REEF: Representation Encoding Fingerprints for Large Language Models

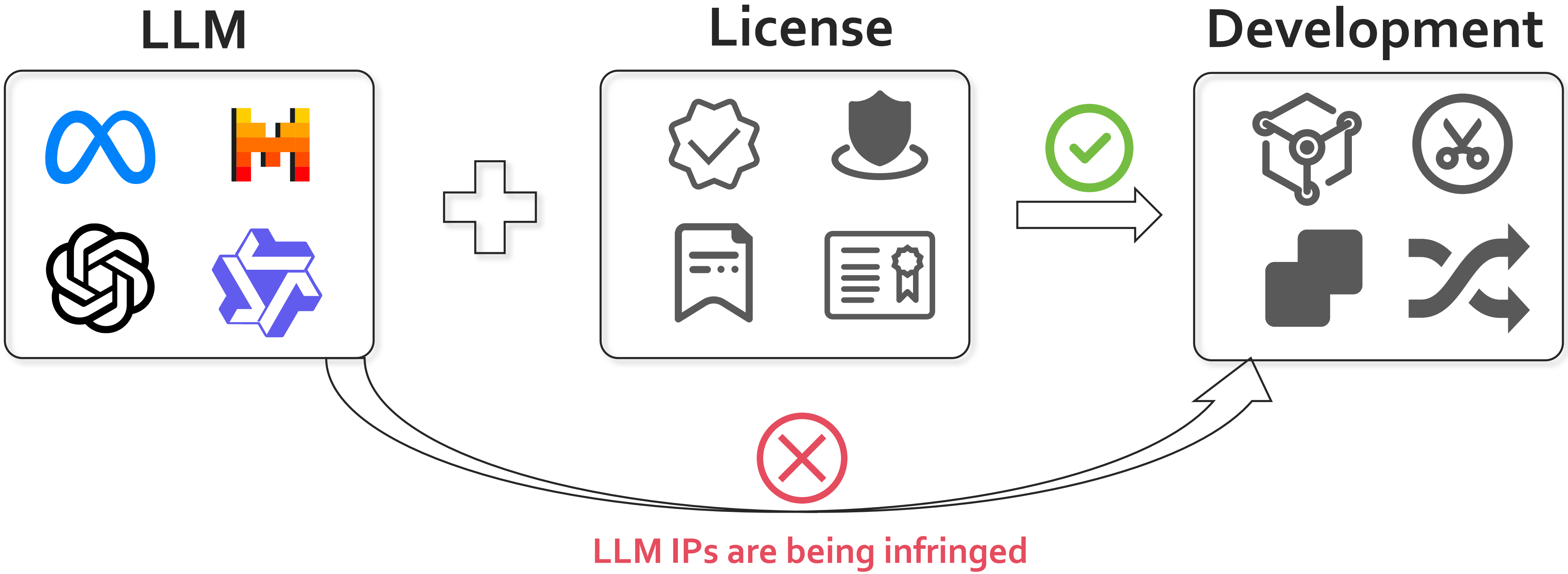
Jie Zhang[1,2]*, Dongrui Liu[1]*, Chen Qian[1,3], Linfeng Zhang[4], Yong Liu[3], Yu Qiao[1], Jing Shao [1]†

[1] Shanghai Artificial Intelligence Laboratory

[2] University of Chinese Academy of Sciences

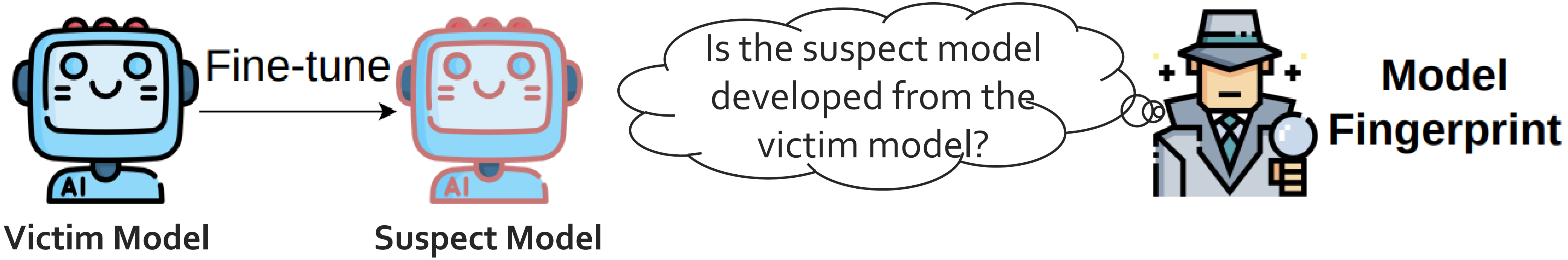
[3] Renmin University of China [4] Shanghai Jiao Tong University

Urging for LLM Intellectual Property Protection



It is urgent to identify *whether the suspect model is a subsequent development of the victim model that serves as the root origin or is developed from scratch*.

LLM Fingerprint for Intellectual Property Protection



Model fingerprinting safeguards model IP by allowing model publishers to authenticate model ownership.

- Harmlessness
- Effectiveness
- Persistence
- Efficiency
- Robustness
- Reliability

Xu, Jiashu, et al. "Instructional fingerprinting of large language models." arXiv preprint arXiv:2401.12255 (2024).

LLM Fingerprint Categories

Injection Fingerprint (*e.g.*, watermarking)

Watermarking artificially inject triggers into the victim model to make it generate specific content for identification.

- Introduce extra training costs
- Impair the model's general capabilities
- Can not be applied to models that have already been open-released

Intrinsic Fingerprint (*e.g.*, model weights)

Weight-based fingerprints calculate the similarity between a suspect model and a victim model's weights for identification.

- Lack of robustness
- Fragile to major changes in weights
- Operations as weight permutations, pruning, and extensive fine-tuning

Askell, Amanda, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

LLM Fingerprint Categories

Injection Fingerprint (*e.g.*, watermarking)

Watermarking artificially inject triggers into the victim model to make it generate specific content for identification.

- Introduce extra training costs
- Impair the model's general capabilities
- Can not be applied to models that have already been open-released

Intrinsic Fingerprint (*e.g.*, model weights)

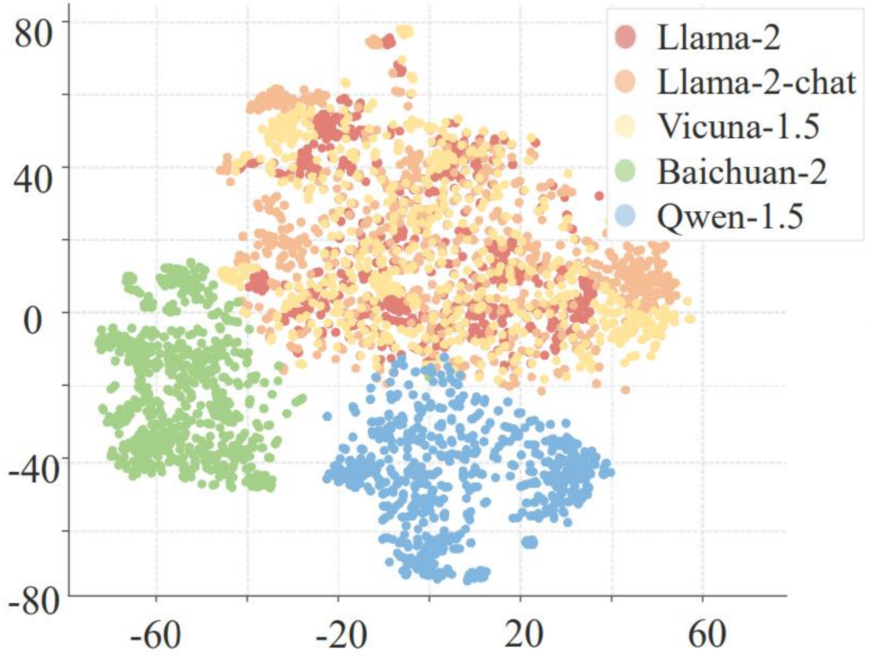
Weight-based fingerprints calculate the similarity between a suspect model and a victim model's weights for identification.

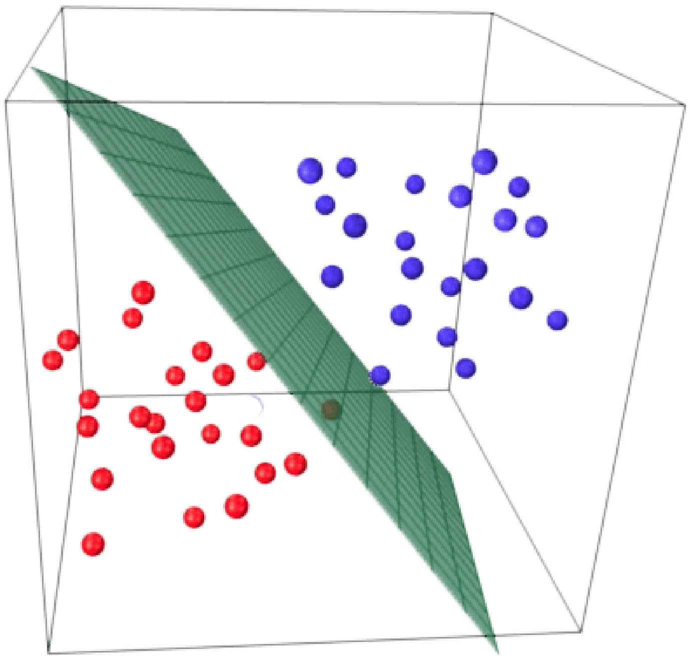
- Lack of robustness
- Fragile to major changes in weights
- Operations as weight permutations, pruning, and extensive fine-tuning

Representation Encoding Fingerprints for LLMs.

Askell, Amanda, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

Potential of Feature Representations as Fingerprints





Discriminability: Feature representations of victim models are similar to representations of the original victim model, while differ from the feature representations of unrelated models

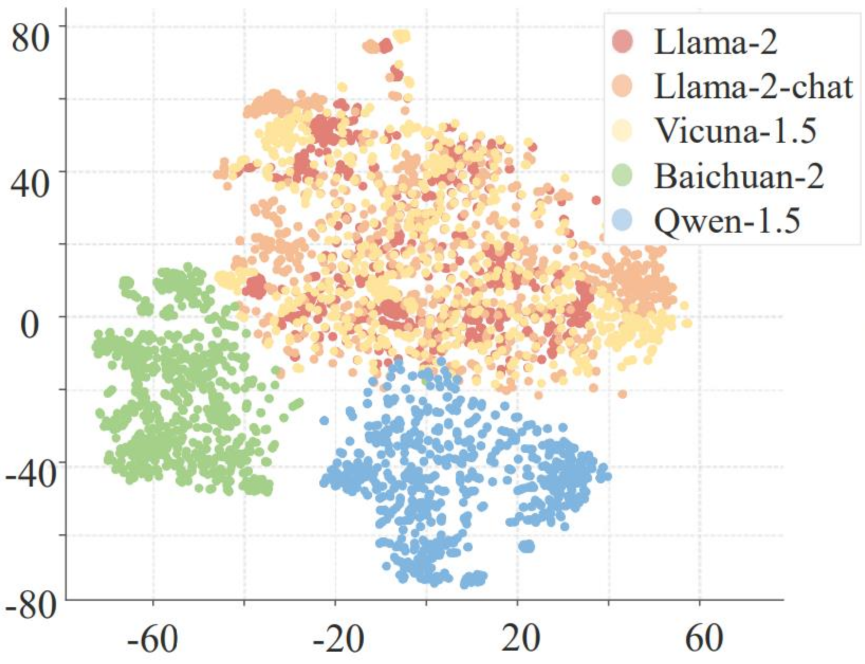
Linear encoding : Some high-level semantic concepts are “linearly” encoded in the representation space of LLMs and can be easily classified, such as safety or unsafety and honest or dishonest[1,2]

Credit: https://mlcourse.ai/book/topic04/topic4_linear_models_part2_logit_likelihood_learning.html

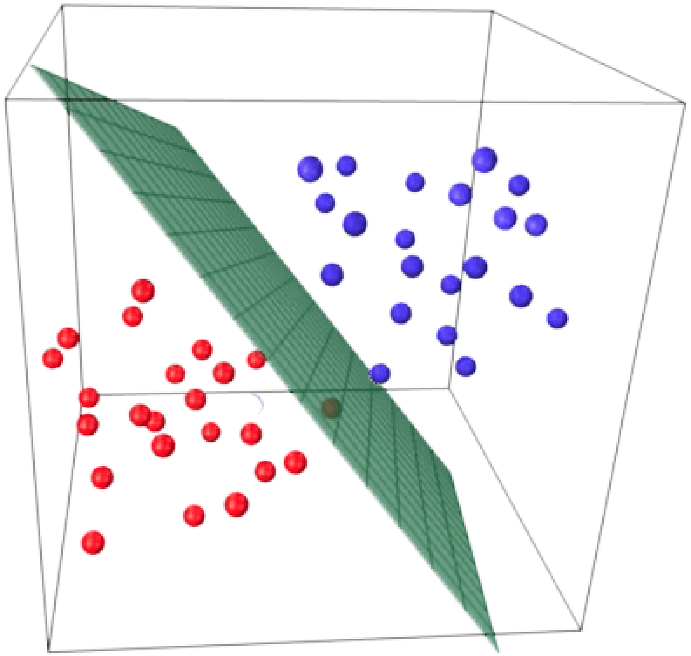
[1] Zou, Andy, et al. "Representation engineering: A top-down approach to ai transparency." arXiv preprint arXiv:2310.01405 (2023).

[2] Qian, Chen, et al. "Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models." arXiv preprint arXiv:2402.19465 (2024).

Potential of Feature Representations as Fingerprints



Discriminability



Linear encoding

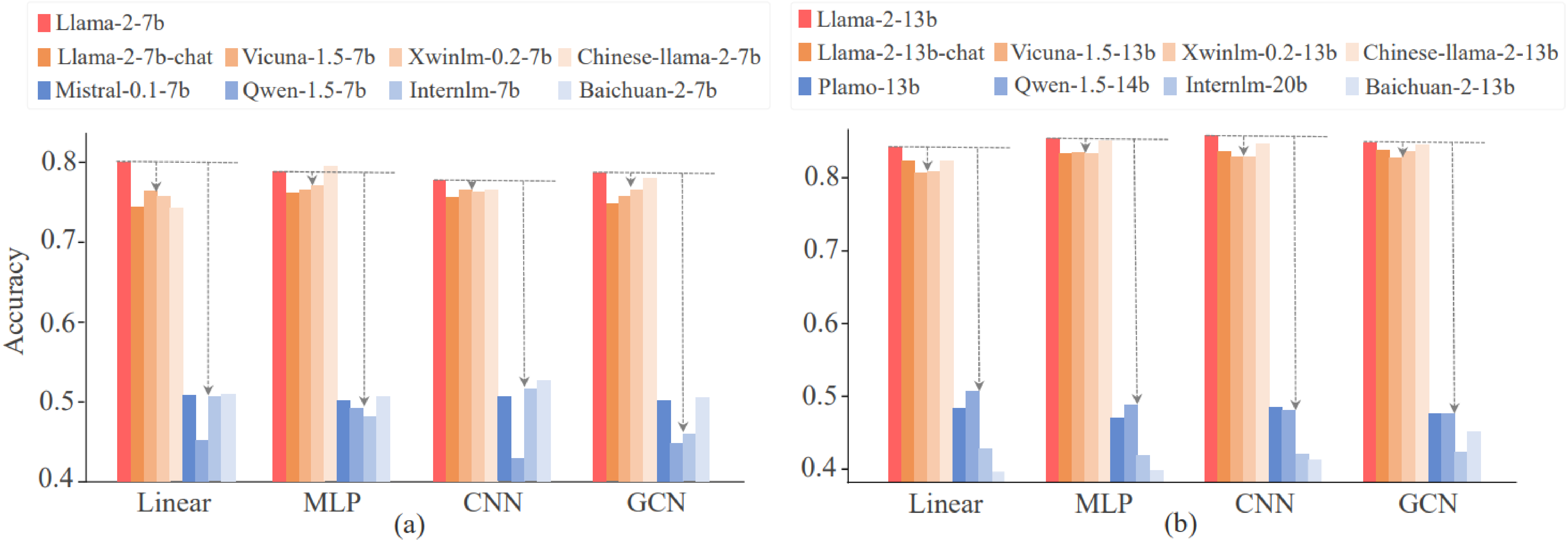
Train a classifier on the high-level semantic representations of the victim model, and examine whether this classifier can be applied to its fine-tuned versions or other different models.

Credit: https://mlcourse.ai/book/topic04/topic4_linear_models_part2_logit_likelihood_learning.html

[1] Zou, Andy, et al. "Representation engineering: A top-down approach to ai transparency." arXiv preprint arXiv:2310.01405 (2023).

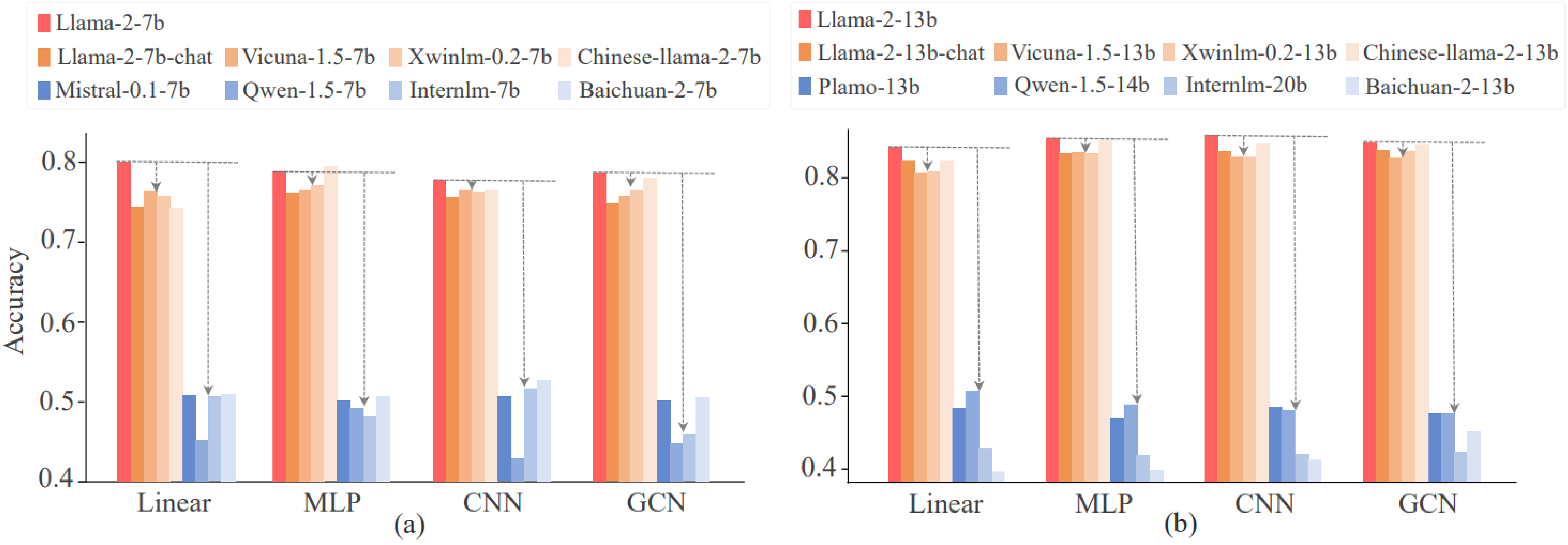
[2] Qian, Chen, et al. "Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models." arXiv preprint arXiv:2402.19465 (2024).

Potential of Feature Representations as Fingerprints: Experiment



Classifiers trained on representations of a victim model can effectively generalize to its variants but not to others.

Potential of Feature Representations as Fingerprints: Experiment



(a) (b)

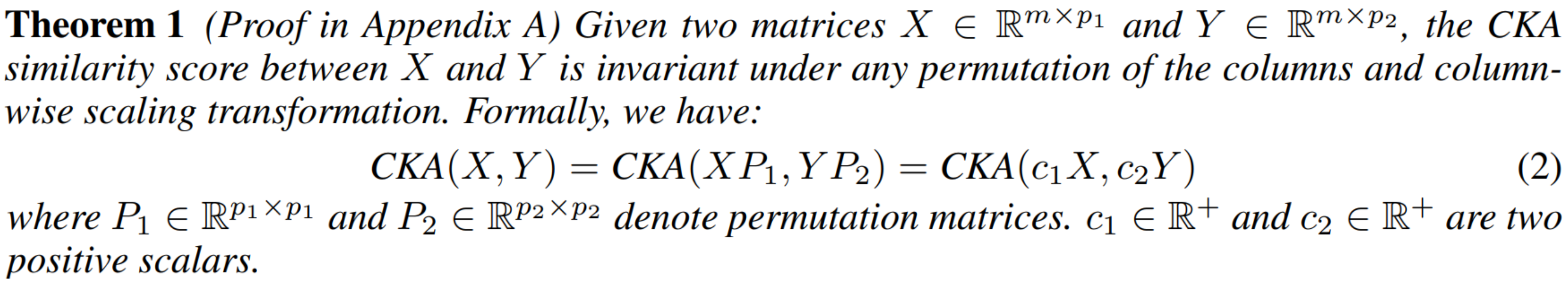
(1) DNNs have fixed input dimensions and cannot be applied to models pruned from the victim model

(2) DNNs are not robust to permutations of the input feature representations

REEF: Representation-based Fingerprinting

REEF identifies whether a suspect model is derived from a root victim model, given the representations of these two models on certain examples.

- $X \in \mathbb{R}^{m*p_1}$: the activation the l-th layer from the suspect model on m examples
- $Y \in \mathbb{R}^{m*p_2}$: the activation the l'-th layer from the suspect model on same m examples
- $s(X, Y)$: a high similarity score indicates that the suspect model is more likely derived from the victim model



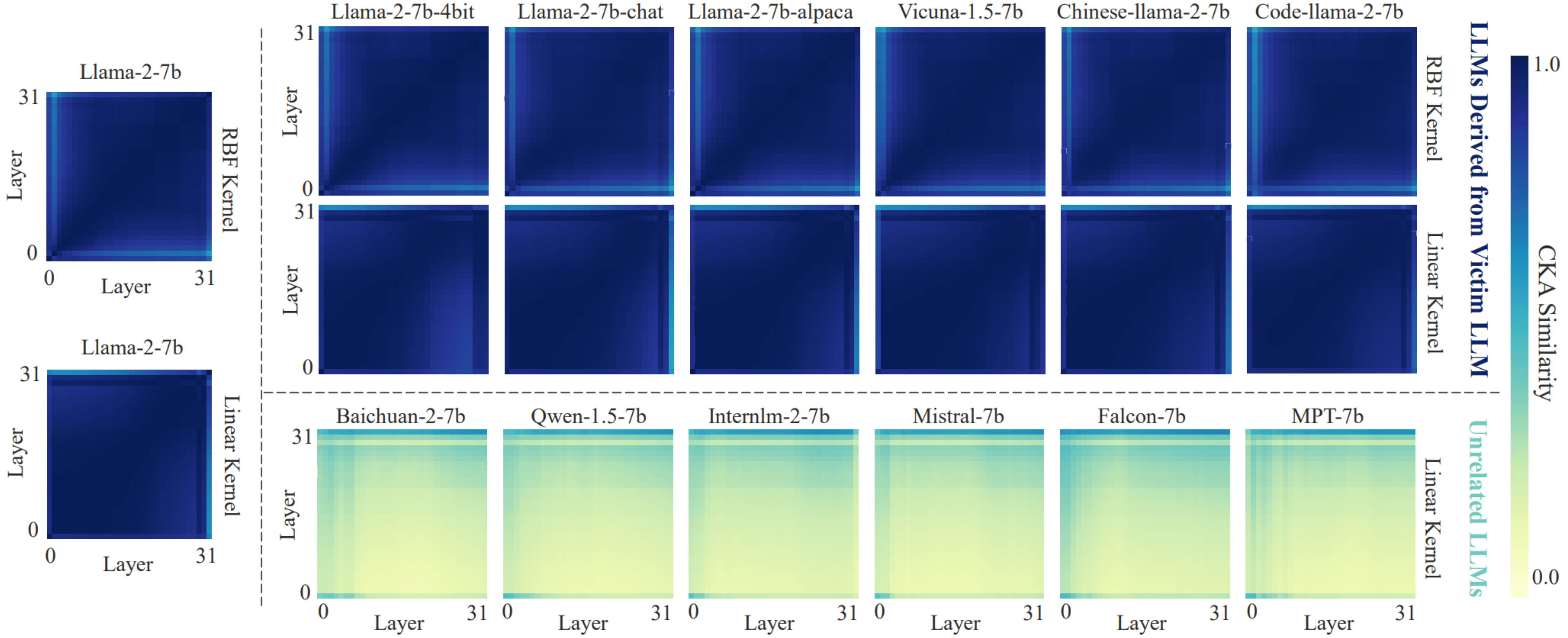
Theorem 1 *(Proof in Appendix A) Given two matrices $X \in \mathbb{R}^{m \times p_1}$ and $Y \in \mathbb{R}^{m \times p_2}$, the CKA similarity score between X and Y is invariant under any permutation of the columns and column-wise scaling transformation. Formally, we have:*

$$CKA(X, Y) = CKA(XP_1, YP_2) = CKA(c_1 X, c_2 Y) \tag{2}$$

where $P_1 \in \mathbb{R}^{p_1 \times p_1}$ and $P_2 \in \mathbb{R}^{p_2 \times p_2}$ denote permutation matrices. $c_1 \in \mathbb{R}^+$ and $c_2 \in \mathbb{R}^+$ are two positive scalars.

Kornblith, Simon, et al. "Similarity of neural network representations revisited." International conference on machine learning. PMLR, 2019.

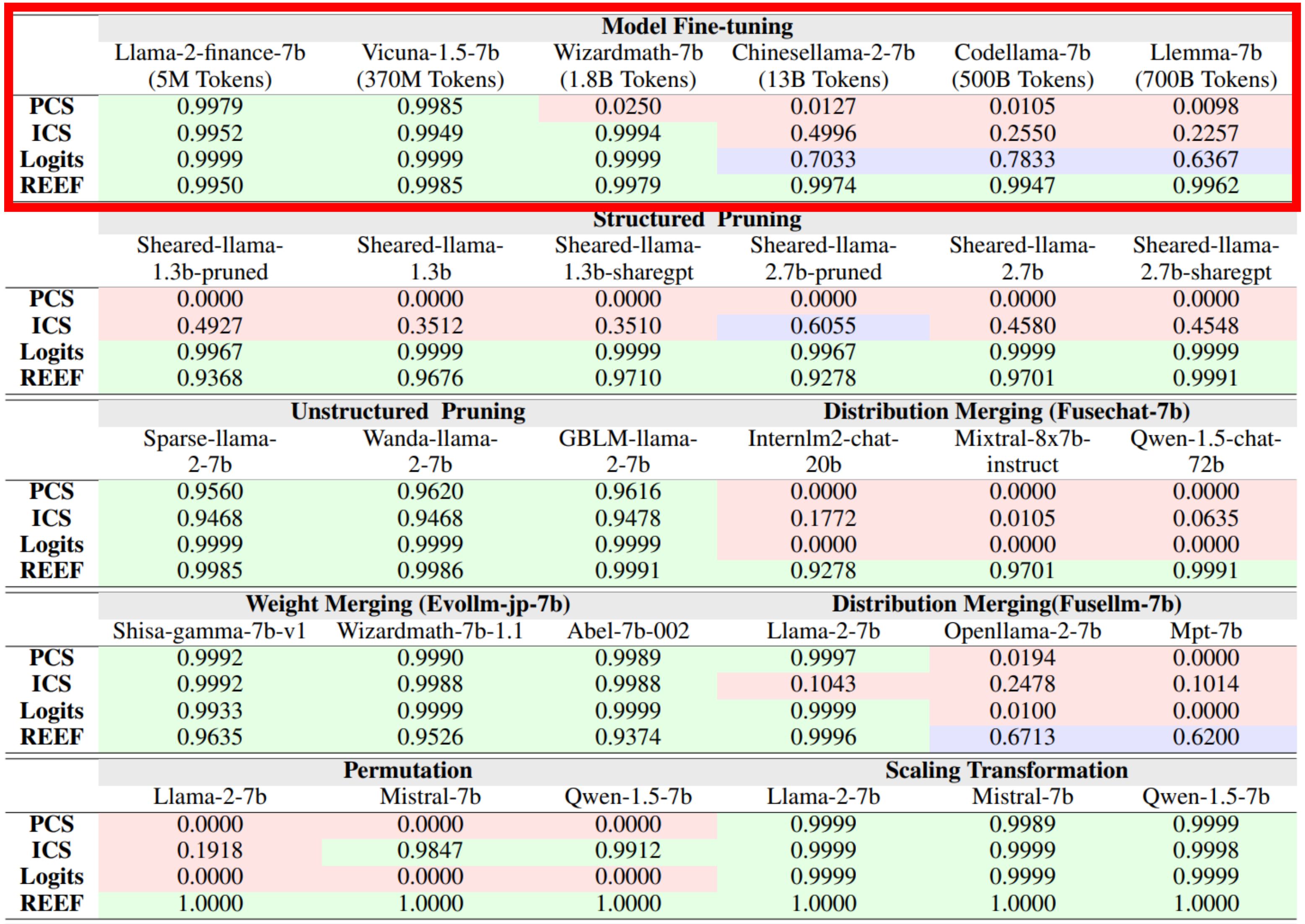
REEF: Effectiveness Verification



(1) **REEF can accurately distinguish** between models derived from the victim model and unrelated models

(2) **Linear and RBF kernels yield similar results** in identifying whether a suspect model is derived from the victim model

REEF: Robustness Verification



	Model Fine-tuning					
	Llama-2-finance-7b (5M Tokens)	Vicuna-1.5-7b (370M Tokens)	Wizardmath-7b (1.8B Tokens)	Chinesellama-2-7b (13B Tokens)	Codellama-7b (500B Tokens)	Llemma-7b (700B Tokens)
PCS	0.9979	0.9985	0.0250	0.0127	0.0105	0.0098
ICS	0.9952	0.9949	0.9994	0.4996	0.2550	0.2257
Logits	0.9999	0.9999	0.9999	0.7033	0.7833	0.6367
REEF	0.9950	0.9985	0.9979	0.9974	0.9947	0.9962

	Structured Pruning					
	Sheared-llama-1.3b-pruned	Sheared-llama-1.3b	Sheared-llama-1.3b-sharegpt	Sheared-llama-2.7b-pruned	Sheared-llama-2.7b	Sheared-llama-2.7b-sharegpt
PCS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ICS	0.4927	0.3512	0.3510	0.6055	0.4580	0.4548
Logits	0.9967	0.9999	0.9999	0.9967	0.9999	0.9999
REEF	0.9368	0.9676	0.9710	0.9278	0.9701	0.9991

	Unstructured Pruning			Distribution Merging (Fusechat-7b)		
	Sparse-llama-2-7b	Wanda-llama-2-7b	GBLM-llama-2-7b	Internlm2-chat-20b	Mixtral-8x7b-instruct	Qwen-1.5-chat-72b
PCS	0.9560	0.9620	0.9616	0.0000	0.0000	0.0000
ICS	0.9468	0.9468	0.9478	0.1772	0.0105	0.0635
Logits	0.9999	0.9999	0.9999	0.0000	0.0000	0.0000
REEF	0.9985	0.9986	0.9991	0.9278	0.9701	0.9991

	Weight Merging (Evollm-jp-7b)			Distribution Merging(Fusellm-7b)		
	Shisa-gamma-7b-v1	Wizardmath-7b-1.1	Abel-7b-002	Llama-2-7b	Openllama-2-7b	Mpt-7b
PCS	0.9992	0.9990	0.9989	0.9997	0.0194	0.0000
ICS	0.9992	0.9988	0.9988	0.1043	0.2478	0.1014
Logits	0.9933	0.9999	0.9999	0.9999	0.0100	0.0000
REEF	0.9635	0.9526	0.9374	0.9996	0.6713	0.6200

	Permutation			Scaling Transformation		
	Llama-2-7b	Mistral-7b	Qwen-1.5-7b	Llama-2-7b	Mistral-7b	Qwen-1.5-7b
PCS	0.0000	0.0000	0.0000	0.9999	0.9989	0.9999
ICS	0.1918	0.9847	0.9912	0.9999	0.9999	0.9998
Logits	0.0000	0.0000	0.0000	0.9999	0.9999	0.9999
REEF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

🟩 similarity greater than 0.8 🟪 similarity between 0.5 and 0.8 🟥 similarity less than 0.5

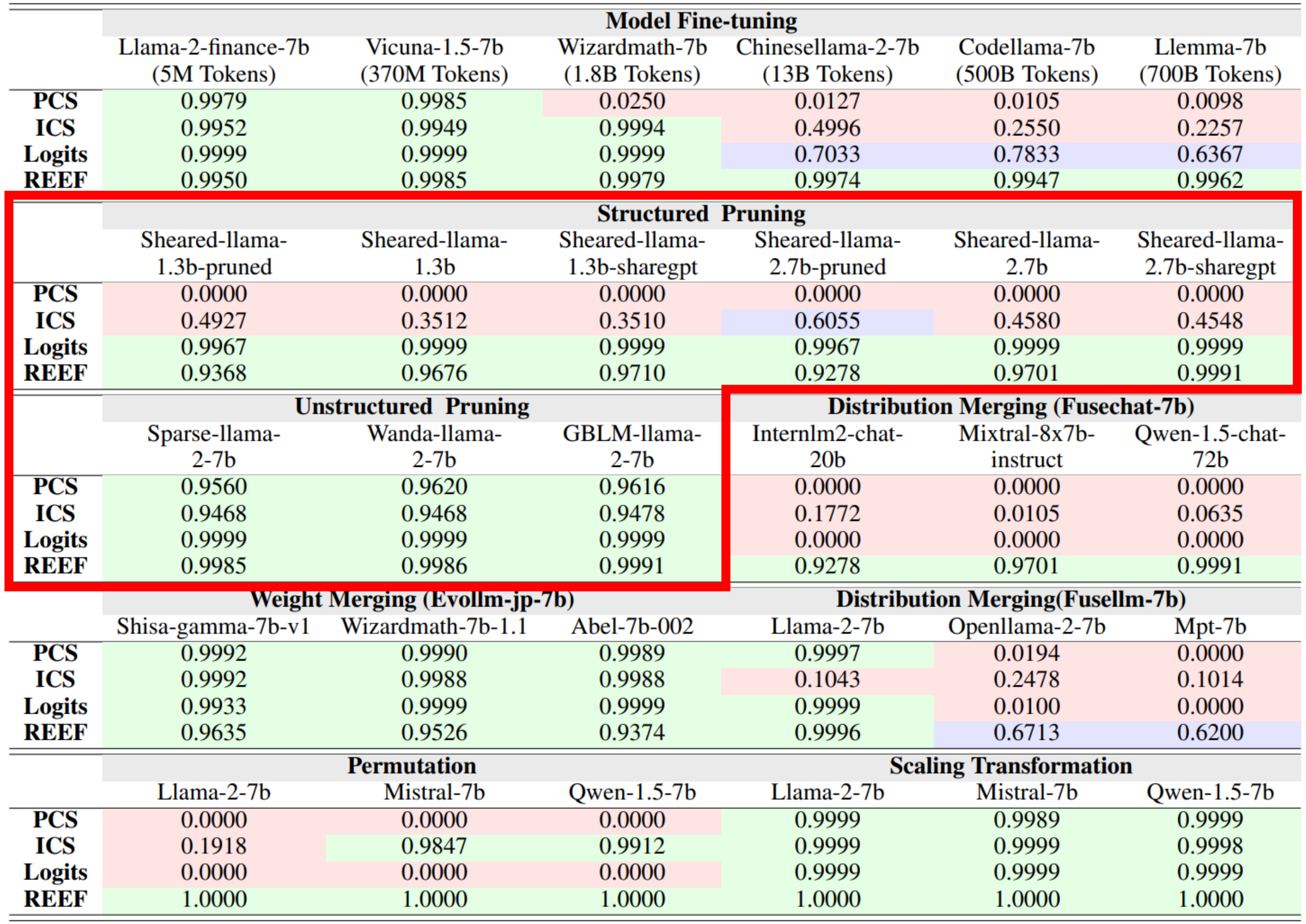
REEF is robustness to extensive fine-tuning

REEF is robust to various pruning strategies; even up to 90% pruning ratio

REEF is robust across both weight and distribution merging scenarios

REEF is invariant and robust to any column-wise permutations and scaling transformations

REEF: Robustness Verification



	Model Fine-tuning					
	Llama-2-finance-7b (5M Tokens)	Vicuna-1.5-7b (370M Tokens)	Wizardmath-7b (1.8B Tokens)	Chinesellama-2-7b (13B Tokens)	Codellama-7b (500B Tokens)	Llemma-7b (700B Tokens)
PCS	0.9979	0.9985	0.0250	0.0127	0.0105	0.0098
ICS	0.9952	0.9949	0.9994	0.4996	0.2550	0.2257
Logits	0.9999	0.9999	0.9999	0.7033	0.7833	0.6367
REEF	0.9950	0.9985	0.9979	0.9974	0.9947	0.9962
	Structured Pruning					
	Sheared-llama-1.3b-pruned	Sheared-llama-1.3b	Sheared-llama-1.3b-sharegpt	Sheared-llama-2.7b-pruned	Sheared-llama-2.7b	Sheared-llama-2.7b-sharegpt
PCS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ICS	0.4927	0.3512	0.3510	0.6055	0.4580	0.4548
Logits	0.9967	0.9999	0.9999	0.9967	0.9999	0.9999
REEF	0.9368	0.9676	0.9710	0.9278	0.9701	0.9991
	Unstructured Pruning			**Distribution Merging (Fusechat-7b)**		
	Sparse-llama-2-7b	Wanda-llama-2-7b	GBLM-llama-2-7b	Internlm2-chat-20b	Mixtral-8x7b-instruct	Qwen-1.5-chat-72b
PCS	0.9560	0.9620	0.9616	0.0000	0.0000	0.0000
ICS	0.9468	0.9468	0.9478	0.1772	0.0105	0.0635
Logits	0.9999	0.9999	0.9999	0.0000	0.0000	0.0000
REEF	0.9985	0.9986	0.9991	0.9278	0.9701	0.9991
	Weight Merging (Evollm-jp-7b)			**Distribution Merging(Fusellm-7b)**		
	Shisa-gamma-7b-v1	Wizardmath-7b-1.1	Abel-7b-002	Llama-2-7b	Openllama-2-7b	Mpt-7b
PCS	0.9992	0.9990	0.9989	0.9997	0.0194	0.0000
ICS	0.9992	0.9988	0.9988	0.1043	0.2478	0.1014
Logits	0.9933	0.9999	0.9999	0.9999	0.0100	0.0000
REEF	0.9635	0.9526	0.9374	0.9996	0.6713	0.6200
	Permutation			**Scaling Transformation**		
	Llama-2-7b	Mistral-7b	Qwen-1.5-7b	Llama-2-7b	Mistral-7b	Qwen-1.5-7b
PCS	0.0000	0.0000	0.0000	0.9999	0.9989	0.9999
ICS	0.1918	0.9847	0.9912	0.9999	0.9999	0.9998
Logits	0.0000	0.0000	0.0000	0.9999	0.9999	0.9999
REEF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

🟩 similarity greater than 0.8 🟪 similarity between 0.5 and 0.8 🟥 similarity less than 0.5

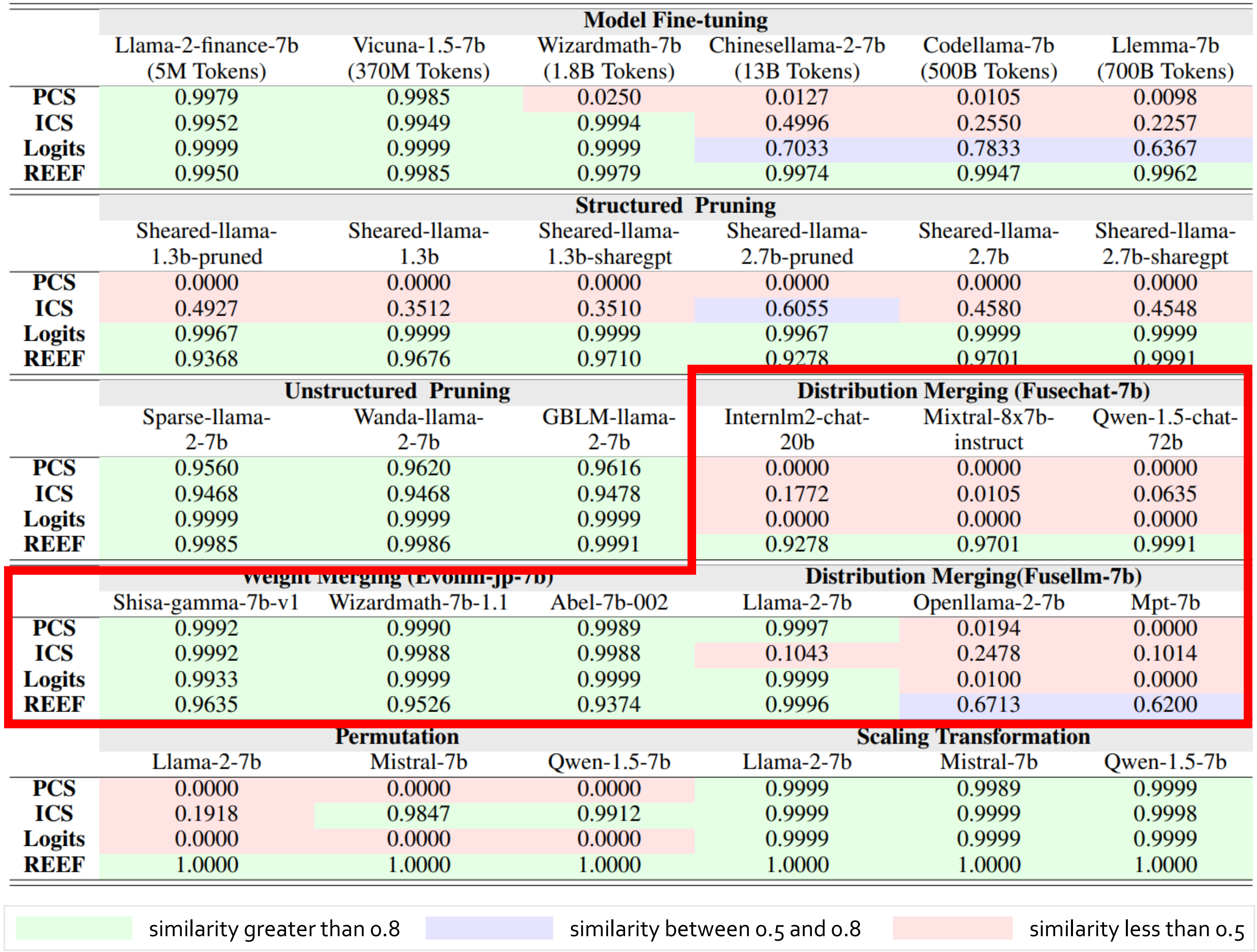
REEF is robustness to extensive fine-tuning

REEF is robust to various pruning strategies; even up to 90% pruning ratio

REEF is robust across both weight and distribution merging scenarios

REEF is invariant and robust to any column-wise permutations and scaling transformations

REEF: Robustness Verification



	Model Fine-tuning					
	Llama-2-finance-7b (5M Tokens)	Vicuna-1.5-7b (370M Tokens)	Wizardmath-7b (1.8B Tokens)	Chinesellama-2-7b (13B Tokens)	Codellama-7b (500B Tokens)	Llemma-7b (700B Tokens)
PCS	0.9979	0.9985	0.0250	0.0127	0.0105	0.0098
ICS	0.9952	0.9949	0.9994	0.4996	0.2550	0.2257
Logits	0.9999	0.9999	0.9999	0.7033	0.7833	0.6367
REEF	0.9950	0.9985	0.9979	0.9974	0.9947	0.9962
	Structured Pruning					
	Sheared-llama-1.3b-pruned	Sheared-llama-1.3b	Sheared-llama-1.3b-sharegpt	Sheared-llama-2.7b-pruned	Sheared-llama-2.7b	Sheared-llama-2.7b-sharegpt
PCS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ICS	0.4927	0.3512	0.3510	0.6055	0.4580	0.4548
Logits	0.9967	0.9999	0.9999	0.9967	0.9999	0.9999
REEF	0.9368	0.9676	0.9710	0.9278	0.9701	0.9991
	Unstructured Pruning			**Distribution Merging (Fusechat-7b)**		
	Sparse-llama-2-7b	Wanda-llama-2-7b	GBLM-llama-2-7b	Internlm2-chat-20b	Mixtral-8x7b-instruct	Qwen-1.5-chat-72b
PCS	0.9560	0.9620	0.9616	0.0000	0.0000	0.0000
ICS	0.9468	0.9468	0.9478	0.1772	0.0105	0.0635
Logits	0.9999	0.9999	0.9999	0.0000	0.0000	0.0000
REEF	0.9985	0.9986	0.9991	0.9278	0.9701	0.9991
	Weight Merging (Evollm-jp-7b)			**Distribution Merging(Fusellm-7b)**		
	Shisa-gamma-7b-v1	Wizardmath-7b-1.1	Abel-7b-002	Llama-2-7b	Openllama-2-7b	Mpt-7b
PCS	0.9992	0.9990	0.9989	0.9997	0.0194	0.0000
ICS	0.9992	0.9988	0.9988	0.1043	0.2478	0.1014
Logits	0.9933	0.9999	0.9999	0.9999	0.0100	0.0000
REEF	0.9635	0.9526	0.9374	0.9996	0.6713	0.6200
	Permutation			**Scaling Transformation**		
	Llama-2-7b	Mistral-7b	Qwen-1.5-7b	Llama-2-7b	Mistral-7b	Qwen-1.5-7b
PCS	0.0000	0.0000	0.0000	0.9999	0.9989	0.9999
ICS	0.1918	0.9847	0.9912	0.9999	0.9999	0.9998
Logits	0.0000	0.0000	0.0000	0.9999	0.9999	0.9999
REEF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

🟩 similarity greater than 0.8 🟦 similarity between 0.5 and 0.8 🟥 similarity less than 0.5

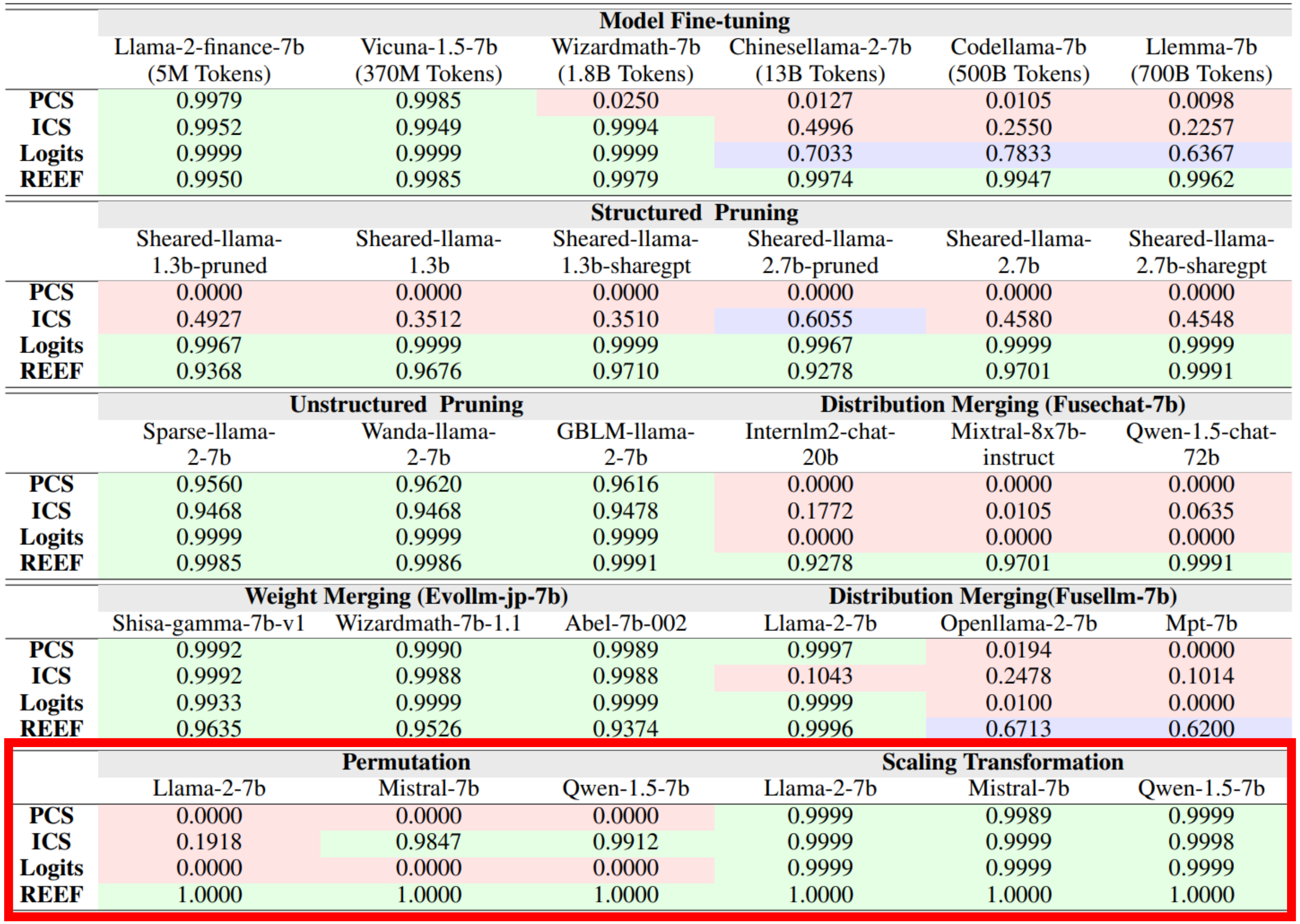
REEF is robustness to extensive fine-tuning

REEF is robust to various pruning strategies; even up to 90% pruning ratio

REEF is robust across both weight and distribution merging scenarios

REEF is invariant and robust to any column-wise permutations and scaling transformations

REEF: Robustness Verification



	Model Fine-tuning					
	Llama-2-finance-7b (5M Tokens)	Vicuna-1.5-7b (370M Tokens)	Wizardmath-7b (1.8B Tokens)	Chinesellama-2-7b (13B Tokens)	Codellama-7b (500B Tokens)	Llemma-7b (700B Tokens)
PCS	0.9979	0.9985	0.0250	0.0127	0.0105	0.0098
ICS	0.9952	0.9949	0.9994	0.4996	0.2550	0.2257
Logits	0.9999	0.9999	0.9999	0.7033	0.7833	0.6367
REEF	0.9950	0.9985	0.9979	0.9974	0.9947	0.9962
	Structured Pruning					
	Sheared-llama-1.3b-pruned	Sheared-llama-1.3b	Sheared-llama-1.3b-sharegpt	Sheared-llama-2.7b-pruned	Sheared-llama-2.7b	Sheared-llama-2.7b-sharegpt
PCS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ICS	0.4927	0.3512	0.3510	0.6055	0.4580	0.4548
Logits	0.9967	0.9999	0.9999	0.9967	0.9999	0.9999
REEF	0.9368	0.9676	0.9710	0.9278	0.9701	0.9991
	Unstructured Pruning			**Distribution Merging (Fusechat-7b)**		
	Sparse-llama-2-7b	Wanda-llama-2-7b	GBLM-llama-2-7b	Internlm2-chat-20b	Mixtral-8x7b-instruct	Qwen-1.5-chat-72b
PCS	0.9560	0.9620	0.9616	0.0000	0.0000	0.0000
ICS	0.9468	0.9468	0.9478	0.1772	0.0105	0.0635
Logits	0.9999	0.9999	0.9999	0.0000	0.0000	0.0000
REEF	0.9985	0.9986	0.9991	0.9278	0.9701	0.9991
	Weight Merging (Evollm-jp-7b)			**Distribution Merging(Fusellm-7b)**		
	Shisa-gamma-7b-v1	Wizardmath-7b-1.1	Abel-7b-002	Llama-2-7b	Openllama-2-7b	Mpt-7b
PCS	0.9992	0.9990	0.9989	0.9997	0.0194	0.0000
ICS	0.9992	0.9988	0.9988	0.1043	0.2478	0.1014
Logits	0.9933	0.9999	0.9999	0.9999	0.0100	0.0000
REEF	0.9635	0.9526	0.9374	0.9996	0.6713	0.6200
	Permutation			**Scaling Transformation**		
	Llama-2-7b	Mistral-7b	Qwen-1.5-7b	Llama-2-7b	Mistral-7b	Qwen-1.5-7b
PCS	0.0000	0.0000	0.0000	0.9999	0.9989	0.9999
ICS	0.1918	0.9847	0.9912	0.9999	0.9999	0.9998
Logits	0.0000	0.0000	0.0000	0.9999	0.9999	0.9999
REEF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

🟩 similarity greater than 0.8 🟦 similarity between 0.5 and 0.8 🟥 similarity less than 0.5

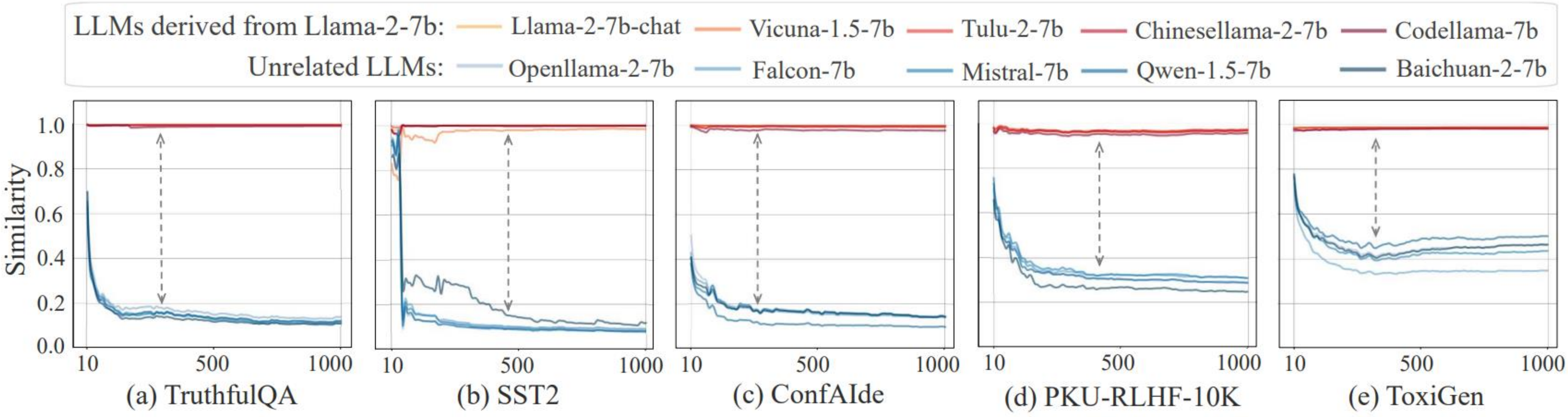
REEF is robustness to extensive fine-tuning

REEF is robust to various pruning strategies; even up to 90% pruning ratio

REEF is robust across both weight and distribution merging scenarios

REEF is invariant and robust to any column-wise permutations and scaling transformations

REEF: Ablation Study



(a) TruthfulQA (b) SST2 (c) ConfAIde (d) PKU-RLHF-10K (e) ToxiGen

(1) REEF is highly efficient regarding the number of samples required for robust model fingerprinting

(2) REEF is effective across various datasets

Takeaways

- **Training-free**: REEF requires no additional training and has no impact on model performance, unlike traditional watermarking methods which may degrade model capabilities.
- **Robustness**: REEF remains effective even after fine-tuning, pruning, or model merging, whereas existing fingerprinting methods (e.g., weight-based) tend to fail.
- **Inference-time Applicability**: REEF can be applied to any LLM, even models that have already been released, and it is impossible to bypass

Thanks for your attention!