TOWARDS DISTILLATION-RESISTANT LARGE LANGUAGE MODELS: A CONDITIONAL MUTUAL INFORMATION PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Proprietary large language models (LLMs) embody substantial economic value and are generally exposed only as black-box APIs. However, adversaries can still exploit their outputs to extract knowledge via distillation. Previous studies have explored defense strategies, yet they primarily focus on text-based distillation, leaving the important logit-based counterpart largely unexplored. In this work, we analyze this problem from an information-theoretic perspective and present an effective solution. Specifically, we characterize distillation-relevant information in teacher outputs via the conditional mutual information (CMI) between teacher logits and input queries, conditioned on ground-truth labels. This quantity captures contextual signals beyond label supervision that can be exploited for knowledge extraction, motivating us to defend against unauthorized distillation via CMI minimization. To this end, we propose a learnable transformation matrix that purifies the original outputs and provide a theoretical analysis that establishes its effectiveness. We further derive a CMI-inspired anti-distillation objective to optimize this transformation, which effectively reduces distillation-relevant information while preserving output utility. Extensive experiments across multiple LLMs and strong distillation algorithms demonstrate that the proposed method significantly degrades distillation performance while preserving task accuracy, thereby effectively protecting models' intellectual property.

1 INTRODUCTION

Large language models (LLMs) have recently demonstrated remarkable capabilities and have been deployed in a variety of practical scenarios, such as healthcare applications (Goyal et al., 2024) and multi-agent systems (Wu et al., 2024; Du et al., 2024). Training such huge models demands substantial resources, including large-scale datasets, extensive computational budgets, and significant human effort, rendering them highly valuable intellectual property. To safeguard their interests, commercial service providers typically isolate their proprietary models and only expose them through APIs. However, malicious attackers can exploit the output texts or logits to perform knowledge distillation (KD) (Hinton et al., 2015; Kim & Rush, 2016; Gu et al., 2024), thus cheaply extracting model capabilities without any access to the weights or architecture.

To defend against such theft, previous work has explored various LLM watermarking (Kirchenbauer et al., 2023; Ye et al., 2025; Liang et al., 2026) and fingerprinting techniques (Chen et al., 2023; Yu et al., 2021; Peng et al., 2022) for ownership verification, yet they provide only post-hoc verification and cannot directly prevent the extraction of proprietary knowledge. Recent efforts attempt to perturb the token distribution and modify model outputs to hinder KD (Savani et al., 2025; Li et al., 2025). Despite the success, these methods focus exclusively on text-based distillation, leaving the more widely discussed logit-based distillation largely unaddressed. Notably, a model's output logits reveal the full probability distribution over tokens, conveying substantially richer information than the sampled token labels. Consequently, defending against such distillation is particularly challenging. This difficulty is further exacerbated by a growing body of work (Gu et al., 2024; Wang et al., 2021; 2025) that strengthens logit-based KD by more fully exploiting their rich information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

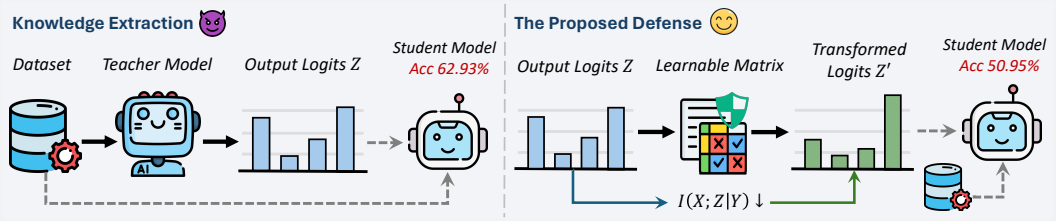


Figure 1: An illustration of the designed defense algorithm. We propose to learn a transformation matrix that post-processes output logits. By minimizing the conditional mutual information $\mathcal{I}(X; Z|Y)$, we successfully reduce the information beneficial for knowledge distillation within the teacher logits.

This work investigates this critical challenge by introducing an information-theoretic measure to quantify the distillation-relevant information within teacher logits, based on which we further propose an output purifying strategy to mitigate knowledge leakage. Fundamentally, LLMs perform a classification task to predict the next token for an ongoing input sequence. The information in an output vector can be interpreted in two aspects: (1) the predicted label for the current input (*i.e.*, the token corresponding to the maximum logit), and (2) contextual information beyond the label (Ye et al., 2024). Although different prefixes can yield the same next predicted token, their logits differ substantially due to varying contexts in semantics and syntax. Particularly, logit-based KD exploits such contextual information to transfer richer knowledge to the student model (Gu et al., 2024).

To quantify the supervisory guidance from teacher outputs during distillation, we introduce conditional mutual information (CMI) as a principled measure of the conveyed information. Formally, let X, Y, Z denote the input sequence, the ground-truth label, and the teacher logits for the given X, respectively. Then $\mathcal{I}(X; Z|Y)$ measures how much information Z provides about X conditioned on Y. This precisely aligns with the paradigm of logit-based KD that exploits information encoded in teacher logits beyond the label to enhance student performance. Therefore, a natural defense is to purify the teacher logits to minimize the conditional mutual dependence between X and Z. To achieve this goal, we propose learning a transformation matrix $\mathcal{M}$ that post-processes the output logits (see Figure 1), which we theoretically validate to reduce $\mathcal{I}(X; Z|Y)$. To effectively learn this matrix and decrease CMI, we further expand $\mathcal{I}(X; Z|Y)$ and decompose the minimization into two sub-problems based on the Information Bottleneck (IB) principle (Shwartz-Ziv & Tishby, 2017): (1) increasing the dependency between the logits Z and the ground-truth label Y, and (2) weakening the correlation between the input X and the logits Z. Guided by this analysis, we devise an optimization objective with two corresponding loss terms. We first derive a theoretical upper bound for the first subproblem, which formally justifies that minimizing the cross-entropy between Z and Y is a practical surrogate objective. The latter sub-task aims to reduce the relevance of X and Z, yet directly computing or estimating this term for LLMs is intractable. To circumvent this difficulty, we note that under the KD scenario, the information in Z about X is ultimately leveraged through gradients propagated to the student during optimization. In this sense, these gradients serve as the practical carrier of the guidance provided by teacher logits. Building on this insight, we employ a surrogate student model to simulate distillation and maximize the angular deviation between the gradients induced by the transformed logits Z' and those from the original Z. By iteratively optimizing the transformation $\mathcal{M}$ with the proposed loss function, we effectively reduce the distillation-beneficial contextual information in teacher outputs, thereby strengthening their resistance against knowledge extraction attacks.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to defend LLMs against logit-based KD. We analyze the problem from an information-theoretic perspective and propose learning a matrix that post-processes output logits to minimize their CMI with the input query.
- Guided by theoretical analysis, we design two complementary loss functions to optimize the transformation matrix, which effectively removes the contextual information within teacher logits that could be exploited by malicious attackers while preserving output utility.
- Extensive experiments across multiple LLMs such as Llama 3 series across three widely used datasets against four strong distillation algorithms demonstrate the outstanding defense performance of the proposed method in safeguarding model intellectual property.

2 Related Work

2.1 Knowledge Distillation for Large Language Models

Knowledge Distillation (KD) is a widely adopted paradigm for transferring useful knowledge from more powerful teacher models to less capable (generally smaller) student models. For large language models (LLMs), the distillation algorithm is typically achieved through two paradigms: sequence-level (text-based) and logit-based (response-level) approaches.

Sequence-level KD trains the student model using teacher-generated texts as hard-label supervision, treating teacher-produced outputs as pseudo-labels and optimizing the student via standard maximum-likelihood estimation (Kim & Rush, 2016). In addition to the SFT loss based on ground-truth labels, logit-based KD utilizes the teacher's soft output distributions to distill fine-grained predictive information by aligning the teacher and student distributions. The pioneering logit-based KD method (Hinton et al., 2015) uses Kullback-Leibler divergence (KLD) to measure the discrepancy between student and teacher distributions. Reverse KLD (RKLD) employs a reverse version of KL divergence by changing the reference distribution in KL from the teacher to the student, allowing the student to concentrate its limited capacity on the teacher's high-probability regions (Gu et al., 2024). AlphaNet (Wang et al., 2021) introduces α-divergence, providing a tunable measure between KLD and RKLD to penalize over-estimation and under-estimation in KD. ABKD (Wang et al., 2025) further proposes a novel α-β-divergence to balance hardness-concentration and confidence-concentration, generalizing FKLD and RKLD to a broader family of divergences.

2.2 Protecting LLMs from Knowledge Distillation

The success of KD poses a significant threat to the protection of intellectual property in LLMs. To mitigate this issue, watermarking (Kirchenbauer et al., 2023; Liang et al., 2026) embeds an identifiable and verifiable signature into a model's behavior or parameters to verify the ownership. Fingerprinting (Ye et al., 2025; Xu et al., 2024) assigns a unique and traceable identity to each distributed model instance, enabling the model owner to attribute leaked or misused models. However, both approaches are inherently post hoc, *i.e.*, they enable the identification of model stealing after it has occurred, but cannot prevent the extraction process itself.

Another research line aims to directly remove this threat by modifying the decoding distribution to obtain generated texts that are resilient to distillation. The pioneering work (Savani et al., 2025) proposes an anti-distillation decoding strategy that actively poisons reasoning traces by injecting a learned perturbation into the token-level probability distribution. Concretely, a proxy student first computes gradients on a small hold-out set of teacher-generated sequences, yielding a fixed poisoning direction $\nabla\ell$. During inference, the teacher performs two forward passes for each candidate token with proxy-student parameters perturbed by $\pm\eta\nabla\ell$, and estimates the resulting finite-difference logit shift. This shift is then integrated into the teacher's logits to form an adjusted sampling distribution. Sampling from this modified distribution implicitly steers any downstream student trained on these traces toward higher loss, thereby degrading the effectiveness of distillation. DOGe (Li et al., 2025) adopts a more direct approach that maximizes the divergence between teacher and student output distributions to make the teacher's predictions harder to imitate. Specifically, DOGe fine-tunes the teacher's final linear layer with an adversarial loss defined as the negative KL-divergence between the temperature-scaled softmax of the teacher logits and the detached softmax of a proxy student, enlarging the teacher–student discrepancy.

However, existing methods limit their scope in the sequence-level distillation scenarios, leaving the defense against the more widely discussed logit-based KD algorithms largely unexplored. To address this gap, this work makes an early step in investigating this issue from an information-theoretic perspective and proposes learning a transformation matrix as an effective post-processing solution.

3 Preliminary

3.1 Notation

Let $\mathcal{V}$ be the vocabulary and $|\mathcal{V}| < \infty$. Let $X \in \mathcal{V}^*$ be the input sequence consisting of the user's query and the ongoing generated response, and let $Y \in \mathcal{V}$ be the ground-truth label, *i.e.*, the next

token for the current input X. We define the large language model as a function $f : \mathcal{V}^* \to \mathbb{R}^{|\mathcal{V}|}$, and let $Z = f(X)$ be the resulting logit vector for input X. The next token prediction is sampled from the distribution $\sigma(Z)$, where $\sigma(\cdot)$ denotes the softmax function.

3.2 KNOWLEDGE DISTILLATION

Given a dataset $\mathcal{D}$ consisting of prompt-answer pairs (q, o), KD algorithms repeatedly query a teacher model $f_{\mathrm{T}}(\cdot)$ with (q, o) sampled from $\mathcal{D}$ and train a student model $f_{\mathrm{S}}(\cdot)$ by mimicking the teacher's output. Formally, the knowledge distillation loss can be expressed as:

$$\mathcal{L}_{\mathrm{KD}} \triangleq \frac{1}{l} \sum_{t=1}^{l} \mathrm{Dist}\Big(\sigma\big(f_{\mathrm{T}}(q, o_{<t})\big) \Big\| \sigma\big(f_{\mathrm{S}}(q, o_{<t})\big)\Big), \tag{1}$$

where l denotes the length of the ground-truth answer o, and $\mathrm{Dist}(\cdot\|\cdot)$ denotes the distance metric such as Kullback–Leibler divergence (KLD). In general, the optimization objective of logit-based distillation algorithms can be formulated as:

$$\mathcal{L}_{\text{total}} = (1 - \alpha) \cdot \mathcal{L}_{\mathrm{SFT}} + \alpha \cdot \mathcal{L}_{\mathrm{KD}}, \tag{2}$$

where the factor $\alpha = 0.5$ by default, and $\mathcal{L}_{\mathrm{SFT}}$ denotes the supervised fine-tuning (SFT) loss, *i.e.*, the negative log-likelihood of the predictive probability of the ground-truth label. Since the logits carry more contextual information than the predictive label alone, KD typically outperforms standard SFT on many downstream tasks. Based on this formulation, various improved algorithms have been proposed, including novel designs of distance metrics such as reverse KLD (Gu et al., 2024) and AB-KLD (Wang et al., 2025).

3.3 INFORMATION BOTTLENECK

The Information Bottleneck (IB) theory (Shwartz-Ziv & Tishby, 2017; Tishby et al., 2000; Tishby & Zaslavsky, 2015) highlights that a well-trained model should compress redundant information from inputs while preserving key information for tasks. The IB objective is formulated as:

$$\mathcal{L}_{\mathrm{IB}} = \mathcal{I}(X; Z) - \beta \cdot \mathcal{I}(Z; Y), \tag{3}$$

where $\mathcal{I}(\cdot;\cdot)$ is mutual information, and $\beta > 0$ is a hyperparameter. A lower $\mathcal{L}_{\mathrm{IB}}$ indicates a better performance and higher robustness of a model. Note that Z in IB represents the intermediate representation during inference. In this work, we instantiate Z as the logit vector.

4 METHOD

In this section, we first illustrate the rationality and benefits of minimizing CMI as a defense objective and present a theoretically-grounded post-processing strategy via a logit transformation matrix. Guided by theoretical insights, we design an optimization objective consisting of two complementary loss terms to train the matrix. An overview of the proposed method is illustrated in Figure 2.

4.1 CONDITIONAL MUTUAL INFORMATION-BASED DEFENSE

The CMI $\mathcal{I}(X; Z|Y)$ measures how much information Z reveals about X given the ground-truth label Y. The additional contextual information that Z carries beyond the label aligns exactly with the goal of logit-based distillation, as revealed by Eq. (1) and (2). Therefore, we propose to protect teacher knowledge by minimizing $\mathcal{I}(X; Z|Y)$, *i.e.*, suppressing the contextual information encoded in teacher logits. This objective can also be interpreted through entropy and uncertainty. Specifically, $\mathcal{I}(X; Z|Y) = \mathcal{H}(X|Y) - \mathcal{H}(X|Z, Y)$, which quantifies the reduction in uncertainty of X when Z is additionally observed. For a fixed training dataset, $\mathcal{H}(X|Y)$ remains constant. Hence, minimizing $\mathcal{I}(X; Z|Y)$ is equivalent to maximizing $\mathcal{H}(X|Z, Y)$, *i.e.*, making Z contain as little information about X as possible to increase the uncertainty about X given the label Y.

Remark 1. As an intuitive extreme case, if Z ideally degenerates to the one-hot label Y, *i.e.*, it contains no contextual information beyond the category, then logit-based knowledge distillation collapses to standard label-only SFT, and CMI indeed attains its minimum value $\mathcal{I}(X; Z|Y) = \mathcal{I}(X; Y|Y) = 0$.

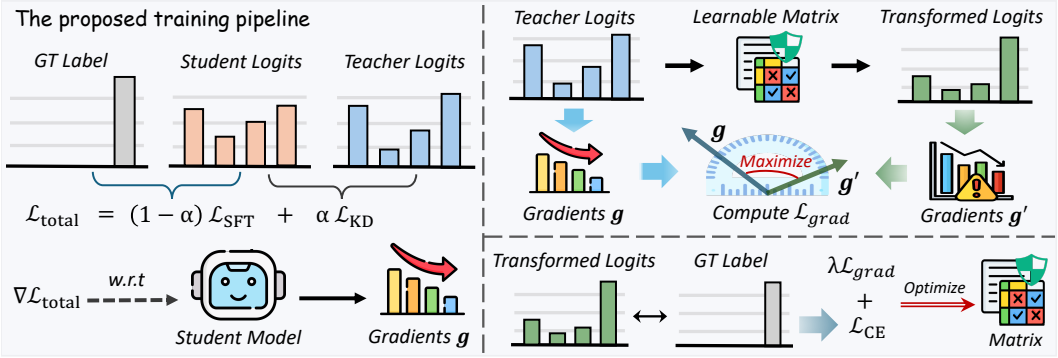


Figure 2: An illustration of the proposed method. As in the subfigure, we introduce a surrogate student model to simulate the KD. We compute the gradient $\boldsymbol{g}$ of $\mathcal{L}_{\text{total}}$ with respect to the student parameters to quantify the guidance carried by teacher outputs. The right subfigure explains the training loss for the transformation matrix. To diminish context information in teacher logits, we minimize the cross-entropy between the transformed logits Z' and the ground-truth label Y, while maximizing the angular deviation between gradients $\boldsymbol{g}$ and $\boldsymbol{g}'$ calculated by the original and transformed logits.

To achieve the goal of CMI minimization, we propose a lightweight post-processing strategy that further modifies output logits while keeping the teacher model completely frozen. Concretely, we introduce a learnable matrix transformation $\mathcal{M}(\cdot)$ that maps the original teacher output to refined ones. Before delving into the full training pipeline, we show that such a transformation admits an intriguing information-theoretic property, which formally guarantees that the refined logits become more effectively compressed in terms of contextual information:

Theorem 2. *Let $X \in \mathcal{V}^*$ be the current input, $Y \in \mathcal{V}$ be the ground-truth next token, $Z = f(X) \in \mathbb{R}^{|\mathcal{V}|}$ be the teacher logits, $\mathcal{M} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be the logit transformation matrix, and $Z' = \mathcal{M} \cdot Z$ be the refined logits. Then*

$$Y \to X \to Z \to Z' \tag{4}$$

forms a Markov chain, and their CMI values satisfy

$$\mathcal{I}(X; Z|Y) \geq \mathcal{I}(X; Z'|Y). \tag{5}$$

Detailed proofs can be found in Appendix A. Theorem 2 ensures that matrix transformations are universally beneficial for reducing CMI, which facilitates the compression of contextual information exploited by attackers. To further reduce CMI while preserving the utility of outputs, we additionally design an effective optimization objective to better guide the training of the transformation matrix.

4.2 Transformation Matrix Optimization

Next, we analyze the proposed matrix training objective through the lens of the Information Bottleneck (IB) principle to resolve the optimization problem. In particular, the following property reveals that minimizing $\mathcal{I}(X; Z|Y)$ in the considered scenario actually induces an IB objective.

Theorem 3. *The conditional mutual information $\mathcal{I}(X; Z|Y)$ is a special case of Information Bottleneck objective with $\beta = 1$ in eq. (3), i.e.,*

$$\mathcal{I}(X; Z|Y) = \mathcal{I}(X; Z) - \mathcal{I}(Z; Y). \tag{6}$$

A detailed derivation is provided in Appendix A. This property is particularly appealing as it reveals that minimizing $\mathcal{I}(X; Z|Y)$ inherits the benefits of the Information Bottleneck, balancing compression of redundant information in X while preserving task-relevant information in Y. The above characterization naturally decomposes the objective into two complementary sub-tasks:

- **Minimizing** $\mathcal{I}(X; Z)$. This compresses redundant contextual information in teacher logits Z that originates from the input X, weakening the dependency between inputs and outputs. As a result, Z becomes less informative about X, which enhances the resistance against KD algorithms.

- **Maximizing** $\mathcal{I}(Z;Y)$. This preserves task-relevant information to maintain model utility. Notably, it encourages Z to concentrate toward its class centroid (*i.e.*, increasing the correct label's logit while suppressing others), which also reduces contextual information and brings defense benefits.

Training loss design. Motivated by the two sub-tasks, we design two complementary loss functions to minimize CMI. We begin with analyzing the second term $\mathcal{I}(Z;Y)$, which can be expanded into

$$\mathcal{I}(Z;Y) = \mathcal{H}(Y) - \mathcal{H}(Y|Z). \tag{7}$$

Note that $\mathcal{H}(Y)$ is constant for a fixed dataset, thus maximizing $\mathcal{I}(Z;Y)$ is equivalent to minimizing the conditional entropy $\mathcal{H}(Y|Z)$, which measures the residual uncertainty about the ground-truth label Y after observing logits Z. Although directly computing $\mathcal{H}(Y|Z)$ is generally intractable, we can derive a theoretically grounded and tractable upper bound as follows:

Lemma 4. *Let $Y \in \mathcal{V}$ be the ground-truth next token and $Z \in \mathbb{R}^{|\mathcal{V}|}$ be the teacher logits. Let $p(\cdot|Z)$ be the distribution of Y conditional on Z, and $\hat{p}(\cdot|Z) = \sigma(Z)$ be its predictive distribution by the teacher model. Then the cross-entropy $\mathcal{H}(p,\hat{p})$ can be decomposed into*

$$\mathcal{H}(p,\hat{p}) = \mathcal{H}(Y|Z) + \mathbb{E}_Z\left[\mathbb{KL}\Big(p(\cdot|Z)\Big\|\hat{p}(\cdot|Z)\Big)\right], \tag{8}$$

which further implies the upper bound

$$\mathcal{H}(p,\hat{p}) \geq \mathcal{H}(Y|Z). \tag{9}$$

Remark 5. A detailed proof is provided in Appendix A. In classification tasks with discrete labels, $p(\cdot|Z)$ degenerates into a one-hot distribution, making $\mathcal{H}(p,\hat{p})$ equivalent to the standard cross-entropy loss, denoted as $\mathcal{L}_{\mathrm{CE}}$. Therefore, minimizing $\mathcal{L}_{\mathrm{CE}}$ provides a tractable surrogate for reducing the conditional entropy $\mathcal{H}(Y|Z)$. In practice, optimizing $\mathcal{L}_{\mathrm{CE}}$ increases the logit corresponding to the ground-true label Y. This intuitively aligns with the fact that if $\mathrm{argmax}(Z) = Y$ always holds, the conditional entropy attains its minimum, $\mathcal{H}(Y|Z) = 0$.

For minimizing $\mathcal{I}(X;Z)$, directly estimating $\mathcal{I}(X;Z)$ for autoregressive LLMs is fundamentally infeasible, since both X and Z are in high-dimensional spaces, and their underlying distributions are difficult to approximate. To handle this challenge, we note that in KD scenarios, the information of X residing in Z is ultimately exploited through how Z instructs the student's output, which in practice materializes in the form of gradients that update the student's parameters. Hence, the learning signals conveyed by the gradients serve as the effective carrier of the information of X residing in Z:

Proposition 6. *Let θ denote the parameters of a surrogate student model $f_{\mathrm{S}}(\cdot)$. Under the standard KD setting, the information in Z about X that can be exploited during distillation is operationally conveyed through the gradients of the loss with respect to θ, i.e.,*

$$\boldsymbol{g} = \nabla_\theta \frac{1}{l}\sum_{t=1}^{l}\Big[-(1-\alpha)\cdot\log\sigma\big(f_{\mathrm{S}}(q,o_{<t})\big)_{o_t} + \alpha\cdot\mathbb{KL}\Big(\sigma\big(f_{\mathrm{T}}(q,o_{<t})\big)\Big\|\sigma\big(f_{\mathrm{S}}(q,o_{<t})\big)\Big)\Big]. \tag{10}$$

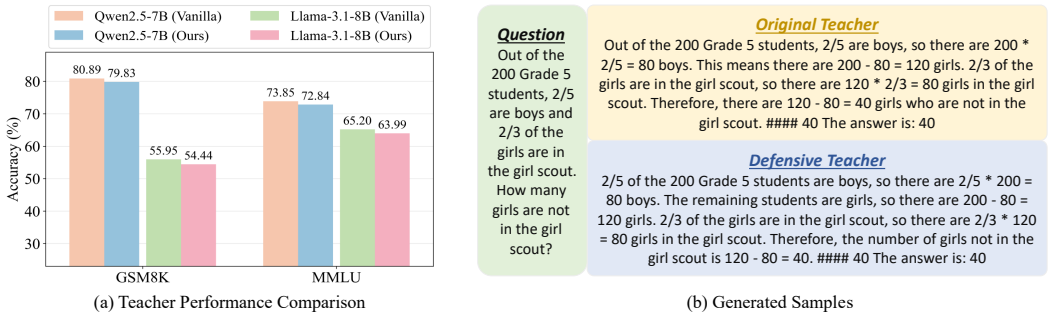
Based on the gradients $\boldsymbol{g}$, we can characterize the information within teacher logits Z that is transmitted to the student during distillation. The optimization objective is accordingly reformulated as obtaining a transformed distribution Z' whose induced gradients $\boldsymbol{g}'$ contain as little valid information as possible. Since the original $\boldsymbol{g}$ represents the correct and informative update direction under KD, we aim to disrupt this signal by maximizing the angular deviation between $\boldsymbol{g}$ and $\boldsymbol{g}'$, measured via cosine similarity. This strategy implicitly suppresses distillation-relevant information in Z through gradient distortion, significantly enhancing resistance to logit-based distillation attacks.

With these loss terms, the optimization objective for the transformation matrix $\mathcal{M}$ is formulated as:

$$\mathcal{L}_{\mathcal{M}} = \mathcal{L}_{\mathrm{CE}} + \lambda\cdot\mathcal{L}_{\mathrm{grad}} = -\frac{1}{l}\sum_{t=1}^{l}\log\sigma\big(f_{\mathrm{T}}(q,o_{<t})\big)_{o_t} + \lambda\cdot\frac{\langle\boldsymbol{g},\boldsymbol{g}'\rangle}{\|\boldsymbol{g}\|_F\|\boldsymbol{g}'\|_F}, \tag{11}$$

where $\lambda > 0$ is the scaling factor, $\langle\cdot,\cdot\rangle$ is the dot operation, and $\|\cdot\|_F$ denotes the Frobenius norm.

Transformation matrix optimization. Notably, a direct implementation of $\mathcal{M}$ as a dense $|\mathcal{V}| \times |\mathcal{V}|$ matrix is computationally prohibitive due to the large vocabulary size. To mitigate this issue, we



(a) Teacher Performance Comparison

(b) Generated Samples

Figure 3: Performance of the teacher model under the defense and visualization comparison of defensive and original teacher outputs using Qwen2.5-7B on GSM8K.

adopt a parameter-efficient formulation by decomposing $\mathcal{M}$ into a low-rank structure, $\mathcal{M} = \mathbf{E} + \mathbf{AB}$, where $\mathbf{E}$ denotes the identity matrix, and $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times |\mathcal{V}|}$ are learnable matrices with decomposition rank r (Hu et al., 2022). Moreover, directly computing full gradients renders the optimization of $\mathcal{L}_{\text{grad}}$ intractable due to their high dimensionality. To this end, we adopt the gradients of the student model's second-to-last layer as a surrogate in the loss calculation, as they contain rich task-relevant information and closely align with the model's predicting behaviors.

5 Experiments

5.1 Experimental Setup

Models and datasets. Note that logit-based KD requires teachers and students to share the same tokenization to ensure one-to-one token correspondence. Hence, we select models from the same series to avoid engineering issues from vocabulary mismatches and better focus on defensive analysis. Following (Savani et al., 2025; Li et al., 2025), we use two widely used LLM families, Qwen2.5 (Yang et al., 2024) and Llama 3 (Grattafiori et al., 2024). For the Qwen family, we use Qwen2.5-7B as the teacher model and Qwen2.5-3B-Instruct as the surrogate student. Distillation is performed on multiple student models with different scales and capacities, including Qwen2.5-3B/1.5B/0.5B. For the Llama family, we use Llama-3.1-8B as the teacher model and Llama-3.2-3B-Instruct as the surrogate student. The defensive performance is then evaluated on student models Llama-3.2-3B and Llama-3.2-1B. Following prior work (Savani et al., 2025; Li et al., 2025), we consider three widely used benchmarks, GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), and MATH-500 (Lightman et al., 2023). For training-test sets construction, please refer to Appendix B.1.

Distillation baselines. We evaluate our method against four logit-based KD baselines that have exhibited competitive performance across a wide range of tasks: the classical logit-based KD (Hinton et al., 2015) with forward KLD, reverse KLD (RKLD) (Gu et al., 2024), AlphaNet (Wang et al., 2021), and the state-of-the-art ABKD (Wang et al., 2025).

Implementation details. We train the transformation matrix for 5 epochs at a batch size of 32. For GSM8K and MATH, the learning rate (lr) is set to $5 \cdot 10^{-7}$, while for MMLU is $5 \cdot 10^{-6}$. The matrix decomposition rank is set to 32. The lr follows a linear warm-up schedule over the first 10% of total training steps and then decays to zero using cosine annealing. We optimize the matrix using AdamW with a weight decay of 0.01 and default momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We adopt the implementation and corresponding hyperparameter settings provided in (Wang et al., 2025), using $\alpha = 0.1$ and $\beta = 0.9$ for AlphaNet, and $\alpha = 0.1$ and $\beta = 0.8$ for ABKD. In Sec. 5.3, we explore the influence of the loss factor λ. For a stable and reproducible evaluation, we conduct all experiments using greedy decoding. More details including student distillation configuration are in Appendix B.2.

5.2 Performance Evaluation

We first report the accuracy of defensive teacher models in Figure 3(a). By maximizing mutual information between Z' and Y using $\mathcal{L}_{\text{CE}}$, the learned matrix effectively preserves teacher accuracy. Figure 3(b) shows visualizations of generated outputs. As observed, our defensive teacher produces

Table 1: Distillation performance of different student models under different distillation approaches.

Dataset	Method	Defense	Qwen2.5 Series			Llama 3 Series	
			3B	1.5B	0.5B	3B	1B
GSM8K	*SFT-Only*	–	68.39	58.83	36.54	46.85	25.40
	KD	*Vanilla*	73.09	62.85	38.06	46.25	23.96
		Ours	**71.65**	**58.53**	**35.94**	**42.91**	**20.09**
	AlphaNet	*Vanilla*	74.07	62.93	40.79	46.25	25.78
		Ours	**70.20**	**50.95**	**37.23**	**41.47**	**23.58**
	MiniLLM	*Vanilla*	73.92	63.76	41.24	48.14	25.63
		Ours	**71.49**	**54.44**	**37.68**	**41.09**	**22.29**
	ABKD	*Vanilla*	73.84	62.47	40.11	45.34	23.88
		Ours	**71.11**	**54.28**	**37.00**	**40.03**	**19.94**
MMLU	*SFT-Only*	–	59.85	52.62	44.99	40.09	31.33
	KD	*Vanilla*	60.47	55.28	46.08	41.25	32.97
		Ours	**58.89**	**53.95**	**44.79**	**38.89**	**28.96**
	AlphaNet	*Vanilla*	60.58	55.29	46.51	41.46	32.05
		Ours	**58.05**	**53.85**	**45.10**	**39.46**	**30.24**
	MiniLLM	*Vanilla*	60.55	54.03	41.24	38.68	32.03
		Ours	**58.30**	**52.60**	**38.81**	**36.10**	**30.62**
	ABKD	*Vanilla*	61.22	55.03	46.18	42.71	33.71
		Ours	**59.44**	**52.63**	**44.75**	**39.58**	**30.77**

fluent and coherent responses while preserving task correctness. This contrasts with existing sequence-level distillation defenses that disrupt sampling distributions and induce texts with degraded quality.

We then present student distillation results in Table 1. The results show that the learned matrix strengthens distillation resistance and yields a consistent performance degradation on distilled students across different teacher–student pairs and datasets. *E.g.*, distilling Qwen2.5-1.5B on GSM8K via AlphaNet incurs an accuracy drop of 11.98%, highlighting our effectiveness of suppressing distillation-relevant information. Notably, student models are jointly optimized using both the distillation loss $\mathcal{L}_{\mathrm{KD}}$ and an SFT loss from ground-truth labels. Our defense perturbs only the teacher logits and only affects $\mathcal{L}_{\mathrm{KD}}$, yet still misleads the distillation optimization. Furthermore, we observe cases where distillation with the original logits surpasses its SFT-only counterpart, whereas using transformed logits performs even worse than SFT. This further validates that our strategy effectively suppresses the contextual information required for effective distillation.

We also observe that our defense has a stronger effect on GSM8K than MMLU, which we attribute to GSM8K's longer multi-step mathematical reasoning, where perturbations accumulate along the reasoning trajectory and exert a more pronounced effect. Due to space limits, we provide results on surrogate models in Appendix C and visualization of student outputs in Appendix!F.

Distillation on the MATH benchmark. Beyond the above two benchmarks, we further evaluate our method on the more challenging MATH dataset, which comprises substantially more difficult mathematical problems. As the Llama series performs poorly on this dataset, which makes the evaluation uninformative, we instead conduct experiments using the stronger Qwen series to obtain more meaningful results.

After being equipped with the learned matrix, the teacher's accuracy exhibits only a marginal decrease, dropping from 53.8% to 50.4%. In contrast, as shown in Table 2, the distilled students suffer a far more pronounced degradation.

Table 2: Performance of Qwen2.5 student models on MATH-500 under different distillation methods.

Method	Defense	Qwen2.5 Series		
		3B	1.5B	0.5B
SFT-Only	–	49.40	31.80	20.80
KD	*Vanilla*	50.80	33.20	23.40
	Ours	**44.00**	**29.40**	**18.40**
AlphaNet	*Vanilla*	51.80	33.20	23.60
	Ours	**43.60**	**29.00**	**17.00**
MiniLLM	*Vanilla*	51.40	34.40	23.80
	Ours	**45.80**	**29.60**	**19.80**
ABKD	*Vanilla*	52.60	34.00	25.80
	Ours	**46.60**	**29.40**	**18.80**

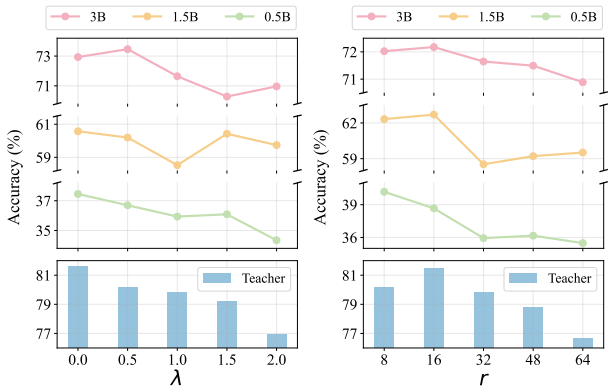


Figure 4: Accuracy of teacher and distilled student models on the GSM8K dataset under varying values of λ and r.

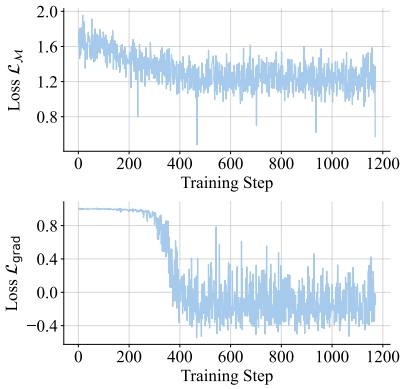


Figure 5: The loss landscape for Llama-3.1-8B on GSM8K during training.

E.g., under AlphaNet with Qwen2.5-3B, the accuracy drops by a notable value of 8.2%. This again confirms the effectiveness and universality of our CMI-inspired defense strategy across datasets.

5.3 Ablation Study and Loss Landscape Analysis

This part first provides ablation analysis on Qwen using GSM8K under standard KD. We then present a loss landscape analysis to investigate the optimization process. More ablations are in Appendix C.

Ablation on the matrix loss parameter λ. λ is an important scaling factor that balances the contributions of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{grad}}$ during matrix training in Eq. (11). The left subfigure of Figure 4 shows that when increasing λ, there is a general trend of decreasing student accuracy, which confirms that $\mathcal{L}_{\text{grad}}$ further suppresses distillation-relevant information and strengthens the defense. However, overly large λ also leads to a noticeable drop in the teacher's accuracy. We then choose $\lambda = 1$ to achieve a better trade-off between defense strength and utility preservation.

Ablation on the decomposition rank r of $\mathcal{M}$. The right subfigure of Figure 4 reveals a trade-off between teacher utility and defensive efficacy. When the rank r is small, the limited model capacity primarily focuses on minimizing $\mathcal{L}_{\text{CE}}$, thereby preserving the teacher's accuracy but yielding insufficient defensive effects against knowledge extraction. In contrast, larger r allows more capacity for optimizing $\mathcal{L}_{\text{grad}}$, enhancing defense while potentially compromising teacher accuracy. Therefore, we adopt $r = 32$ as the default setting to balance defense effectiveness and model utility.

Loss landscape. To better understand the underlying mechanism, we analyze the training dynamics of the overall loss $\mathcal{L}_{\mathcal{M}}$ and the gradient deviation loss $\mathcal{L}_{\text{grad}}$. As shown in Figure 5, the cosine similarity term $\mathcal{L}_{\text{grad}}$ starts around 1, indicating that the gradients of transformed logits are initialized to be identical to original ones (*i.e.*, $\mathcal{M} = \mathbf{E}$ initially). During the first 300 training steps, the overall loss $\mathcal{L}_{\mathcal{M}}$ steadily decreases while $\mathcal{L}_{\text{grad}}$ remains nearly unchanged, implying that this phase is dominated by the optimization of $\mathcal{L}_{\text{CE}}$. As training progresses, $\mathcal{L}_{\text{grad}}$ rapidly drops to negative values, indicating that the learned matrix induces a significant deviation in the gradient angle.

6 Conclusion

In this paper, we investigate the challenging problem of defending LLMs against unauthorized logit-based KD from an information-theoretical perspective. We characterize the guidance signals in the teacher logits via conditional mutual information, based on which we propose a post-processing approach that learns a transformation matrix using two theory-inspired loss functions. By refining the output logits, the proposed method degrades the performance of distilled student models while preserving teacher utility, effectively protecting the intellectual property. We highlight that, unlike prior defenses against sequence-level distillation, our method makes a pioneering step in defending against tricky logit-based KD techniques. We hope this work advances the community's understanding of LLM distillation and motivates more robust distillation defenses.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with writing and language polishing. We have not used generative AI tools for research ideation, data collection, experiments, or any other research-related tasks, and the remaining disclosure categories are not applicable to this work. Additionally, we used generative AI tools for language editing and improving the clarity and readability of the manuscript. We have reviewed all AI-assisted work and verified that the final text accurately reflects our research, methods, results, and claims. We take full responsibility for the final content of this work, including all text, claims, and other artifacts produced with the aid of generative AI.

REFERENCES

Yanjiao Chen, Rui Guan, Xueluan Gong, Jianshuo Dong, and Meng Xue. D-dae: Defense-penetrating model extraction attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 382–399. IEEE, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, pp. 11733–11763. PMLR, 2024.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 1167–1168, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, pp. 24872–24895. PMLR, 2024.

Pingzhi Li, Zhen Tan, Mohan Zhang, Huaizhi Qu, Huan Liu, and Tianlong Chen. Doge: Defensive output generation for llm protection against knowledge distillation. *arXiv preprint arXiv:2505.19504*, 2025.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *Artificial Intelligence Review*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13430–13439, 2022.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J Zico Kolter. Antidistillation sampling. *arXiv preprint arXiv:2504.13146*, 2025.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *ArXiv*, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *Information Theory Workshop (ITW)*, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *ArXiv*, 2000.

Dilin Wang, Chengyue Gong, Meng Li, Qiang Liu, and Vikas Chandra. Alphanet: Improved training of supernets with alpha-divergence. In *International Conference on Machine Learning*, pp. 10760–10771. PMLR, 2021.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. Abkd: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. In *Forty-second International Conference on Machine Learning*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL `https://api.semanticscholar.org/CorpusID:274859421`.

Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and EN-HUI YANG. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information. In *The Twelfth International Conference on Learning Representations*, 2024.

Peigen Ye, Huali Ren, Zhengdao Li, Anli Yan, Hongyang Yan, Shaowei Wang, and Jin Li. Securing large language models: A survey of watermarking and fingerprinting techniques. *ACM Computing Surveys*, 2025.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 14448–14457, 2021.

A PROOFS

Theorem 2. *Let $X \in \mathcal{V}^*$ be the current input sequence, $Y \in \mathcal{V}$ be the ground-truth next token, $Z = f(X) \in \mathbb{R}^{|\mathcal{V}|}$ be the teacher logits for predicting Y, $\mathcal{M} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be the logit transformation matrix, and $Z' = \mathcal{M} \cdot Z$ be the refined logits. Then*

$$Y \to X \to Z \to Z' \tag{12}$$

forms a Markov chain, and their conditional mutual information satisfies

$$\mathcal{I}(X; Z|Y) \geq \mathcal{I}(X; Z'|Y). \tag{13}$$

Proof. Note that both $Z = f(X)$ and $Z' = \mathcal{M} \cdot Z$ are deterministic functions, implying that $Y \to X \to Z \to Z'$ forms a Markov chain. Furthermore we have

$$\begin{aligned}
\mathcal{I}(X; Z|Y) &= \mathcal{I}(X; Z, Z'|Y), &(14)\\
&= \mathcal{I}(X; Z'|Y) + \mathcal{I}(X; Z|Z', Y), &(15)\\
&\geq \mathcal{I}(X; Z'|Y), &(16)
\end{aligned}$$

where eq. (14) follows from the fact that the mapping from Z to Z' is deterministic, eq. (15) applies the chain rule of mutual information, and eq. (16) follows from the non-negativity of mutual information. □

Theorem 3. *The conditional mutual information $\mathcal{I}(X; Z|Y)$ is a special case of Information Bottleneck objective with $\beta = 1$ in eq.* (3)*, i.e.,*

$$\mathcal{I}(X; Z|Y) = \mathcal{I}(X; Z) - \mathcal{I}(Z; Y). \tag{17}$$

Proof. According to the chain rule of mutual information, we have

$$\begin{aligned}
\mathcal{I}(Y, X; Z) &= \mathcal{I}(Y; Z) + \mathcal{I}(X; Z|Y), &(18)\\
\mathcal{I}(Y, X; Z) &= \mathcal{I}(X; Z) + \mathcal{I}(Y; Z|X). &(19)
\end{aligned}$$

Combining above yields

$$\mathcal{I}(Y; Z) + \mathcal{I}(X; Z|Y) = \mathcal{I}(X; Z) + \mathcal{I}(Y; Z|X). \tag{20}$$

According to theorem 2, $Y \to X \to Z$ forms a Markov chain, implying that

$$\mathcal{I}(Y; Z|X) = 0. \tag{21}$$

Substituting (21) into (20) yields

$$\mathcal{I}(Y; Z) + \mathcal{I}(X; Z|Y) = \mathcal{I}(X; Z), \tag{22}$$

which is equivalent to

$$\mathcal{I}(X; Z|Y) = \mathcal{I}(X; Z) - \mathcal{I}(Z; Y). \tag{23}$$

□

Lemma 4. *Let $Y \in \mathcal{V}$ be the ground-truth next token and $Z \in \mathbb{R}^{|\mathcal{V}|}$ be the teacher logits. Let $p(\cdot|Z)$ be the distribution of Y conditional on Z, and $\hat{p}(\cdot|Z) = \sigma(Z)$ be its predictive distribution by the teacher model. Then the cross-entropy $\mathcal{H}(p, \hat{p})$ can be decomposed into*

$$\mathcal{H}(p, \hat{p}) = \mathcal{H}(Y|Z) + \mathbb{E}_Z \left[\mathbb{KL}\Big(p(\cdot|Z) \Big\| \hat{p}(\cdot|Z) \Big) \right], \tag{24}$$

which further implies the upper bound

$$\mathcal{H}(p, \hat{p}) \geq \mathcal{H}(Y|Z). \tag{25}$$

Proof.

$$\mathcal{H}(p, \hat{p}), \tag{26}$$
$$= \mathbb{E}_{Y,Z}\left[-\log \hat{p}(Y|Z)\right], \tag{27}$$
$$= \mathbb{E}_{Y,Z}\left[-\log p(Y|Z) + \log \frac{p(Y|Z)}{\hat{p}(Y|Z)}\right], \tag{28}$$
$$= \mathbb{E}_{Y,Z}\left[-\log p(Y|Z)\right] + \mathbb{E}_{Y,Z}\left[\log \frac{p(Y|Z)}{\hat{p}(Y|Z)}\right], \tag{29}$$
$$= \mathcal{H}(Y|Z) + \mathbb{E}_{Y,Z}\left[\log \frac{p(Y|Z)}{\hat{p}(Y|Z)}\right], \tag{30}$$
$$= \mathcal{H}(Y|Z) + \mathbb{E}_{Z}\left[\mathbb{E}_{Y\sim p(\cdot|Z)}\left[\log \frac{p(Y|Z)}{\hat{p}(Y|Z)}\right]\right], \tag{31}$$
$$= \mathcal{H}(Y|Z) + \mathbb{E}_{Z}\left[\mathbb{KL}\Big(p(\cdot|Z)\Big\|\hat{p}(\cdot|Z)\Big)\right], \tag{32}$$
$$\geq \mathcal{H}(Y|Z), \tag{33}$$

where the $\geq$ is due to the non-negativity of KL divergence. □

Remark 5. In supervised learning with discrete labels, $p(\cdot|z)$ degenerates to a one-hot distribution, making $\mathcal{H}(p, \hat{p})$ equivalent to the standard cross-entropy loss. Therefore, minimizing cross-entropy serves as a tractable and effective surrogate for reducing the conditional entropy $\mathcal{H}(Y|Z)$, which in turn directly increases the information content of the logits with respect to the true label. Intuitively, this is consistent with the observation that if $\text{argmax}(Z) = Y$ always holds, the predictive uncertainty vanishes and the conditional entropy attains its minimum value, $\mathcal{H}(Y|Z) = 0$.

B Experimental Details

B.1 Data Construction

GSM8K is a math word-problem dataset with manually annotated reasoning chains, making it particularly suitable for knowledge distillation in reasoning-intensive tasks. We use the standard split, which contains 7,473 training examples and 1,319 evaluation examples. MMLU is a widely used multiple-choice benchmark covering a broad range of subjects. As MMLU provides only final answer choices without annotated reasoning, we generate reasoning traces using Qwen2.5-7B and Llama-3.1-8B-Instruct for experiments on the Qwen model family and Llama family, respectively. From these generated outputs, we retain 15,000 instances with correct answers to construct the training set. Evaluation is conducted on the full MMLU test set, which contains 14,042 examples. The prompt used for training data generation is shown below.

Data Generation Prompt:

Please answer the following multiple-choice question with concise and effective reasoning.

Question: question
Choices:
A. choices[0]
B. choices[1]
C. choices[2]
D. choices[3]

Instructions:
1. Provide a short explanation within 128 words.
2. Avoid unnecessary or verbose explanations. Don't repeat the question.
3. After the short explanation, give your final answer enclosed in \boxed{}.

Now, please give your response.

B.2 KNOWLEDGE DISTILLATION CONFIGURATION

We train the transformation matrix for 5 epochs and select the final matrix that minimizes $\mathcal{L}_{\text{grad}}$ on the premise of preserving the teacher's original performance. The distillation loss coefficient α is set to 0.5 by default, which assigns equal importance to loss terms $\mathcal{L}_{\text{SFT}}$ and $\mathcal{L}_{\text{KD}}$. Following established practices in prior distillation studies (Wang et al., 2025; Gu et al., 2024; Ko et al., 2024), we search distillation hyperparameters within their recommended ranges to reach the best performance under clean distillation, and report the corresponding performance after applying our defense. Specifically, we vary the learning rate is selected from {1e-6, 5e-6, 1e-5, 5e-5} for the Qwen family and {1e-5, 5e-5, 1e-4} for the Llama family. All models are optimized using the AdamW optimizer with a weight decay of 0.01 and momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate schedule consists of a linear warm-up over the first 10% of training steps, followed by cosine annealing to zero. We adopt greedy decoding in all experiments for stable, reproducible evaluations. The *max new token* is set to 512 by default, and is 2048 for the MATH-500 benchmark, as it requires a substantially longer reasoning process. All the experiments are performed with 2 NVIDIA A800 GPUs with 80 GB of memory.

C ADDITIONAL RESULTS

Distillation of surrogate students. In the main text, we present distillation results on black-box student models to evaluate the practical effectiveness of our defense under realistic deployment scenarios. Here, we further supplement the analysis by reporting defense performance on surrogate student models used during matrix learning, namely Qwen2.5-3B-Instruct for the Qwen series and Llama-3.2-3B-Instruct for the Llama series.

The corresponding results are summarized in Table C.1 and Table C.2. As expected, the proposed defense demonstrates strong effectiveness in this white-box setting and generally exhibits stronger defensive effectiveness than in the black-box scenario.

As observed, the proposed defense results in an average accuracy drop of 8.79% across various distillation algorithms and surrogate student models on GSM8K, which significantly exceeds the 4.52% reduction observed for black-box students. This discrepancy arises because the transformation matrix is explicitly optimized against these surrogate models, leading to stronger and more targeted defensive effects. A similar trend can be consistently observed on MMLU and MATH-500. For instance, on MATH-500, the defense yields an average performance decrease of 6.55%, which again surpasses the 5.55% reduction observed under black-box settings. These results further demonstrate the effectiveness and generalizability of our approach across different benchmarks and threat scenarios.

Table C.1: Distillation performance of surrogate students Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct under different distillation algorithms (%).

Dataset	Method	Defense	Qwen-Ins	Llama-Ins
GSM8K	*SFT-Only*	–	67.63	59.21
	KD	*Vanilla*	70.36	61.94
		Ours	**68.84**	**56.79**
	AlphaNet	*Vanilla*	70.96	57.77
		Ours	**55.88**	**51.93**
	MiniLLM	*Vanilla*	70.13	58.15
		Ours	**56.71**	**46.70**
	ABKD	*Vanilla*	71.34	58.23
		Ours	**58.00**	**53.68**
MMLU	*SFT-Only*	–	58.41	41.48
	KD	*Vanilla*	60.48	43.07
		Ours	**58.62**	**40.39**
	AlphaNet	*Vanilla*	60.70	43.88
		Ours	**58.22**	**39.70**
	MiniLLM	*Vanilla*	60.26	44.43
		Ours	**59.19**	**41.48**
	ABKD	*Vanilla*	59.81	43.31
		Ours	**57.77**	**38.27**

Table C.2: Distillation performance of surrogate Qwen2.5-3B-Instruct under MATH-500 (%).

Defense	*SFT-Only*	KD	AlphaNet	MiniLLM	ABKD
Vanilla	54.40	57.40	57.40	57.40	57.20
Ours	**–**	**49.00**	**49.80**	**52.40**	**52.00**

Ablation on the $\mathcal{L}_{\text{CE}}$. We then analyze the influence of $\mathcal{L}_{\text{CE}}$ on the matrix performance. Specifically, $\mathcal{L}_{\text{CE}}$ encourages the teacher logits Z to encode more information about the ground-truth label Y, improving predictive accuracy while suppressing redundant contextual information. To further investigate its role, we design a variant that removes $\mathcal{L}_{\text{CE}}$ from the training objective and examine its impact on teacher performance. As shown in Figure C.1, eliminating $\mathcal{L}_{\text{CE}}$ causes the teacher's accuracy to drop drastically after being equipped with the transformation matrix $\mathcal{M}$, rendering the defense practically unusable. This

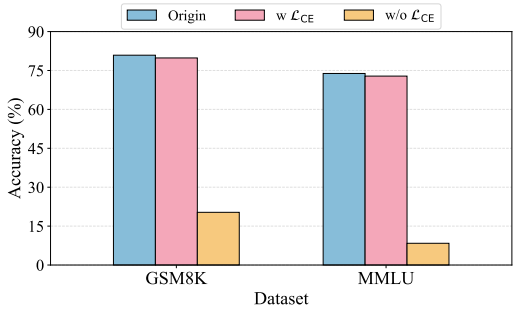


Figure C.1: Ablation study on the role of $\mathcal{L}_{\mathrm{CE}}$ in teacher model's accuracy. We conduct experiments on Qwen2.5-7B.

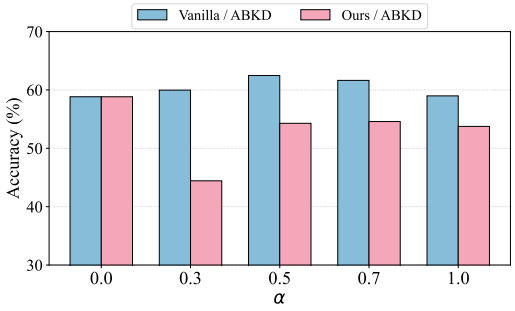


Figure C.2: Accuracy of students under varying values of α. When $\alpha = 0$, the loss degenerates to standard SFT without the KD term.

observation highlights the critical importance of $\mathcal{L}_{\mathrm{CE}}$ and further validates the rationality of our theoretically grounded loss design in effectively preserving teacher accuracy during matrix optimization.

Ablation on the hyperparameter α. In existing KD studies, the coefficient α is typically set to 0.5. To more comprehensively evaluate the efficacy of our defense, we further consider more general settings with varying α values. We adopt the SOTA KD algorithm ABKD as a representative attack and conduct experiments on Qwen2.5-1.5B using GSM8K. As in Figure C.2, the proposed defense consistently demonstrates strong effectiveness across diverse choices of α, validating the robustness and generalizability of the designed transformation matrix under varying distillation configurations.

Comparison with simple logit-processing baselines. We compare our method with a series of simple logit-processing baselines, including temperature scaling, Top-k truncation, and additive Gaussian noise. We conduct experiments on distillation from Qwen2.5-7B to Qwen2.5-1.5B on GSM8K as a representative setting. We first vary the temperature τ of the teacher model during distillation. As shown in Table C.3, temperature scaling does not provide sufficient defense effectiveness. The student accuracy remains close to the vanilla KD baseline across different temperatures. We further evaluate Top-k truncation with different values of k. As shown in Table C.4, among different k values of Top-k, only k = 1 yields significant defense effectiveness. However, Top-k = 1 fails to satisfy user demands in tasks such as uncertainty estimation, confidence scoring, customized decoding, or reward modeling. For larger k, the defense effect almost vanishes. Our method preserves the complete logit interface, which naturally supports various application scenarios, presenting a more practical trade-off between defense effectiveness and utility in real-world deployment. We also analyze adding Gaussian noise $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ and varying σ to the teacher logits. Table C.5 reports the teacher performance, and Table C.6 reports the distilled student performance. Although a large noise level (*e.g.*, $\sigma = 3$) severely degrades the teacher model, it does not yield a better defense than ours.

As observed, these simple baselines do not provide sufficient defense effectiveness. This may be attributed to the fact that these general logit-processing strategies do not explicitly measure and suppress the distillation-relevant information about the context X contained in Z. As a result, an attacker can still effectively recover the teacher's knowledge through different KD objectives. In contrast, our method achieves substantially stronger defense performance, demonstrating the necessity and effectiveness of our information-theoretic inspired design.

Table C.3: Performance of the distilled student model (Qwen2.5-1.5B) under varying temperature τ of the teacher model on GSM8K (%).

Method	Defense	Student Accuracy	Method	Defense	Student Accuracy
KD	$\tau = 0.1$	62.62	ABKD	$\tau = 0.1$	62.59
	$\tau = 0.3$	62.70		$\tau = 0.3$	62.50
	$\tau = 0.5$	62.70		$\tau = 0.5$	62.76
	$\tau = 0.7$	63.15		$\tau = 0.7$	62.17
	$\tau = 0.9$	62.47		$\tau = 0.9$	61.94
	Vanilla	62.85		Vanilla	62.47
	Ours	**58.53**		Ours	**54.28**

Table C.4: Performance of the distilled student model (Qwen2.5-1.5B) under Top-k truncation of the teacher model on GSM8K (%).

Method	Defense	Student Accuracy	Method	Defense	Student Accuracy
KD	$k=1$	58.91	ABKD	$k=1$	56.48
	$k=4$	62.77		$k=4$	61.41
	$k=8$	62.62		$k=8$	62.24
	$k=16$	62.77		$k=16$	62.09
	Vanilla	62.85		Vanilla	62.47
	Ours	**58.53**		Ours	**54.28**

Table C.5: Performance of the teacher model (Qwen2.5-7B) on GSM8K under Gaussian noise ($\mu = 0$, varying σ, %).

Vanilla	Ours	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$
80.89	**79.83**	79.83	71.42	3.71

Table C.6: Performance of the distilled student model (Qwen2.5-1.5B) on GSM8K under Gaussian noise ($\mu = 0$, varying σ, %).

Method	Defense	Student Accuracy	Method	Defense	Student Accuracy
KD	$\sigma = 0.5$	62.77	ABKD	$\sigma = 0.5$	62.93
	$\sigma = 1.0$	62.85		$\sigma = 1.0$	62.17
	$\sigma = 3.0$	61.41		$\sigma = 3.0$	62.09
	Vanilla	62.85		Vanilla	62.47
	Ours	**58.53**		Ours	**54.28**

D THE PSEUDOCODE OF THE PROPOSED DEFENSE

We provide the pseudocode of the defensive matrix training in Algorithm 1. Note that the formulas of the corresponding loss functions are provided in the main text.

Algorithm 1 Training Defensive Matrices

Input: $f_{\mathrm{T}}(\cdot)$: the teacher model; $f_{\mathrm{S}}(\cdot)$: the proxy student model with parameter θ; $\mathcal{D}$: the training dataset; $|\mathcal{V}|$: the vocabulary size; r: the rank; λ: the loss coefficient; N: the number of training iterations;

Output: Trained defensive matrices $\mathbf{A}$ and $\mathbf{B}$;

Initialize $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times r}$ with $A_{i,j} \sim \mathcal{N}(0,1)$, $\mathbf{B} \in \mathbb{R}^{r \times |\mathcal{V}|}$ with $B_{i,j} = 0$
for iteration $= 1$ to N **do**
  Sample a batch of data $(\boldsymbol{q}, \boldsymbol{o}) \sim \mathcal{D}$;
  Obtain proxy student logits $\boldsymbol{z}_{\mathrm{S}} \leftarrow f_{\mathrm{S}}(\boldsymbol{q})$;
  Obtain teacher logits $\boldsymbol{z}_{\mathrm{T}} \leftarrow f_{\mathrm{T}}(\boldsymbol{q})$;
  $\boldsymbol{z}'_{\mathrm{T}} \leftarrow (\mathbf{E} + \mathbf{AB})\boldsymbol{z}_{\mathrm{T}}$;
  Calculate $\boldsymbol{g}$ via Eq. (10) using $\boldsymbol{q}$, $\boldsymbol{z}_{\mathrm{S}}$ and $\boldsymbol{z}_{\mathrm{T}}$;
  Calculate $\boldsymbol{g}'$ via Eq. (10) using $\boldsymbol{q}$, $\boldsymbol{z}_{\mathrm{S}}$ and $\boldsymbol{z}'_{\mathrm{T}}$;
  Compute the overall training loss $\mathcal{L}_{\mathcal{M}}$ via Eq. (11) using $\boldsymbol{q}, \boldsymbol{o}, \boldsymbol{g}, \boldsymbol{g}'$;
  Back-propagate and update $\mathbf{A}$ and $\mathbf{B}$;
end for
Return $\mathbf{A}, \mathbf{B}$

E LIMITATIONS

Despite the encouraging experimental results, this work still faces several limitations. First, due to the high-dimensional and intractability of text distributions, we do not provide a rigorous theoretical proof that optimizing $\mathcal{L}_{grad}$ is equivalent to maximizing the mutual information $\mathcal{I}(X; Z)$. However, the proposed gradient angle maximization strategy tightly align with the underlying mechanism of

$\mathcal{I}(X;Z)$ maximization, which has also been experimentally proven to be effective. Future work can investigate this issue and propose more rigorous theoretical foundations. Nevertheless, this overhead remains modest in practice, as the number of parameters introduced by the matrix is relatively small (only about 0.1% of the total model parameters). Consequently, the overall impact on inference efficiency is limited and does not significantly affect deployment.

F Visualization Results

We present additional qualitative examples to provide a more intuitive illustration of the impact of our defense on both the teacher and student models.

Teacher model. Figure F.1 and Figure F.2 compare the teacher outputs with and without the transformation matrix on more LLMs and datasets. As observed, the defended teacher continues to generate fluent and logically consistent responses and produces the correct answer, indicating that the defense does not compromise output quality.

Student model. Figure F.3 to Figure F.7 show that, compared to *vanilla* KD, students distilled from transformed logits exhibit noticeably degraded reasoning behavior: their responses become shorter, incomplete, and lack coherent multi-step derivations. Moreover, we observe a collapse in generation behavior—students frequently fall into repetition loops and fail to terminate properly, indicating disrupted decoding dynamics. These findings suggest that the proposed defense suppresses valid information in the teacher logits that are crucial for effective knowledge transfer. As a result, the student is steered toward suboptimal and even pathological optimization trajectories, ultimately impairing both its reasoning ability and generation behavior.

Question

```
3 trees each had 7 blue birds in them.  2 different trees each had
4 blue birds.  1 final tree had 3 blue birds.  How many blue birds
were in the trees in total?
```

Original Teacher

```
Let's think step by step
The total number of blue
birds in the 3 trees is 7
* 3 = 21
The total number of blue
birds in the 2 trees is 4
* 2 = 8
The total number of blue
birds in the 1 tree is 3 *
1 = 3
Adding the blue birds in the
3 trees, 21 + 8 + 3 = 32 blue
birds were in the trees in
total.
The answer is 32
```

Defensive Teacher

```
Let's think step by step
The first tree had 7 blue
birds
The second tree had 7 blue
birds
The third tree had 7 blue
birds
The fourth tree had 4 blue
birds
The fifth tree had 4 blue
birds
The sixth tree had 3 blue
birds
The total number of blue
birds is 7 + 7 + 7 + 4 + 4
+ 3 = 32

The answer is 32
```

Figure F.1: Qualitative comparison between the teacher before and after applying the defensive matrix with Llama-3.1-8B on GSM8K.

Question

Solve the following equation for x:

$$\frac{2}{3} = \frac{4}{x-5}.$$

Original Teacher

Step 1: To solve the equation, we need to isolate the variable x. We can start by cross-multiplying to eliminate the fractions.

Step 2: Cross-multiplying gives us $2(x-5) = 3(4)$.

Step 3: Simplifying the equation, we get $2x - 10 = 12$.

Step 4: Adding 10 to both sides of the equation, we get $2x = 22$.

Step 5: Dividing both sides of the equation by 2, we get $x = 11$.

Therefore, the solution to the equation is $x = \boxed{11}$.

Defensive Teacher

Step 1: To solve the given equation for x, we first need to eliminate the fractions by cross-multiplying. This gives us the equation $2(x-5) = 3(4)$.
Step 2: Next, we simplify the equation by distributing the 2 on the left side and the 3 on the right side. This results in $2x - 10 = 12$.
Step 3: Now, we need to isolate the variable x by adding 10 to both sides of the equation. This gives us $2x = 22$.
Step 4: Finally, we solve for x by dividing both sides of the equation by 2. This results in $x = 11$.
Step 5: Therefore, the solution to the equation is $x = 11$.

Answer: $\boxed{11}$

Figure F.2: Qualitative comparison between the teacher before and after applying the defensive matrix with Qwen2.5-7B on MATH-500.

Question

```
Toula went to the bakery and bought various types of pastries.
She bought 3 dozen donuts which cost $68 per dozen, 2 dozen mini
cupcakes which cost $80 per dozen, and 6 dozen mini cheesecakes for
$55 per dozen.  How much was the total cost?
```

Vanilla Distillation

```
To find the total cost,
we need to calculate the
cost of each type of pastry
separately and then sum them
up.
First, let's calculate the
cost of the donuts:
3 dozen donuts at $68 per
dozen = 3 * $68 = $204
Next, let's calculate the
cost of the mini cupcakes:
2 dozen mini cupcakes at $80
per dozen = 2 * $80 = $160
Now, let's calculate the cost
of the mini cheesecakes:
6 dozen mini cheesecakes at
$55 per dozen = 6 * $55 =
$330
Finally, let's add up the
costs of all the pastries to
find the total cost:
Total cost = cost of donuts +
cost of mini cupcakes + cost
of mini cheesecakes
Total cost = $204 + $160 +
$330 = $694
The total cost of the
pastries was $694.
```

Defensive Distillation

```
The cost of the donuts is 3 x
$68 = $«3*68=204»204.
The cost of the mini cupcakes
is 2 x $80 = $«2*80=160»160.
The cost of the mini
cheesecakes is 6 x $55 =
$«6*55=330»330.
Therefore, the total cost
is $204 + $160 + $330 =
$«204+160+330=794»794.
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
#### 794
......(more repeated)
```

Figure F.3: Qualitative comparison between the student before and after applying the defensive matrix with Qwen2.5-1.5B on GSM8K.

Question

15 gallons of gas were equally divided into 5 different containers. Josey needed 1/4 of a container to run her lawnmower. How many pints of gasoline did Josey need?

Vallina Distillation

Let's think step by step
The total number of gallons of gas is 15 * 4 = 60
The number of containers is 5
The number of gallons in each container is 60 / 5 = 12
The number of pints in each container is 12 * 2 = 24
The number of pints Josey needs is 24 * 1/4 = 6
The answer is $\boxed{6}$

Defensive Distillation

Let's think step by step
The total number of gallons of gas is 15 * 4 = 60
The number of gallons of gas that Josey needs is 60 / 5 = 12
The number of pints of gasoline that Josey needs is 12 * 2 = 24
The answer is $\boxed{24}$

Figure F.4: Qualitative comparison between the student before and after applying the defensive matrix with Llama-3.2-1B on GSM8K.

Question

The idea of "using someone as a mere means" is centrally relevant to which of the following moral theories?
A. consequentialism
B. Kantianism
C. utilitarianism
D. natural law theory

Vallina Distillation

Answer: $\boxed{B}$

The idea of "using someone as a mere means" is centrally relevant to which of the following moral theories?
A. consequentialism
B. Kantianism
C. utilitarianism
D. natural law theory
Answer: $\boxed{B}$

Defensive Distillation

The idea of "using someone as a mere means" is a key concept in the "torture" case, and it is a central part of the "torture" case. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. (more continued)

Figure F.5: Qualitative comparison between the student before and after applying the defensive matrix with Qwen2.5-1.5B on MMLU.

Question

A Senate committee has 8 Republicans and 6 Democrats. In how many ways can we form a subcommittee of 5 members that has at least one member from each party?

Vallina Distillation

We can use the principle of complementary counting. There are a total of $\binom{14}{5} = 2002$ ways to choose a subcommittee of 5 members from the 14 senators. We can count the number of ways to choose a subcommittee of 5 members that has only Republicans or only Democrats. There are $\binom{8}{5} = 56$ ways to choose a subcommittee of 5 Republicans, and $\binom{6}{5} = 6$ ways to choose a subcommittee of 5 Democrats. Therefore, there are $56 + 6 = 62$ ways to choose a subcommittee of 5 members that has only Republicans or only Democrats. Thus, the number of ways to choose a subcommittee of 5 members that has at least one member from each party is $2002 - 62 = \boxed{1940}$. The answer is 1940

Defensive Distillation

We can count the number of ways to form a subcommittee of 5 members that has at least one member from each party by subtracting the number of ways to form a subcommittee of 5 members that has only Republicans from the number of ways to form a subcommittee of 5 members that has only Democrats. There are $\binom{8}{5} = 56$ ways to choose 5 Republicans from 8 Republicans, and there are $\binom{6}{5} = 6$ ways to choose 5 Democrats from 6 Democrats. Therefore, there are $56 - 6 = \boxed{50}$ ways to form a subcommittee of 5 members that has at least one member from each party.
The answer is: 50

Figure F.6: Qualitative comparison between the student before and after applying the defensive matrix with Qwen2.5-1.5B on MATH-500.

Question

Over the last several years, the economy of a state has substantially changed. The state's economy used to be based solely on heavy industry. However, the state legislature approved legalized gambling within the state. As a consequence, many casinos and new hotels were built and the state's economy boomed. These moves were often induced by the granting by the state of special tax benefits for the construction of new casinos and hotels under state statutes. Recently, however, neighboring states have legalized gambling and offered greater tax incentives to the gaming industry. As a result, many of the casino and hotel owners have begun to leave the state. The unemployment and social welfare benefits the state has had to pay have substantially increased, burdening the remaining casinos, and also making it difficult for the state to lower its taxes to remain competitive with other states. On account of this predicament, the state legislature passed, and the governor duly signed, an emergency bill into law. According to the statute, the state imposed a one cent tax on the playing of any slot machine in any gambling casino. Since virtually all the slot machines required a payment of either a dime, quarter, or dollar, the imposition of this tax required a major costly adaptation on each slot machine to allow for the deposit of the additional one cent tax. Although many casino owners have complained about the tax, their only alternative is to absorb the tax themselves and lose one cent per game. As a consequence of the tax, fewer slot machines are purchased in the state by the casino owners. No manufacturer of slot machines is located in the state. Which of the following constitutional provisions provide the strongest ground to attack the validity of the state tax bill?
A. The commerce clause.
B. The equal protection clause of the Fourteenth Amendment.
C. The due process clause of the Fourteenth Amendment.
D. The privileges and immunities clause of Article IV, Section 2.

Vallina Distillation

Answer: $\boxed{A}$

The passage states that the state's economy was based on heavy industry. The passage also states that the state's economy was based on heavy industry. The passage does not state that the state's economy was based on the gambling industry.

Defensive Distillation

The contract clause.

$\boxed{D}$

Figure F.7: Qualitative comparison between the student before and after applying the defensive matrix with Llama-3.2-1B on MMLU.