Knowledge Distillation Detection for Open-weights Models

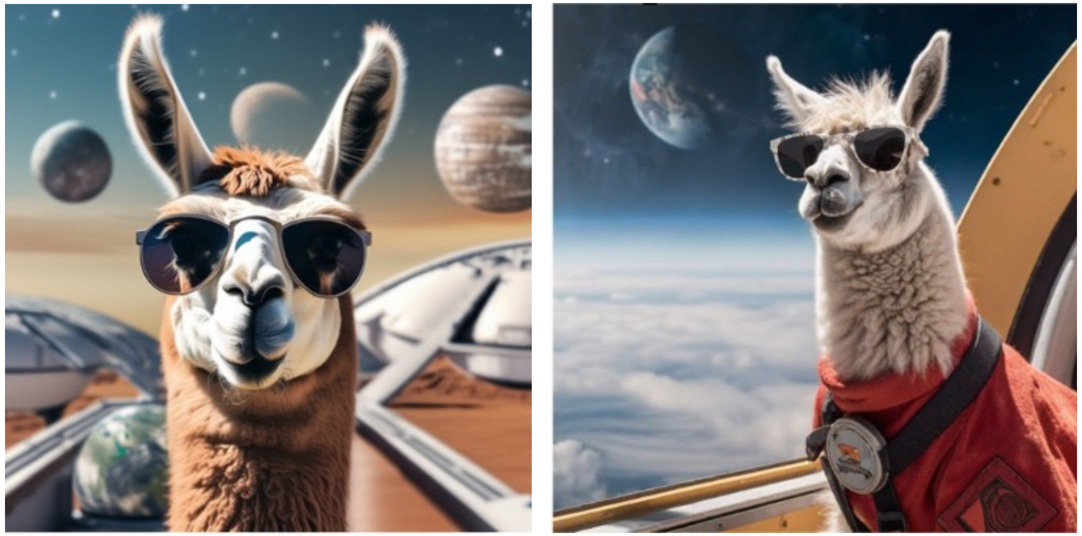
Qin Shi Amber Yijia Zheng Qifan Song Raymond A. Yeh

Purdue University

PURDUE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

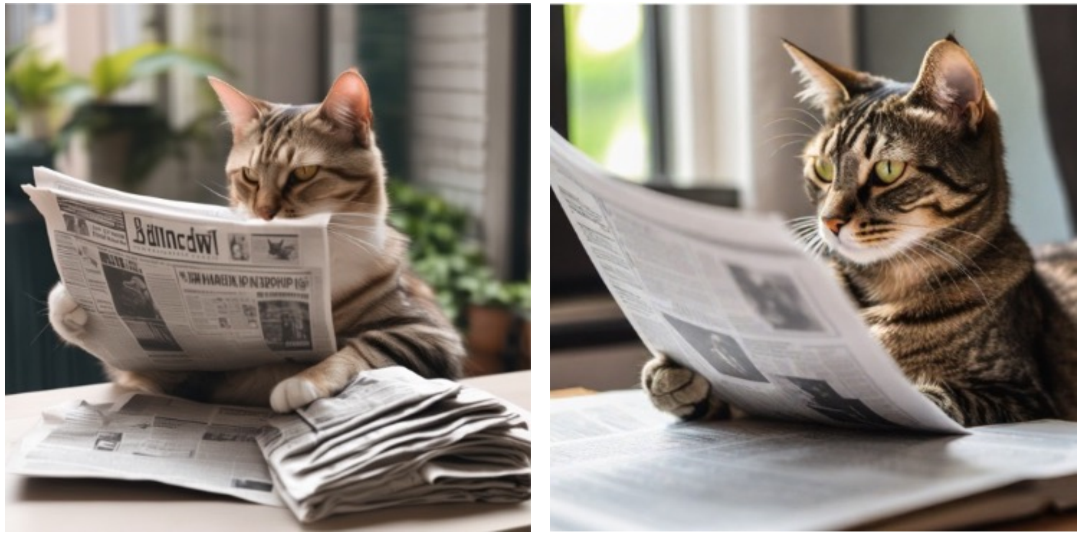
Advancement of Distillation Method



Teacher

Student

A photo of llama wearing sunglasses standing on the deck of a spaceship with the Earth in the background.



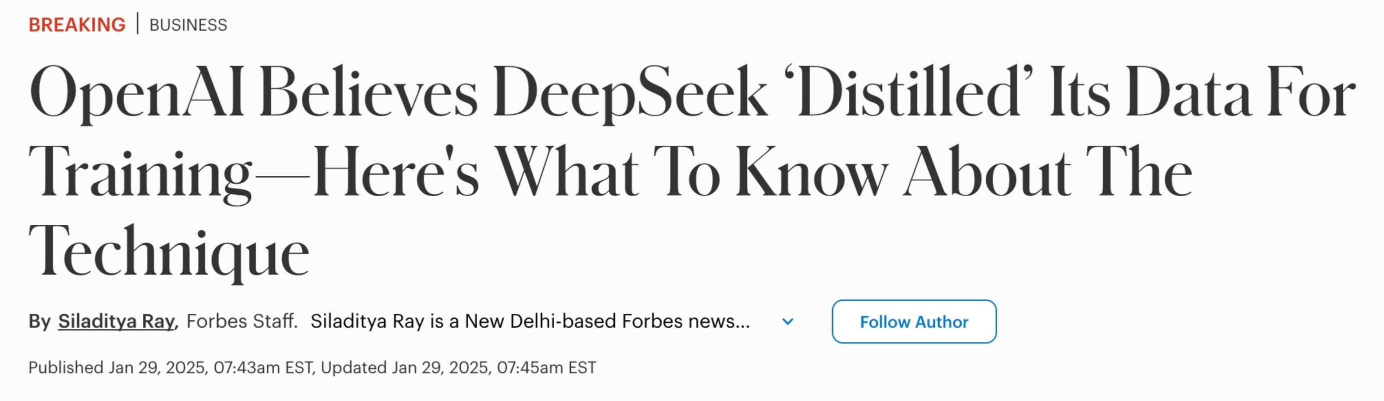
Teacher

Student

A cat reading a newspaper

The Risk of Distillation

Unauthorized knowledge distillation can enable the cloning of open-source models



BREAKING | BUSINESS

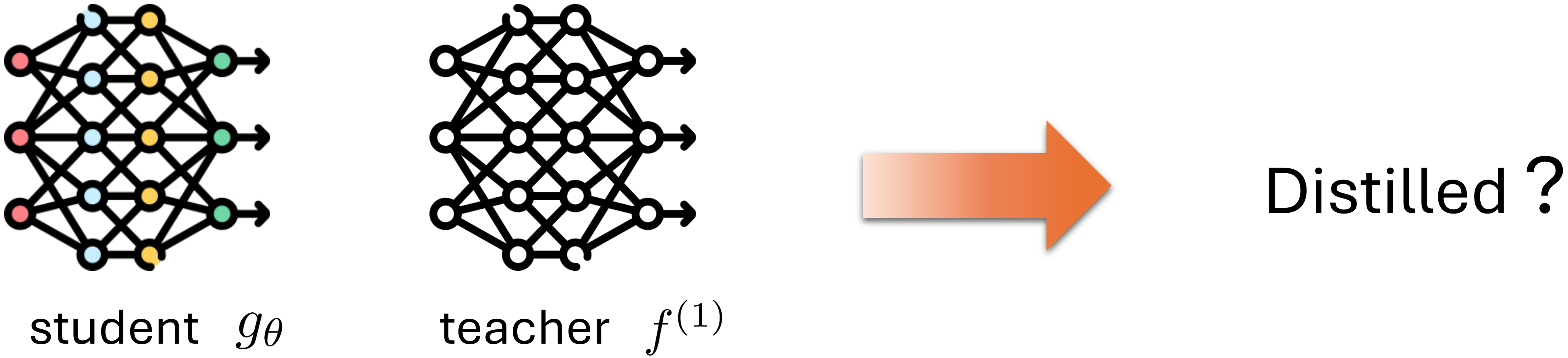
OpenAI Believes DeepSeek 'Distilled' Its Data For Training—Here's What To Know About The Technique

By Siladitya Ray, Forbes Staff. Siladitya Ray is a New Delhi-based Forbes news...

Published Jan 29, 2025, 07:43am EST, Updated Jan 29, 2025, 07:45am EST

Raising concerns about data leakage and model attribution

Knowledge Distillation Detection



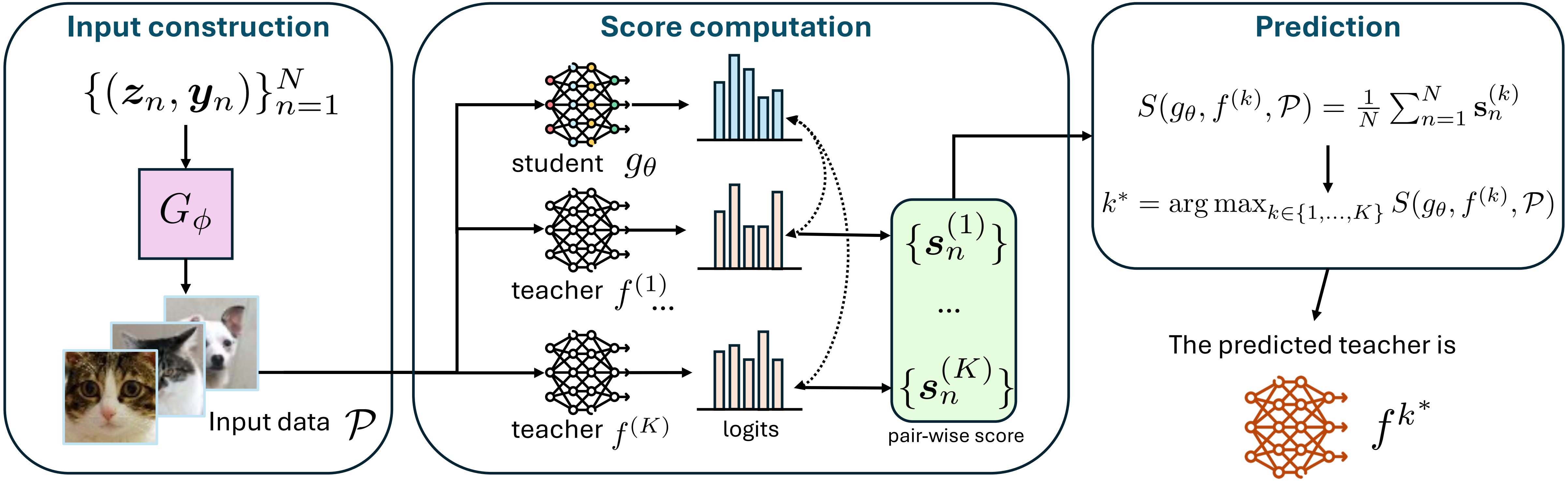
student g_θ

teacher $f^{(1)}$

Distilled ?

- Focus on open-weight student models
- Only the teacher’s API are available
- Without knowing training data, distillation method

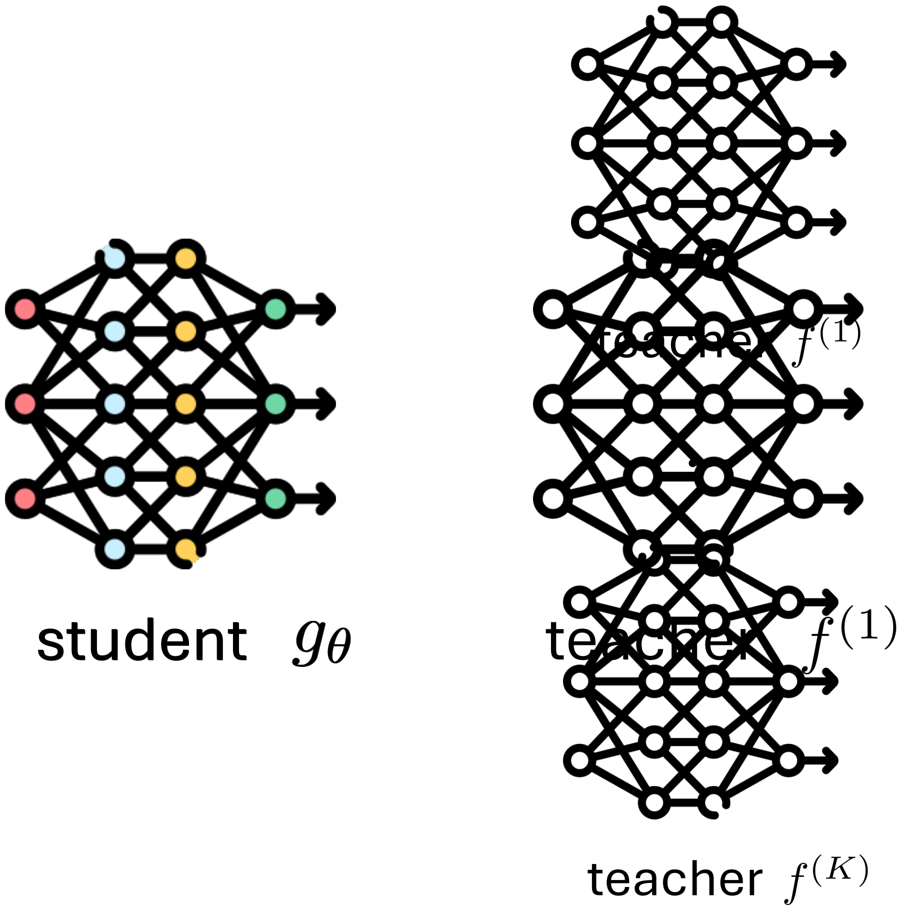
Our Framework

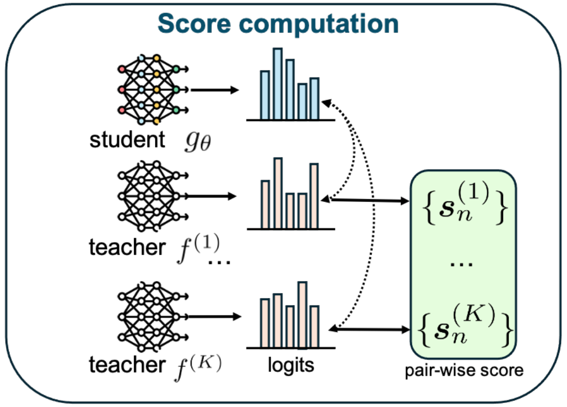

Input construction
$\{(\boldsymbol{z}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$
G_ϕ
Input data $\mathcal{P}$
Score computation
student g_θ
teacher $f^{(1)}$
...
teacher $f^{(K)}$
logits
$\{\boldsymbol{s}_n^{(1)}\}$
...
$\{\boldsymbol{s}_n^{(K)}\}$
pair-wise score
Prediction
$S(g_\theta, f^{(k)}, \mathcal{P}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{s}_n^{(k)}$
$k^* = \arg\max_{k\in\{1,\ldots,K\}} S(g_\theta, f^{(k)}, \mathcal{P})$
The predicted teacher is
f^{k^*}

Input construction: data synthesis; designed prompt; ......

Score computation: KL divergence; ACS; CLIP; test statistics; ......

Multiple Choice Formula





Who is the
Distilled ?
teacher ?

Evaluation Metrics

Accuracy

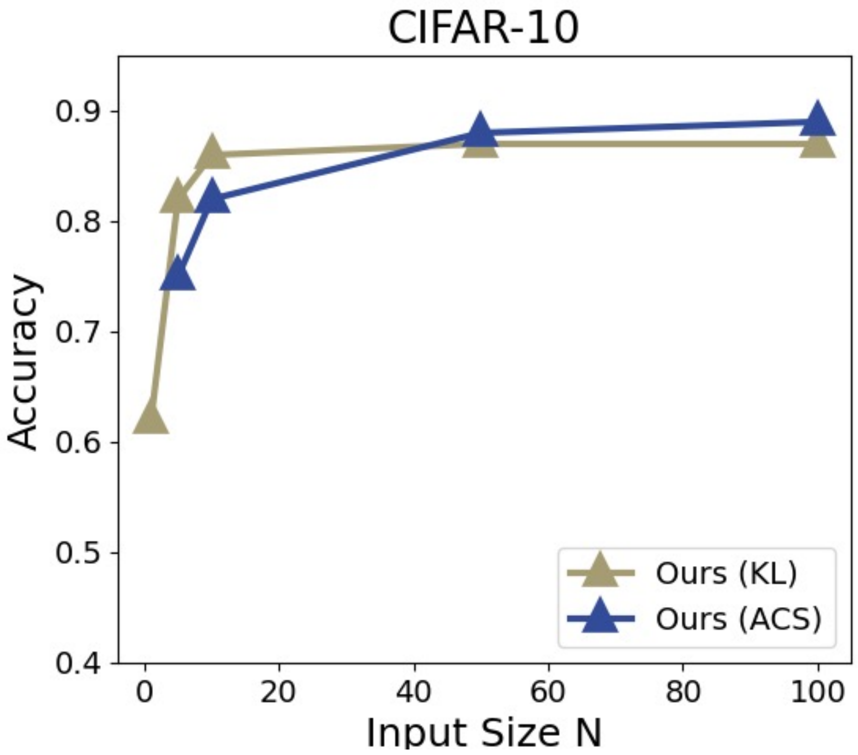
- Acc. $= \frac{\#(\text{Predicted teacher} = \text{True teacher})}{\#\text{Students}}$
- Reports how often the method correctly identifies the teacher used for distillation.

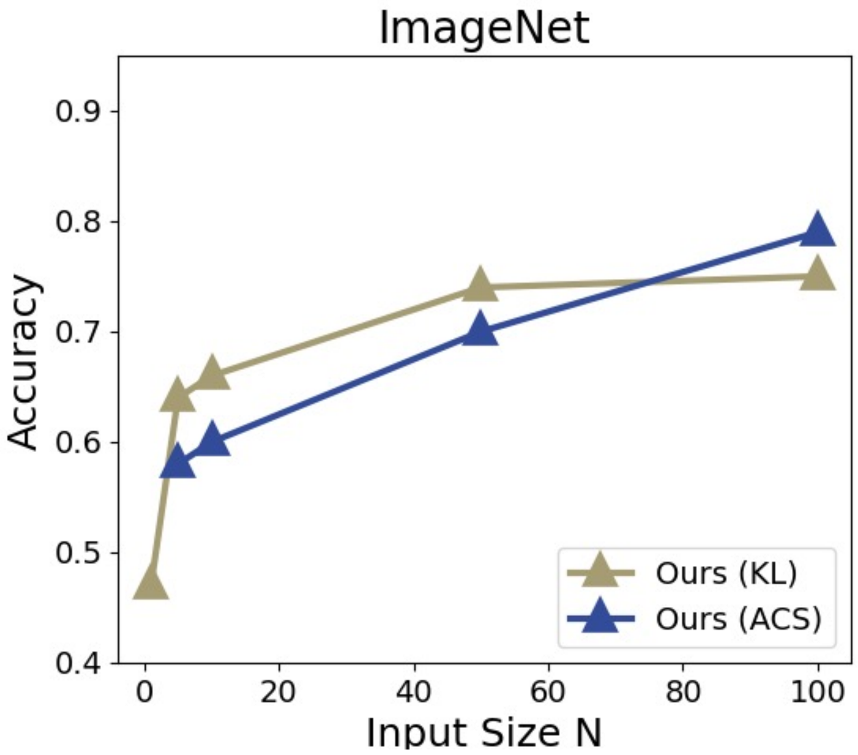
Area Under the Curve (AUC):

- For each student model, treat the true teacher as the positive class and others as negative.
- Compute the AUC for each students and report the average value.
- Reports how well the scores rank the true teacher higher than incorrect ones.

Image Classification Specific Details

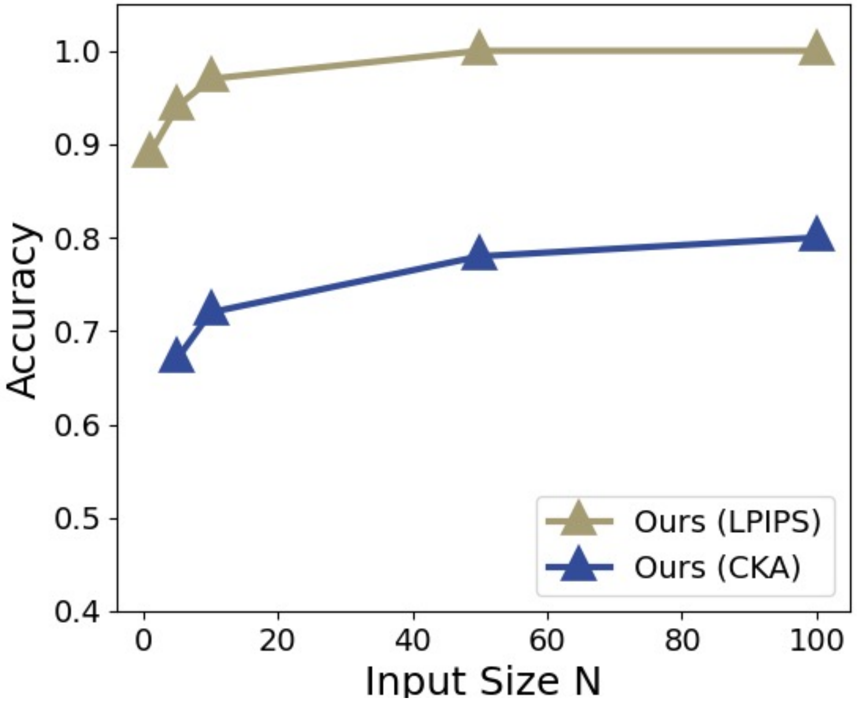
- Input construction: Training a generator
- Score computation: (1) KL divergence (2) Aligned Cosine Similarity (ACS)





Text-to-image Generation Specific Details

- Input construction: Empty String
- Score computation: (1) LPIPS (2) Centered Kernel Alignment (CKA)



Takeaways

- Unauthorized distillation poses real risks to data privacy and model ownership.
- This makes **intellectual property auditing** especially important for commercial models
- We propose the first model-agnostic framework for detecting distillation in open-weights model.

Paper: https://arxiv.org/pdf/2510.02302
Code : https://github.com/shqii1j/distillation_detection