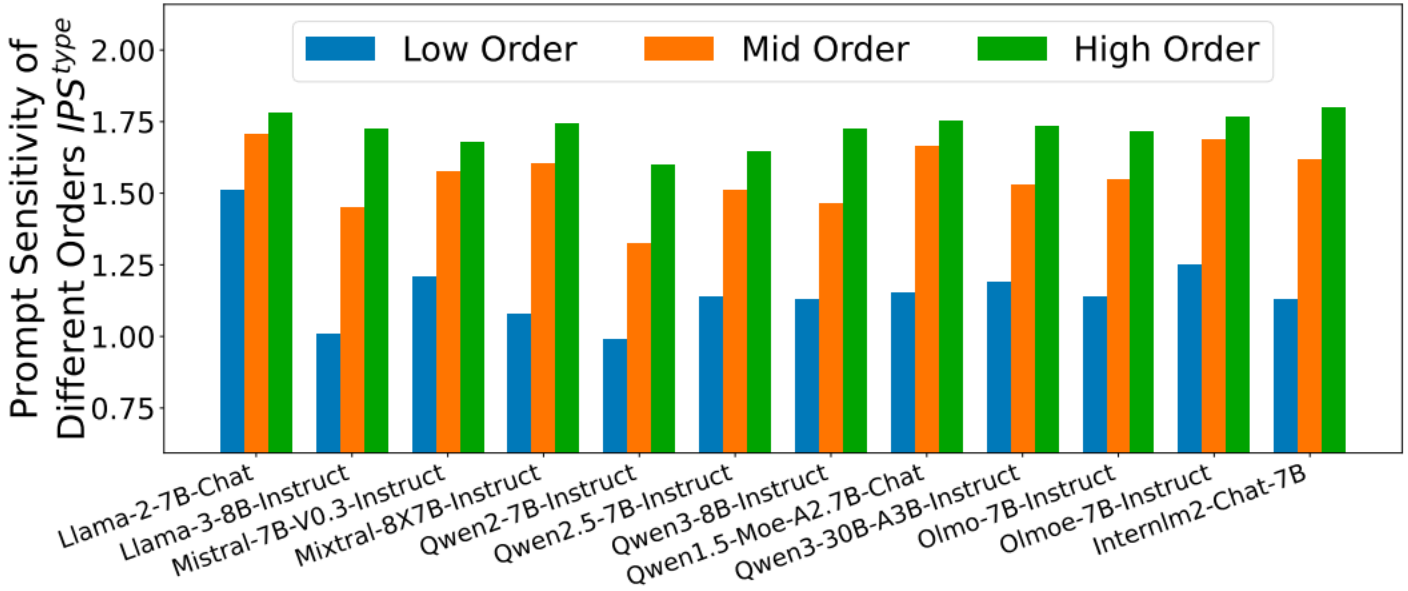


Factor 4: few-shot learning vs. 0-shot learning



Explore the Underlying Mechanisms of Improved Stability for All Factors

We analyze the prompt sensitivity of different types of interactions to explore the underlying mechanisms. Specifically, we partition interactions into three distinct groups based on their **orders** (**complexities**): low-order, mid-order, and high-order.

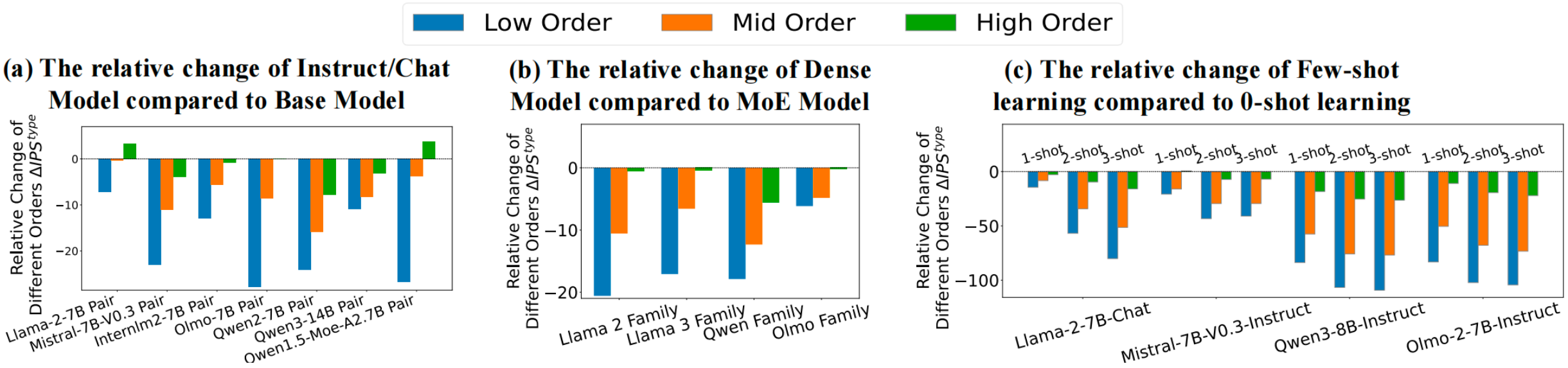


Results show that the prompt sensitivity of **low-order** interactions is the **lowest**, followed by mid-order, while high-order interactions exhibit the highest prompt sensitivity.

Explore the Underlying Mechanisms of Improved Stability for All Factors

We investigate how the four aforementioned factors influence the prompt sensitivity of low-, mid-, and high-order interactions.

Conclusion: All four factors achieve lower prompt sensitivity primarily by reducing the sensitivity of **low-order** interactions, while the prompt sensitivity of high-order interactions remains at a relatively high level.



Thanks For Watching