A Unified Approach to Interpreting Knowledge Distillation for Large Language Models via Interactions

Qingzhuo Wang*, Ruiyang Qin*, Zhenxin Qin, Wen Shen†, Zhihua Wei†

Tongji University

(* Equal Contribution)
(† Correspondence)

Motivation

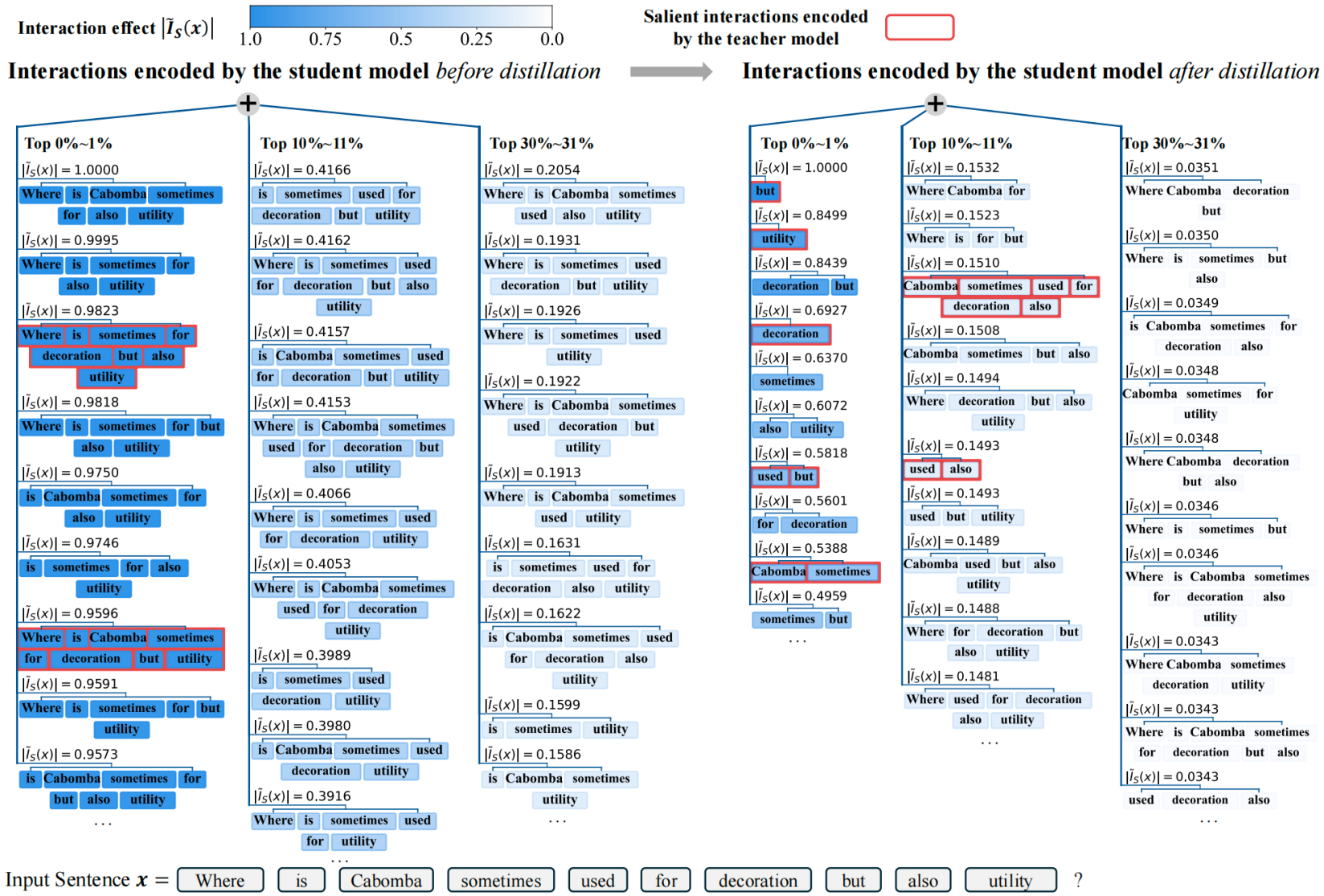
- **Motivation**: Explain the common mechanism behind why **K**nowledge **D**istillation (KD) works for LLMs.
- **Why it matters**: KD is essential for deploying compact LLMs under computation and data constraints.
- **Limitation of prior work**: Strong empirical KD methods exist, but a unified and faithful explanation of why KD methods work is missing.
- **Our Core insight**: We propose a unified approach to explore the common mechanism of various KD methods using interactions.

Comparing the interaction patterns encoded by the student model before and after distillation.

Conclusions:

The essence of distillation lies in the **sparsification** of interactions.

The student model after distillation **retain more salient interactions** learned from the teacher model than the student model before distillation, while setting many other interactions to **nearly zero effects**.



Preliminaries: Interactions

Given **a LLM** 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 LLM.

$$v(x) = \log \frac{\bar{p}}{1-\bar{p}} \in R, \text{ where } \bar{p} = \left(\prod_{l=1}^{L} p\left(y_l^* \middle| x, y_{<l}^*\right)\right)^{1/L}$$

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

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 = "I am a green hand means that I am a"*

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 LLM's output $v(\cdot)$.

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

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

Interaction-Based Metrics for Analyzing KD

We normalize the interaction effect as $\widetilde{I_S} = \frac{I_S}{Max}$, where Max is the maximum absolute values of all interactions.

Interaction Sparsity

- *Gini coefficient*. Let $\mathcal{I} = \{|\widetilde{I_S}| : S \subseteq N, S \neq \emptyset\}$ denote the set of absolute value of all normalized interaction effects extracted from the input x. Let M = $|\mathcal{I}| = 2^n - 1$ denote the number of all the interactions. Let the values in $\mathcal{I}$ be sorted in ascending order as $u_1 \leq u_2 \leq \cdots \leq u_M$. For the input x, the Gini coefficient $G(x)$ is defined as:

$$G(x) = \frac{2\sum_{i=1}^{M} i\, u_i}{M\sum_{i=1}^{M} u_i} - \frac{M+1}{M}$$

- *Shannon entropy*. The entropy $H(x)$ is defined as $H(x) = -\sum_S p_S \log p_S$, where $p_S = |\widetilde{I_S}| / \sum_{S'} |\widetilde{I_{S'}}|$ is the normalized probability.

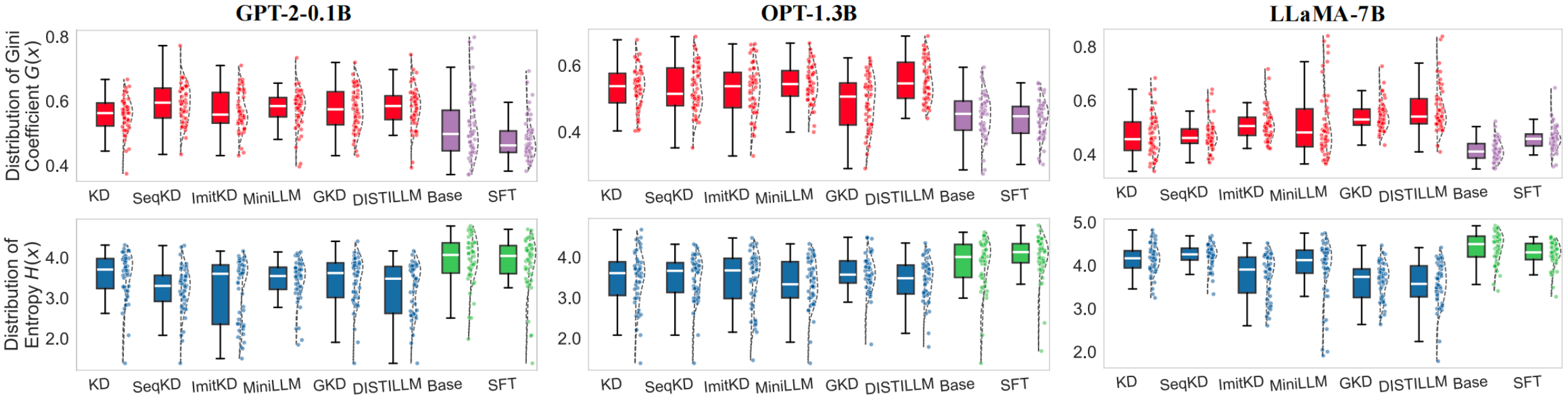
Interaction Alignment

We define a threshold ratio $k \in (0,1]$. Let $\Omega^{(k)}_{\text{teacher}}$ and $\Omega^{(k)}_{\text{student}}$ represent the sets containing the indices of the $top - \lfloor k \times M \rfloor$ interactions with the largest absolute effects in the teacher and student models, respectively.

$$Overlap@k(x) = \frac{\left|\Omega^{(k)}_{\text{teacher}} \cap \Omega^{(k)}_{\text{student}}\right|}{\lfloor k \times M \rfloor}$$

Exploring the Common Mechanism of KD Methods

Distillation enhances interaction sparsity.



Distillation enhances interaction alignment.

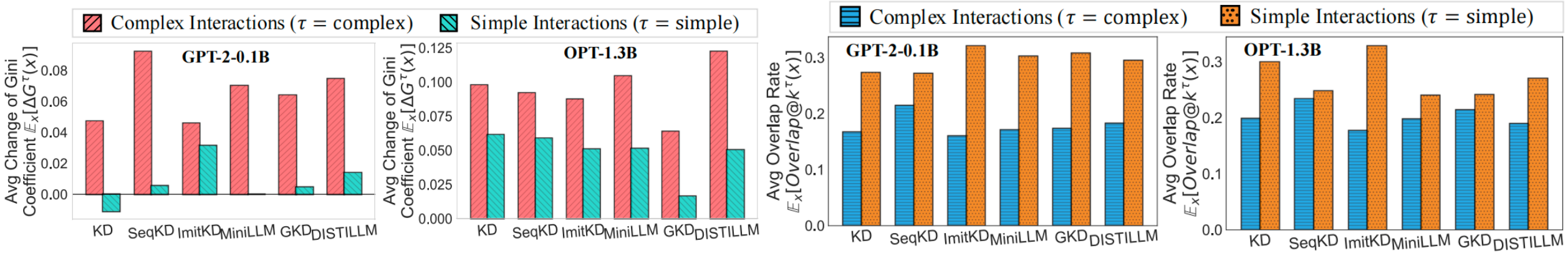
$\mathbb{E}_x\left[Overlap@k(x)\right]$	Model	non-distilled	distilled (with different KD methods)					
		Base	KD	SeqKD	ImitKD	MiniLLM	GKD	DISTILLM
$k = 0.1$	GPT-2-0.1B	19.81%	24.84%	26.96%	23.05%	23.68%	26.69%	26.38%
	OPT-1.3B	21.96%	25.24%	26.48%	21.47%	26.54%	25.23%	25.71%
	LLaMA-7B	21.35%	24.18%	21.55%	22.02%	27.62%	29.43%	31.60%

Exploring the Underlying Reasons behind the Common Mechanism

The mechanism of KD can be summarized as "*discarding the dross and selecting the essential*."

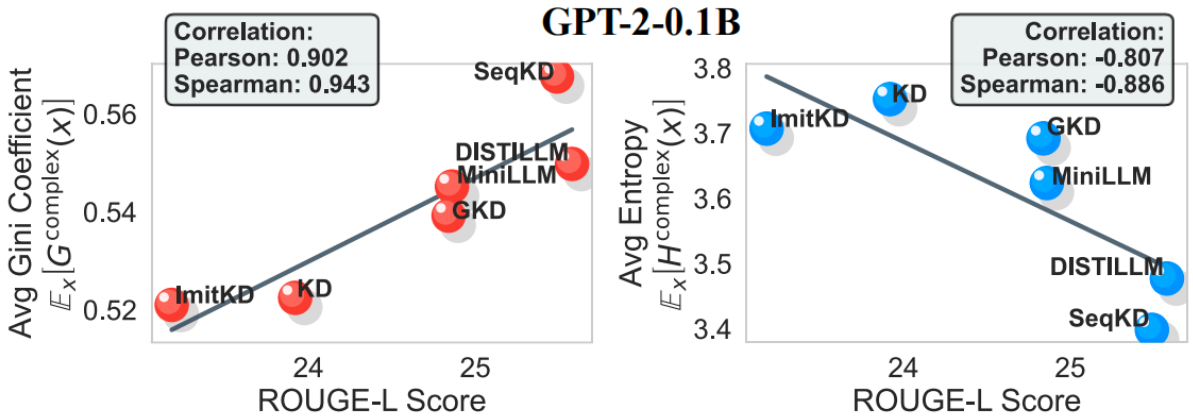
We further identify which specific types of interactions are primarily treated as "dross" to be discarded, and which specific types of interactions are treated as "essential" to be selected. Based on this, we partition interactions into two types: **simple interactions** and **complex interaction**.

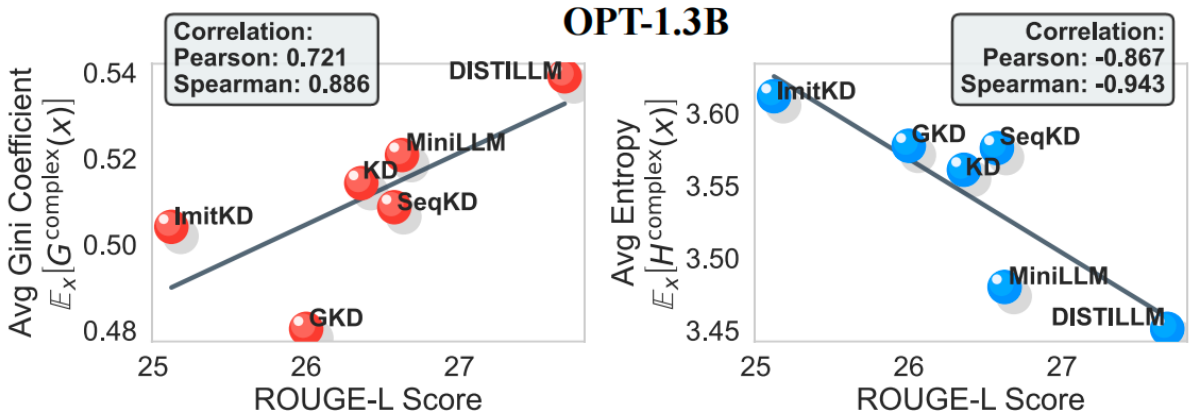
Conclusion: the underlying reason of the common mechanism is that the distillation process makes the student model **retain more simple interactions** to stabilize the student model's general capabilities, while **discarding more complex interactions** to filter out non-generalizable noise.

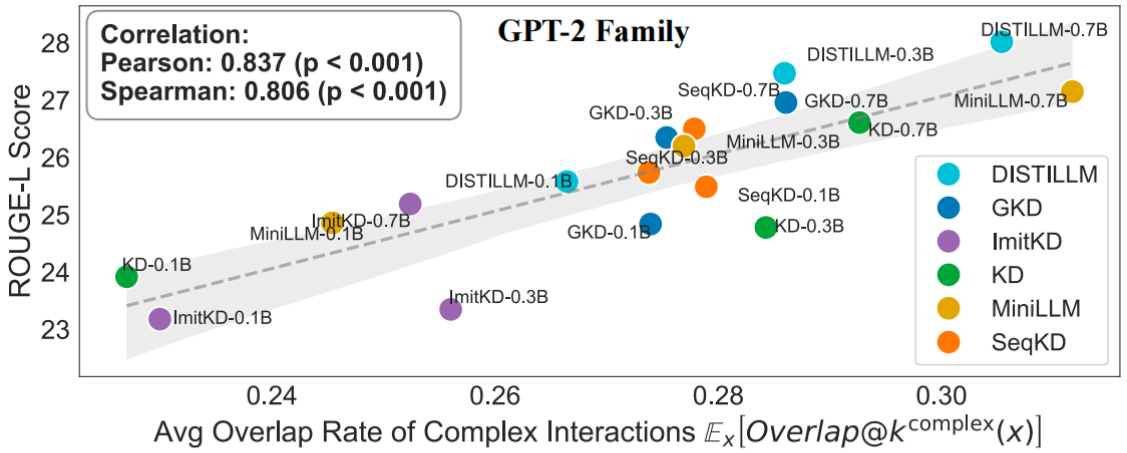


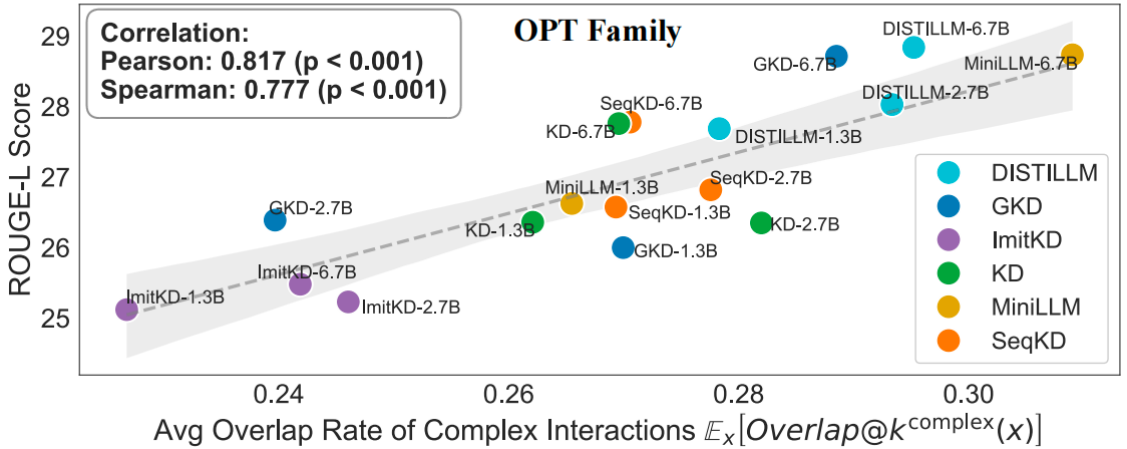
Explaining the Performance Variance across Different KD methods

Conclusion: the performance variance across KD methods is related to their abilities in **handling complex interactions**. In general, superior performance is achieved when student models **exhibit higher sparsity and higher alignment** (overlap rate) with the teacher model in complex interactions.









Guiding KD via Interactions

We introduce a **plug-and-play** loss function called **C**omplex **I**nteraction **P**enalty (CIP), which is designed to suppress complex interactions encoded by the student model during distillation. We define the loss function $\mathcal{L}_{\text{CIP}}$ as follows.

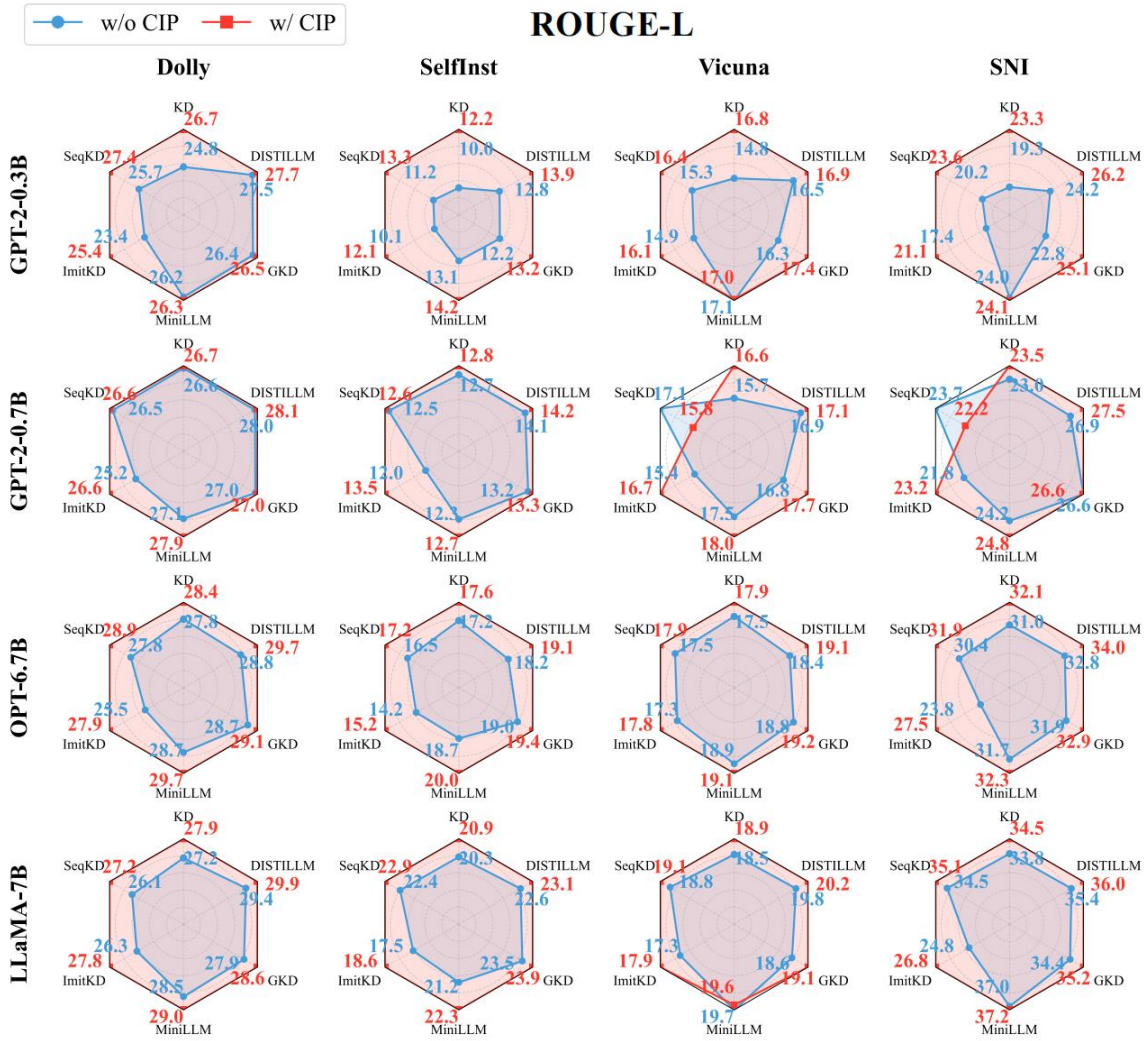
$$\mathcal{L}_{\text{CIP}} = E_x \left[E_{S \in \Omega} \text{complex}[|I(S)|] \right]$$

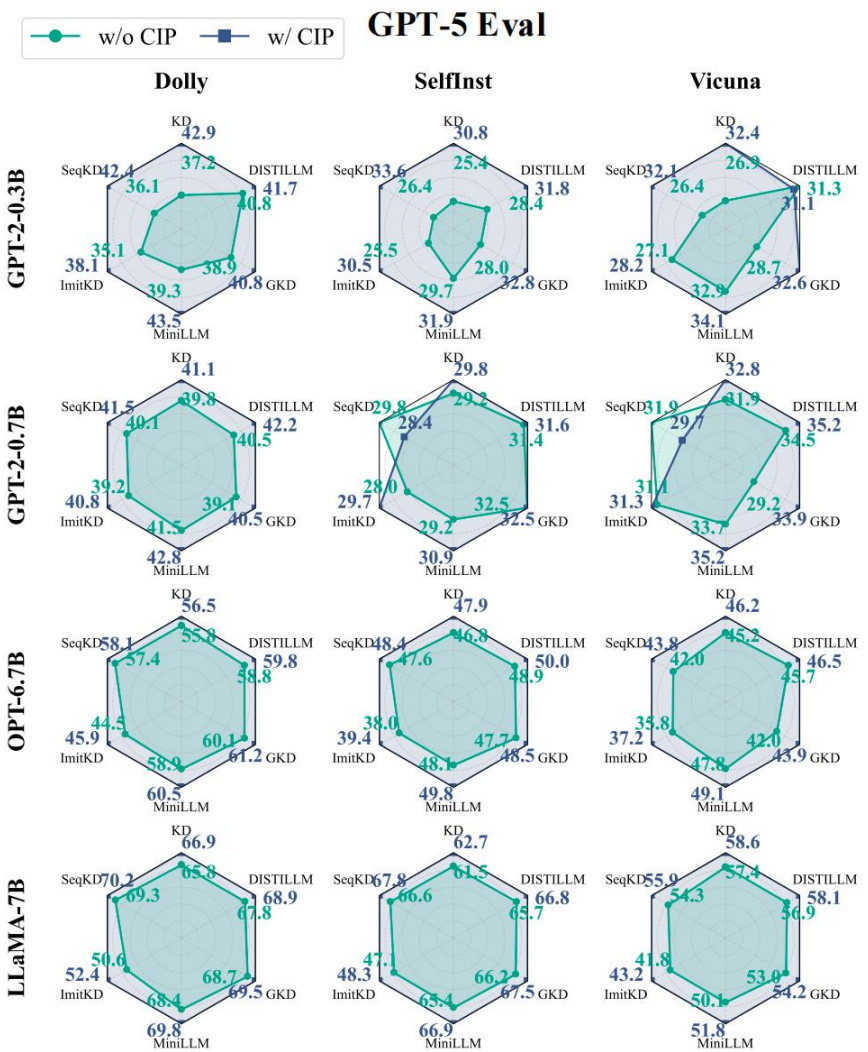
To alleviate computational complexity while preserving efficacy, we adopt a sampling method to approximate $\mathcal{L}_{\text{CIP}}$. Specifically, given an input sequence x with n words indexed by $N = \{1, \ldots, n\}$, we randomly partition the set N into m disjoint subsets $S_1, S_2, \ldots, S_m$, such that $\bigcup_{i=1}^{m} S_i = N$ and $S_i \cap S_j = \emptyset$. Based on this, we define the approximated loss term $\mathcal{L}'_{\text{CIP}}$ as follows.

$$\mathcal{L}'_{\text{CIP}} = E_x \left[E_{K \subseteq \{\mathbb{1}, \ldots, m\}, K \neq \emptyset} \left[\left| I\left(\bigcup_{i \in K} S_i \right) \right| \right] \right]$$

Performance of Adding the CIP Loss

Adding the CIP loss yields improvements in ROUGE-L and GPT-5 scores in most of the settings, regardless of the specific model or KD method.





Thanks For Watching