On Teacher Hacking in Language Model Distillation

Daniil Tiapkin [1]   **Daniele Calandriello** [2]   **Johan Ferret** [2]   **Sarah Perrin** [2]
Nino Vieillard [2]   **Alexandre Ramé** [2]   **Mathieu Blondel** [2]

Abstract

Post-training of language models (LMs) increasingly relies on the following two stages: (i) knowledge distillation, where the LM is trained to imitate a larger teacher LM, and (ii) reinforcement learning from human feedback (RLHF), where the LM is aligned by optimizing a reward model. In the second RLHF stage, a well-known challenge is reward hacking, where the LM over-optimizes the reward model. Such phenomenon is in line with Goodhart's law and can lead to degraded performance on the true objective. In this paper, we investigate whether a similar phenomenon, that we call teacher hacking, can occur during knowledge distillation. This could arise because the teacher LM is itself an imperfect approximation of the true distribution. To study this, we propose a controlled experimental setup involving: (i) an oracle LM representing the ground-truth distribution, (ii) a teacher LM distilled from the oracle, and (iii) a student LM distilled from the teacher. Our experiments reveal the following insights. When using a fixed offline dataset for distillation, teacher hacking occurs; moreover, we can detect it by observing when the optimization process deviates from polynomial convergence laws. In contrast, employing online data generation techniques effectively mitigates teacher hacking. More precisely, we identify data diversity as the key factor in preventing hacking. Overall, our findings provide a deeper understanding of the benefits and limitations of distillation for building robust and efficient LMs.
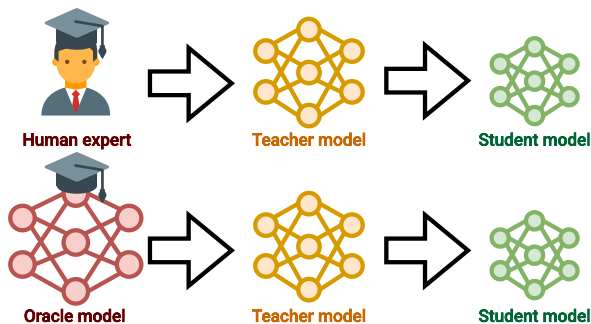
Figure 1. **Overview of our controlled experimental setup.** Usually, the teacher model is trained on expert data before being distilled into the student LM. In the controlled setup of this paper, the teacher is itself distilled from an additional oracle model. This oracle model allows us to measure the quality of the distillation process into the student, and to reveal "teacher hacking".

1. Introduction

Distillation for post-training LMs. Language models (LMs) have achieved remarkable success across a wide range of natural language processing tasks, such as translation, summarization, and reasoning. Notably, large LMs demonstrate impressive generalization capabilities, but their high computational cost poses a significant challenge, particularly when deployed on resource-constrained devices. Efficiency considerations motivate the training of smaller LMs, that would ideally provide similar performance at a fraction of the computational cost. To this end, the most popular approach is knowledge distillation (KD) (Hinton et al., 2015), in which a smaller student LM is trained to imitate the larger teacher LM. Distillation is increasingly studied (Agarwal et al., 2024; Gu et al., 2024; Kim et al., 2024) and used, notably for the post-training pipelines of LMs (as demonstrated by Zephyr (Tunstall et al., 2023), Gemma-2 (Riviere et al., 2024), and DeepSeek-V3 (Liu et al., 2024)), just before the final reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022) phase.

Teacher as an imperfect proxy. However, a key understudied limitation of KD is that the teacher model does not represent the ground-truth distribution but instead acts as

---

[1]CMAP, École Polytechnique, Palaiseau, France; Work done during an internship at Google DeepMind. [2]Google DeepMind. Correspondence to: Daniil Tiapkin <daniil.tiapkin@polytechnique.edu>.

an imperfect proxy for it (Menon et al., 2021; Zhang et al., 2023). This viewpoint draws parallels to a well-studied phenomenon in RLHF known as reward hacking (Amodei et al., 2016; Pan et al., 2022; Gao et al., 2023). Reward hacking is a manifestation of Goodhart's law arising when LMs over-optimize the reward model, trained to represent human preference, thus also an imperfect proxy for the true task objective. The consequences of reward hacking in RLHF can be significant, leading to models misleading humans and producing unsafe behaviors (Hendrycks et al., 2021; Wen et al., 2025).

Teacher hacking. Inspired by the analogy with reward hacking in RLHF, we define teacher hacking as a possible phenomenon during distillation in which the student LM learns to imitate the teacher, not by better approximating the true data distribution, but by exploiting imperfections in the teacher. This raises natural questions: *(1) Does teacher hacking occur in practice?* and if so, *(2) when does it appear?*, and *(3) what strategies can mitigate it?*

Controlled experimental setup. To answer these questions, we propose a controlled experimental setup. Specifically, we introduce an oracle model that represents the ground-truth; see Figure 1 for an illustration. In this setup, the distance between the student and oracle LMs serves as a "golden metric". Conversely, the distance between student and teacher LMs, optimized during the fine-tuning of the student, defines a "proxy metric". Our setup is semi-synthetic, in the sense that prompts are sampled from real datasets but responses are sampled from the oracle LM to learn the teacher, and from the teacher LM to learn the student.

We summarize our main contributions as follows.

- We introduce and formally define the phenomenon of teacher hacking in LM distillation, providing a novel controlled experimental framework to systematically measure its importance.

- We show that teacher hacking occurs when distillation is performed on a fixed offline dataset. Notably, we empirically found that the proxy metric initially follows a polynomial convergence law, and that teacher hacking emerges when the optimization process deviates from this law.

- We analyze strategies to mitigate teacher hacking. Notably, we show that teacher hacking is absent when using online data generation methods, sampling directly from the teacher or the student during distillation. We relate this success to increased diversity in the dataset. We then propose cheaper practical alternative strategies to mitigate teacher hacking, such as increasing the diversity across prompts or using a larger amount of offline teacher-generated data.

2. Preliminaries

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the spaces of all possible prompts and responses, assumed to be sentences in the vocabulary Σ. For two sets A and B, $\Delta(A)$ denotes the space of probability measures over A, and $\Delta(A|B)$ represents the space of conditional probability measures over A given B. An auto-regressive language model $\pi \in \mathsf{LM}_\Sigma(\mathcal{X})$ is defined as the conditional probability of the next token or the end-of-sequence token given a prompt $x \in \mathcal{X}$ and a partial generation $y \in \mathcal{Y}$, expressed as $\pi(\omega|x, y)$, where $\omega \in \Sigma$. In practice, the probability $\pi(\omega|x, y)$ is defined using a softmax with temperature $\pi(\omega|x, y) \propto \exp(\frac{1}{\tau} z(\omega|x, y))$, where $z(\cdot|x, y)$ are the logits output by the neural network and τ is a temperature parameter.

For any language model π, the induced distribution over responses is given by $p_\pi(y|x) \triangleq \prod_{i=1}^{|y|} \pi(y_i|x, y_{:i})$, where $|y|$ denotes the length of the response y, i.e., the number of tokens (non-empty characters) in y, and $y_{:i} \triangleq (y_1, y_2, \ldots, y_{i-1})$.

Distances between language model distributions. Let $d \in \Delta(\mathcal{X})$ be a distribution over prompts induced by the particular task dataset. To measure the convergence of one distribution, induced by a language model π, to a distribution induced by a language model π', we use the (expected) forward and reverse KL between the corresponding conditional measures p_π and $p_{\pi'}$, defined as $\mathrm{KL}_{\mathrm{seq}}(\pi, \pi') \triangleq \mathbb{E}_{x \sim d(\cdot)}[\mathrm{KL}(p_\pi(\cdot|x), p_{\pi'}(\cdot|x))]$. By the properties of the KL divergence, the sequence-level divergence can be estimated very efficiently using token-level KL divergences. Additionally, we use a sequence-level Jensen-Shannon divergence $\mathrm{JS}_{\mathrm{seq}}(\pi, \pi')$. We provide more details in Appendix B.

Supervised fine-tuning. Let $\rho \in \Delta(\mathcal{Y}|\mathcal{X})$ be a conditional response distribution that encodes the ground-truth response distribution. Solving downstream tasks such as summarization, translation, reasoning, or instruction following is fundamentally equivalent to approximating ρ. Therefore, the ultimate goal of any post-training pipeline is to approximate ρ in order to address these tasks effectively.

One of the common approaches to this problem is supervised fine-tuning (SFT). Let us assume that we have a dataset of pairs (x, y) for $x \sim d(\cdot)$ and $y \sim \rho(\cdot|x)$. Then, to find a language model π such that its conditional distribution p_π approximates ρ, it is common to use a simple log-loss

$$\mathcal{L}_{\mathrm{SFT}}(\pi) \triangleq \mathbb{E}_{x \sim d(\cdot), y \sim \rho(\cdot|x)}[-\log p_\pi(y|x)]. \quad (1)$$

This loss is equal, up to a constant factor, to an expected sequence-level forward KL divergence between ρ and p_π: $\mathbb{E}_{x \sim d(\cdot)}[\mathrm{KL}(\rho, p_\pi)]$.

Language model distillation. We suppose that we have access to a teacher language model, denoted $\pi_t \in \mathsf{LM}_\Sigma(\mathcal{X})$, such that it approximates the ground-truth distribution ρ, that is, $p_t = p_{\pi_t} \approx \rho$. The goal of language model distillation is to train a student language model, denoted π_s, so as to approximate the teacher model, that is, $p_s = p_{\pi_s} \approx p_t$. We emphasize that the teacher-induced distribution p_t is not equal to a ground-truth ρ but only approximates it.

We usually distinguish between *hard* and *soft* distillation. In hard distillation, the student is only trained from the teacher's predicted next tokens, i.e., the student only sees the most likely tokens according to the teacher. In soft distillation, the student is trained from the teacher's predicted next token distributions, giving much more information to the student. In this work, we focus on soft distillation, and the loss function for this procedure takes the form

$$\mathcal{L}_{\mathrm{KD}}(\pi_s) \triangleq \mathbb{E}\left[\frac{1}{|y|}\sum_{i=1}^{|y|}\ell_{\mathrm{token}}(\pi_s(\cdot|x, y_{:i}), \pi_t(\cdot|x, y_{:i}))\right],$$

(2)

where $x \sim d(\cdot)$ and $y \sim \nu(\cdot|x)$, $\nu \in \Delta(\mathcal{Y}|\mathcal{X})$ is a *data source*, and ℓ_{token} is a token-level loss function between two distributions over the vocabulary.

The token-level loss and the data source should satisfy two assumptions: (i) it is non-negative and satisfies the property $\ell_{\mathrm{token}}(p, q) = 0$ if and only if $p = q$ for any two distributions $p, q \in \Delta(\Sigma)$, and (ii) the support of $\nu(\cdot|x)$ includes the support of teacher-induced conditions measure $p_t(\cdot|x)$ for almost all x, i.e. $d(x) > 0$. Given these two assumptions, it is easy to show that a language model π minimizes the loss $\mathcal{L}_{\mathrm{KD}}(\pi_s) = 0$ if and only if $\pi_s = \pi_t$.

In particular, considering $\nu(\cdot|x)$ induced by an offline dataset generated by a teacher and using $\ell_{\mathrm{token}}(p, q) = \mathrm{KL}(p, q)$, we achieve the same expected loss as in the case of SFT. This approach corresponds to the works of (Hinton et al., 2015; Sanh et al., 2020). However, we could consider different token-level losses, such as reverse KL divergence (Gu et al., 2024), generalized Jensen-Shannon divergence (Agarwal et al., 2024), or skewed KL divergence (Ko et al., 2024). The data source can be induced by sampling online from the teacher model (Kim & Rush, 2016), sampling online from the student model (Gu et al., 2024), or combining offline and online data (Lin et al., 2020; Agarwal et al., 2024; Ko et al., 2024).

Offline vs. online data sources. In this paper, we distinguish between two different types of data sources: offline and online. Offline data sources are based on a fixed dataset, denoted as $\mathcal{D}_{\mathrm{offline}} = \{(x_i, y_i)\}_{i=1}^M$, where x_i is sampled from $\mathcal{D}_{\mathrm{prompt}}$, and $y_i \sim p_t(\cdot|x_i)$ are responses generated by the teacher model. Importantly, the dataset $\mathcal{D}_{\mathrm{offline}}$ does not need to have a one-to-one correspondence between prompts

and responses. Instead, it may include multiple responses per prompt, and the number of prompts is not necessarily equal to the total number of available prompts N. Each training batch is sampled directly from this fixed dataset.

Online data sources, by contrast, involve generating responses dynamically through an online sampling procedure. For each training batch of prompts $\{x_j\}_{j=1}^B$ of size B, sampled from $\mathcal{D}_{\mathrm{prompt}}$, a corresponding model (either the teacher or the student) generates a batch of responses $\{y_j\}_{j=1}^B$. While sampling from the student model is often referred to as on-policy generation in the literature (Agarwal et al., 2024), we use the term online student to emphasize its parallel to the online teacher data source.

The distinction between offline and online data is particularly evident during subsequent epochs: for offline data sources, responses remain fixed across epochs, while for online data sources, new responses are generated for each epoch, independently drawn from the same distribution.

Teacher hacking. As we already mentioned, the main goal of the post-training pipeline is to approximate the ground-truth distribution ρ by the student model. This goal could be achieved by SFT given access to sufficiently many samples from ρ. However, recent works have shown that distilling from a teacher model can actually work better since the whole next-token distribution contains much richer information than only sampled tokens. An understudied problem is that the teacher model is only an imperfect proxy for ρ. This may lead to a problematic situation where the student LM learns to imitate the teacher, not by better approximating the true data distribution but by exploiting imperfections in the teacher. We call this phenomenon teacher hacking, and give a more formal definition below.

> **Definition 1** (Teacher hacking). Let $\{p_s^{(k)}\}_{k=1}^\infty$ be a sequence of conditional response distributions induced during the training of a student model, p_t the distribution induced by the teacher model, and ρ the target human expert conditional distribution. We say that $\{p_s^{(k)}\}_{k=1}^\infty$ exhibits the teacher hacking phenomenon with respect to a distance measure $\mathrm{dist}\colon \Delta(\mathcal{Y}|\mathcal{X}) \times \Delta(\mathcal{Y}|\mathcal{X}) \to \mathbb{R}_+$ if, as $k \to +\infty$, $\mathrm{dist}(p_s^{(k)}, p_t)$ decreases while $\mathrm{dist}(p_s^{(k)}, \rho)$ increases.

A simple example where this would occur is when the student model is initially closer to the target distribution ρ than the teacher model. However, in more realistic scenarios that are closer to real-world applications, the teacher model is larger and provides a better approximation of ρ than the student model.
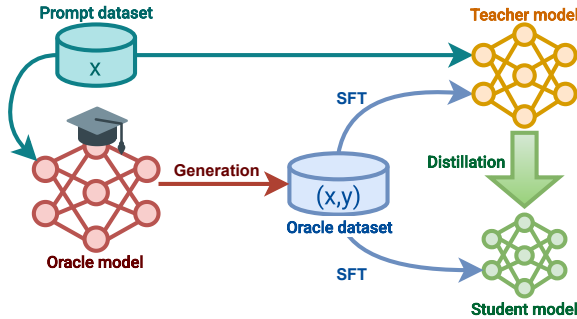
Figure 2. **Overview of the training pipeline.** Two stages: (1) prompts x from a task-specific real dataset are used by the oracle model to generate the oracle pairs (x, y), and afterwards, this dataset is used to get initial SFT checkpoints for both teacher and student model; (2) prompts from the same distribution are used to perform knowledge distillation, where the teacher model serves as a proxy to train the student model.

Overfitting vs. teacher hacking. We would like to clarify the difference between two related phenomena: classical overfitting and the newly-introduced teacher hacking. In the case of overfitting, the model continues to minimize the loss on the training set but fails to do so on the held-out validation set, which it never observes. In contrast, teacher hacking occurs when the model successfully achieves its objective from the teacher's perspective, even on the validation dataset, but fails to improve in terms of approximating the ground-truth behavior. The ultimate outcome of both phenomena is similar: the model fails to generalize to the ground truth, but the reasons differ.

3. Methodology

To analyze the effect of teacher hacking, we require a method to estimate the distance between the student model and the ground-truth distribution. For this purpose, we introduce the *oracle model*, denoted as $\mu \in \mathsf{LM}_\Sigma(\mathcal{X})$, which is assumed to induce the target distribution ρ, i.e., $p_\mu = \rho$.

Golden and proxy metrics. As outlined in Definition 1, evaluating the teacher hacking phenomenon requires computing two sets of metrics.

Golden metrics are computed using the oracle model and reflect the performance with respect to the true objective. Specifically, we use three types of divergences: the forward KL divergence $\mathrm{KL}_{\mathrm{seq}}(\mu, \pi_\mathsf{s})$, the reverse KL divergence $\mathrm{KL}_{\mathrm{seq}}(\pi_\mathsf{s}, \mu)$, and a Jensen-Shannon-like distance $\mathrm{JS}_{\mathrm{seq}}(\pi_\mathsf{s}, \mu)$, which is closely related to the Jensen-Shannon divergence between the conditional distributions over the response space. These metrics are estimated using a held-out validation set of prompts, with sampling performed from the respective models.

Proxy metrics, in contrast, do not rely on the oracle model and instead measure the alignment between the student and teacher models. These metrics use the same types of distances: the forward KL divergence $\mathrm{KL}_{\mathrm{seq}}(\pi_\mathsf{t}, \pi_\mathsf{s})$, the reverse KL divergence $\mathrm{KL}_{\mathrm{seq}}(\pi_\mathsf{s}, \pi_\mathsf{t})$, and the Jensen-Shannon divergence $\mathrm{JS}_{\mathrm{seq}}(\pi_\mathsf{s}, \pi_\mathsf{t})$. Similar to the golden metrics, proxy metrics are estimated using the validation set of prompts and sampling from the models involved.

Training. The training procedure for our experiments consists of two stages.

In the first stage, supervised fine-tuning is performed on both the teacher and student models using a small oracle-generated dataset $\mathcal{D}_{\mathrm{oracle}} = \{(x_i, y_i)\}_{i=1}^{N_{\mathrm{oracle}}}$. The prompts x_i are sampled from the task distribution $d(\cdot)$, and the responses $y_i \sim \rho(\cdot|x_i)$ are generated using the oracle model. Our setup is semi-synthetic, in the sense that prompts are sampled from real datasets but responses (seen as labels) are sampled from the oracle LM to learn the teacher, and from the teacher LM to learn the student. This stage is the only place where direct information from the oracle model is propagated to the teacher and student models. The fine-tuning process optimizes a usual SFT loss (1) that in expectation equals to a sequence-level distance $\mathrm{KL}_{\mathrm{seq}}(\mu, \pi)$. The best checkpoint is selected based on the estimate of this quantity over the validation set.

In the second stage, distillation is conducted from the teacher to the student model by optimizing the soft-distillation loss (2). The distillation process uses a training dataset of unlabeled prompts $\mathcal{D}_{\mathrm{prompt}} = \{x_i\}_{i=1}^N$, where $N \gg N_{\mathrm{oracle}}$ and different data sources that define a distribution of $y_i \sim \nu(\cdot|x)$ in the loss. The final pipeline is summarized in Figure 2.

Evaluation. To investigate the teacher hacking phenomenon, we analyze two key types of curves: (1) the dependence of the training loss, proxy metrics, and golden metrics on the number of epochs completed, and (2) the proxy-golden curve, which illustrates the relationship between the golden metric (only accessible in our controlled experimental setup) and the proxy metric. Both proxy and golden metrics are computed using a held-out validation set of prompts. The final pipeline is summarized in Figure 3.

The epoch-dependence plots provide insights into scaling law phenomena (Kaplan et al., 2020) and help to understand overall training dynamics. Proxy-golden curves are crucial for visually assessing the presence of teacher hacking: a U-shaped curve serves as a clear indicator. Indeed, we expect the proxy metric to be reduced during training, and if the golden metric first decreases and then increases, it directly shows teacher hacking. These proxy-golden plots can be compared to plots from Gao et al. (2023), with one essential
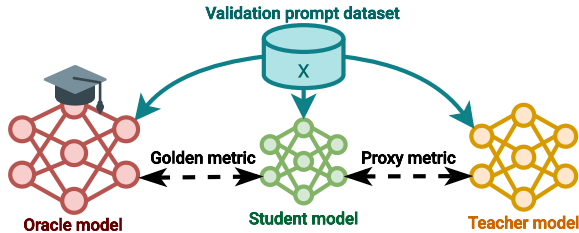
Figure 3. **Overview of the evaluation pipeline.** We use the validation prompt dataset to measure the golden metric (the distance between the oracle and the student models) and the proxy metric (the distance between the teacher and the student models).

Figure 4. **Proxy-Golden plot (offline data source).** We distill a T5-large teacher into a T5-base student on the XSUM dataset. The token-level training loss is the forward KL, the proxy metric is the distance to the teacher distribution and the golden metric is the distance to the ground-truth (oracle) distribution (available thanks to our semi-synthetic controlled experimental setup). In this plot, the x-axis (proxy metric) indicates optimization progress, and the y-axis shows the ground-truth performance (golden metric): lower is better. Teacher hacking occurs in the case of offline data source: the orange curve has a U-type shape, indicating that during optimization, the orange metric starts increasing, whereas the proxy metric continues to decrease.

difference: our approach measures optimization progress as the distance to the teacher model rather than the distance from an initial reference policy.

4. Experimental Results

Oracle, teacher, and student models. Our experiments use a family of encoder-decoder language models based on T5 (Raffel et al., 2020; Roberts et al., 2022). The oracle model is the Flan-T5-XL (Chung et al., 2024), a 3B-parameter model fine-tuned on the Flan dataset (Wei et al., 2021; Longpre et al., 2023) for instruction-based tasks. For the teacher and student models, we use pretrained checkpoints of T5-1.1 in three configurations: small (77M parameters), base (250M parameters), and large (800M parameters). We always use temperature sampling with a temperature parameter $\tau = 1$ for generations from any model.

Datasets. Our experiments use three datasets for training and evaluation: the XSum summarization dataset (Narayan et al., 2018), the WMT-14 en-de translation dataset (Bojar et al., 2014), and the instruction-following dataset Natural Instructions (Mishra et al., 2022; Wang et al., 2022). In alignment with our experimental setup, we use only the prompt data from these datasets, supplemented with task-specific instructions as needed for each task.

For the first stage of the training pipeline, where the oracle dataset is build and used for SFT, we use $N_{\text{oracle}} = 25\,000$, $50\,000$, and $100\,000$ prompts from the XSum, WMT-14 en-de, and Natural Instructions datasets, respectively. For the second stage, which involves the knowledge distillation procedure, we use $N = 200\,000$, $450\,000$, and $500\,000$ prompts from these datasets. A single epoch is defined as one complete pass through all N examples, corresponding to $\lceil N/B \rceil$ training steps, where B denotes the batch size. For XSum and WMT-14 en-de, we use batch size $B = 32$; for Natural Instructions, we use batch size $B = 64$.
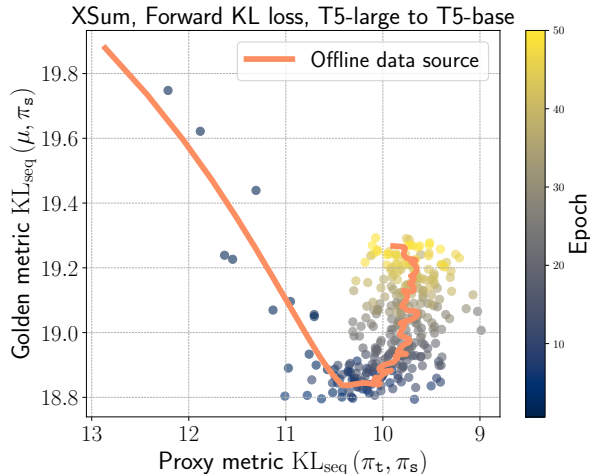
4.1. Does Teacher Hacking Appear?

We begin by investigating whether teacher hacking appears.

Setup. For the first experiment, we use only *offline data sources*: responses are pre-generated as $y_i \sim p_{\text{t}}(x_i)$ for all $x_i \in \mathcal{D}_{\text{prompt}}$, and the dataset remains fixed throughout training. The learning rate for optimization is selected via a grid search over $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$.

The distillation procedure starts from the SFT checkpoints of the teacher and student models. Training is carried out over 50 epochs to analyze long-term convergence behavior.

Results. The results of distilling the T5-large teacher model into the T5-base student on the XSum dataset, using forward KL loss, along with the corresponding golden and proxy metrics, are shown in Figure 4.

In this plot, the x-axis represents the optimization progress in terms of the distance to the teacher model (from left to right), while the y-axis shows the golden metric. The scatter plot shows the exact values of proxy and golden metrics, where the color demonstrates at which epoch this measurement was performed. The curve itself shows a relationship between smoothed values of the proxy and
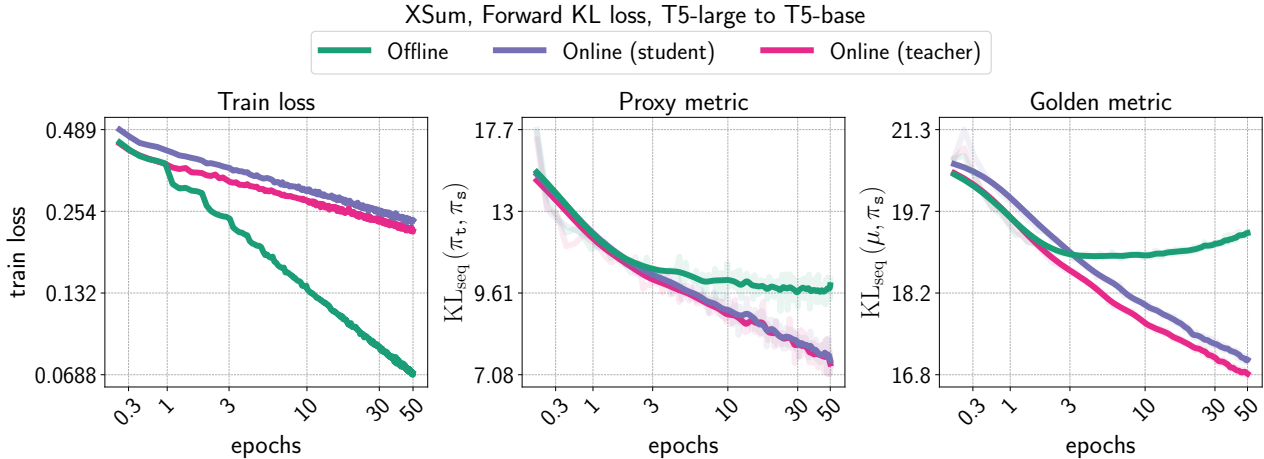
Figure 5. **Impact of using offline vs. online data sources.** When using a fixed offline dataset, though the proxy metric continues to decrease, this is not visible in the golden metric, which continues to increase, a phenomenon we call teacher hacking. However, when using online response sampling, both from the teacher model or from the student model, this phenomenon does not occur.

golden metric, where smoothing is performed by Gaussian smoothing.

For offline data source, the plot exhibits a U-shaped curve. This behavior indicates teacher hacking: as optimization progresses, the ground-truth performance (golden metric) initially improves but eventually deteriorates. Overall, the conclusion of this experiments is following.

> **Observation 1.** Teacher hacking exists and emerges after extended training on a fixed offline dataset.

4.2. When Does Teacher Hacking Appear?

In this subsection, we investigate the conditions under which teacher hacking occurs.

Setup. For this experiment, we evaluate three distinct data sources: (1) *Offline data source*; (2) *Online teacher data source*: for each batch of prompts sampled from $\mathcal{D}_{\mathrm{prompt}}$, a new response $y_i \sim p_{\mathrm{t}}(x_i)$ is dynamically generated by the teacher model; (3) *Online student data source*: responses are generated on-the-fly as $y_i \sim p_{\mathrm{s}}(x_i)$ using the current student model.

As in the previous experiment, we use the forward KL divergence as a token-level loss. We refer to Appendix A for different token-level loss functions, such as reverse KL divergence and Jensen-Shannon divergence.

We analyze the dynamics of proxy and golden metrics for each data source. For the proxy and golden metrics, we apply Gaussian smoothing to smooth the noisy behavior of the curves. We would also like to emphasize that the difference in scaling between the training loss and proxy/golden

metrics is due to averaging over sentence length, which is used in training loss but not in proxy/golden metrics. The training loss indicates the token-level training convergence, whereas the proxy/golden metrics show the sentence-level validation convergence.

Results. The comparison results between offline and online data sources are shown in Figure 5. Additional comparisons for other combinations of datasets, student/teacher model sizes, and loss functions are provided in Appendix A.

We analyze the epoch dependence of the training loss, proxy metric, and golden metric. We present the dependencies on a log-log scale to track possible polynomial convergence laws: the polynomial dependence of the metric on the training time. Overall, we make the following observations.

(i) In the offline data scenario, we validate the presence of teacher hacking: the proxy metric decreases while the golden metric increases after a certain point. This phenomenon does not occur with online data sources.

(ii) We notice that the behavior of all curves for online and offline data sources is different. Overall, the training loss for the offline data sources decreases faster since the training loss is optimized multiple times over the same data, but the performance on the proxy/golden metrics is worse overall.

(iii) The proxy metric for online data sources follows a linear trend on the log-log scale, indicating a polynomial convergence law. In contrast, teacher hacking in offline data coincides with deviations in the proxy metric compared to online data. It gives a mechanism to *detect* teacher hacking using only the proxy metric, that is measurable even in real scenarios (not only in our controlled experimental setup).
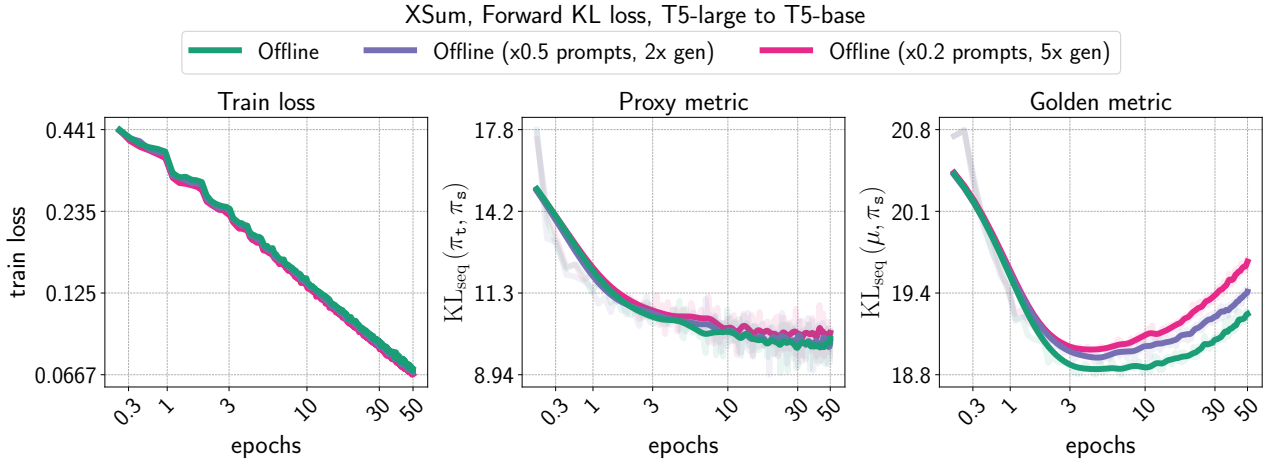
6

Figure 6. **Impact of diversity of offline data sources.** We regulate the diversity of the dataset by decreasing the number of prompts in $2/5$ times and providing $2/5$-times more generations for each existing prompt, while preserving the size of the dataset. Whereas the dynamics of the train loss and proxy metric are almost the same, the effect of teacher hacking becomes more evident with a less diverse dataset.

These results highlight that teacher hacking can harm ground-truth performance, particularly when training involves multiple epochs on the same dataset. However, we would like to emphasize that this issue is not present when training is limited to a small number of epochs (e.g., 1–3), as the golden metric remains stable in these cases. We summarize the conclusions of these first experiments as follows.

> **Observation 2.** Employing online data generation or limiting training to a few epochs effectively prevents teacher hacking.

4.3. How to Mitigate Teacher Hacking?

In the next experiment, we evaluate different methods for modifying the diversity and amount of offline data, in order to investigate how the properties of the offline data affect the teacher hacking phenomenon.

Setup. For this experiment, we evaluate different approaches to constructing an offline data source. We define an *ordinary* offline data source as one that uses all available prompts and a single generation for each prompt: $\mathcal{D}_{\text{offline}} = \{(x_i, y_i) \mid y_i \sim p_t(\cdot \mid x_i)\}_{i=1}^{N}$, where $\mathcal{D}_{\text{prompt}} = \{x_i\}_{i=1}^{N}$.

Our first objective is to study how the diversity of the offline dataset impacts the teacher hacking phenomenon under a fixed dataset generation budget. Let $k \in \mathbb{N}$ be a natural number. To construct a dataset with reduced diversity, we sub-sample $\lceil N/k \rceil$ prompts from $\mathcal{D}_{\text{prompt}}$ and generate k responses for each sampled prompt using the teacher model. The resulting dataset maintains the same generation budget of N total responses but exhibits reduced diversity because the k generations for the same prompt x are closer to each

other compared to k generations for k different prompts. In our experiments, we apply this technique for $k = 2$ and $k = 5$.

Second, we investigate how increasing the generation budget influences teacher hacking. For a fixed integer $m \in \mathbb{N}$, we generate m responses for each prompt in $\mathcal{D}_{\text{prompt}}$, resulting in a dataset of size $m \times N$. Despite the larger dataset size, we define epochs as passing through N data points to ensure comparability across experiments. This setup enables us to interpolate between using an offline data source and an online teacher data source. We use values $m = 2$ and $m = 3$ for our experiments. The rest of the experimental setup follows the description in Section 4.1.

Diversity of offline data sources. We begin by examining the impact of dataset diversity on training dynamics and the previously observed teacher hacking phenomenon. The results are shown in Figure 6. Notably, while the training loss and proxy metric dynamics remain nearly the same, the golden metric behaves differently: lower dataset diversity leads to worse golden metric performance, making the teacher hacking effect more evident.

Generation budget. Next, we examine the effect of a larger generation budget on training dynamics, as shown in Figure 7. Increasing the number of generations per prompt uniformly improves both the proxy and golden metrics over time and, oppositely, alters the training loss dynamics. Overall, this suggests an interpolation between the behavior of an ordinary offline and online teacher data sources.

Discussion. The results suggest two practical strategies to mitigate the effects of teacher hacking:
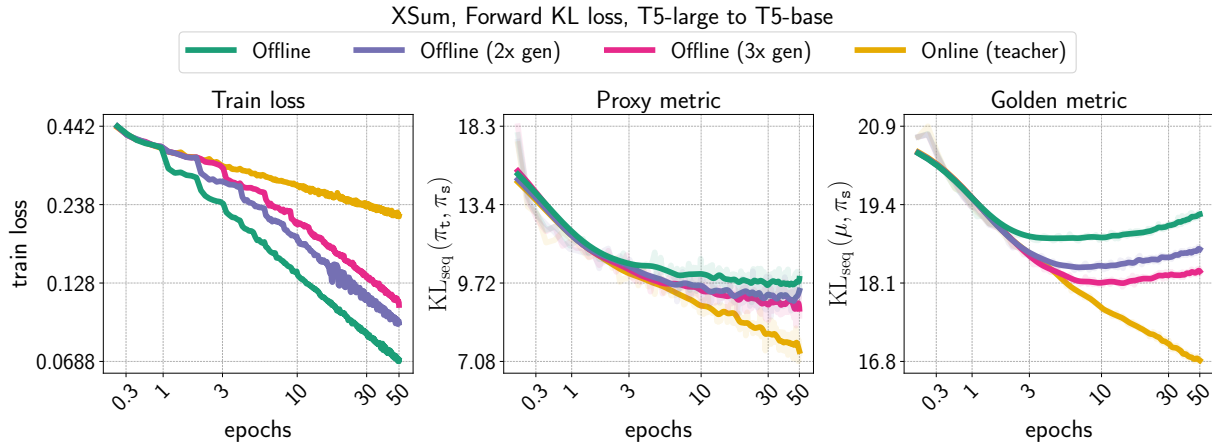
Figure 7. **Impact of generation budget for offline data sources.** As the number of generations per prompt increases, both proxy and golden metrics improve, suggesting that the effect of teacher hacking is decreasing.

> **Observation 3.** *Prioritize Prompt Diversity.* When the generation budget for the distillation dataset is fixed, focusing on increasing the diversity of prompts can help reduce the impact of teacher hacking.

> **Observation 4.** *Expand the Dataset with Multiple Completions.* If the prompt dataset is fixed, increasing the generation budget by generating multiple completions per prompt also helps diminish the effects of teacher hacking.

5. Related Work

Goodhart's law states: "When a measure becomes a target, it ceases to be a good measure" (Strathern, 1997). In particular, it manifests itself as reward hacking in RLHF (Amodei et al., 2016; Gao et al., 2023; Weng, 2024). A line of works studied reward hacking under controlled experimental setups (Gao et al., 2023; Rafailov et al., 2024). Our setup closely resembles that of Gao et al. (2023), where two types of reward models (RMs) are used: a golden reward model, which substitutes the ground-truth reward function, and a proxy reward model, trained on golden RM-preferred generations as ground-truth preferences. Specifically, we employ oracle and teacher models in the same roles as the golden and proxy RMs, respectively: the teacher model is trained on oracle-generated data, while the final student model is trained using the teacher's next-token distribution.

Given possible negative consequences of reward hacking (Hendrycks et al., 2021; Wen et al., 2025), another line of research attempts to mitigate its effects through better reward modeling or more robust training procedures (Chen et al., 2024; Rame et al., 2024; Liu et al., 2025; Miao et al.,

2024).

Knowledge distillation (KD) was initially introduced as a method to compress a large teacher model into a smaller student model without much loss in performance (Buciluă et al., 2006; Hinton et al., 2015); it has then been successfully used to create smaller language models such as DistilBERT (Sanh et al., 2020), Zephyr (Tunstall et al., 2023), Gemma-2 (Riviere et al., 2024), Gemini-1.5 Flash (Georgiev et al., 2024), and DeepSeek-V3 (Liu et al., 2024). One of the actively used approaches to language model distillation is sequence-level KD (Kim & Rush, 2016). In our case, this approach corresponds to utilizing offline data sources and forward KL token-level loss (thanks to properties of the KL). Other approaches focus on matching different quantities of the teacher model by the student model, such as hidden states (Jiao et al., 2020; Zhang et al., 2024) or attention scores (Wang et al., 2020).

Most of the previously-mentioned works implicitly or explicitly assume that the replication of the teacher model represents the final goal of the distillation process, contrary to the approach mentioned by Menon et al. (2021), where the teacher is only an imperfect approximation of the true data distribution. Based on this perspective, Zhang et al. (2023) developed a perturbed loss that can be seen as training from a "proxy teacher". However, they do not further investigate the consequences of over-optimizing the objective.

6. Conclusion

In this paper, we introduce and examine the phenomenon of teacher hacking in language model distillation by designing a semi-synthetic controlled experimental setup. This allows us to measure its effects, and validate experimentally its presence when using a fixed offline dataset for the

distillation procedure.

Fortunately, as a practical outcome of our study, we were able to identify several strategies to mitigate teacher hacking: (1) utilize online generations during the distillation process, (2) when the generation budget is fixed, prioritize increasing the diversity of the prompt dataset, and (3) if the prompt dataset is fixed and online generations are not feasible, generate multiple offline completions per prompt ahead of time to expand the dataset. We hope that these practical and methodological insights provide valuable guidance in extending the applicability and effectiveness of language model distillation in real-world scenarios.

Impact statement

This paper presents work on language model distillation, which is actively used in the training of many modern language models. We identify a possible shortcoming of existing distillation procedures, called teacher hacking, that can lead to the transfer of unsafe behaviors from teacher to student. Additionally, we proposed several strategies to reduce the effect of this phenomenon. We believe that understanding and identifying such issues have positive societal consequences and allow the development of more reliable and safe language models.

References

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Garea, S. R., Geist, M., and Bachem, O. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. Findings of the 2014 workshop on statistical machine translation. In Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., and Specia, L. (eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, June 2014. Association for Computa-

tional Linguistics. doi: 10.3115/v1/W14-3302. URL https://aclanthology.org/W14-3302/.

Buciluă, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Chen, L., Zhu, C., Chen, J., Soselia, D., Zhou, T., Goldstein, T., Huang, H., Shoeybi, M., and Catanzaro, B. ODIN: Disentangled reward mitigates hacking in RLHF. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7935–7952. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24bn.html.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gu, Y., Dong, L., Wei, F., and Huang, M. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kim, G., Jang, D., and Yang, E. PromptKD: Distilling student-friendly knowledge for generative language models via prompt tuning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6266–6282, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp. 364. URL https://aclanthology.org/2024.findings-emnlp.364/.

Kim, Y. and Rush, A. M. Sequence-level knowledge distillation. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Ko, J., Kim, S., Chen, T., and Yun, S.-Y. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lsHZNNoC7r.

Lin, A., Wohlwend, J., Chen, H., and Lei, T. Autoregressive knowledge distillation through imitation learning. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main. 494. URL https://aclanthology.org/2020.emnlp-main.494/.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-V3 Technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Liu, T., Xiong, W., Ren, J., Chen, L., Wu, J., Joshi, R., Gao, Y., Shen, J., Qin, Z., Yu, T., Sohn, D., Makarova, A., Liu, J. Z., Liu, Y., Piot, B., Ittycheriah, A., Kumar, A., and Saleh, M. RRM: Robust reward model training mitigates reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=88AS5MQnmC.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Menon, A. K., Rawat, A. S., Reddi, S., Kim, S., and Kumar, S. A statistical perspective on distillation. In *International Conference on Machine Learning*, pp. 7632–7642. PMLR, 2021.

Miao, Y., Zhang, S., Ding, L., Bao, R., Zhang, L., and Tao, D. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Rafailov, R., Chittepu, Y., Park, R., Sikchi, H. S., Hejna, J., Knox, B., Finn, C., and Niekum, S. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Rame, A., Vieillard, N., Hussenot, L., Dadashi-Tazehozi, R., Cideron, G., Bachem, O., and Ferret, J. WARM: On the benefits of weight averaged reward models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42048–42073. PMLR, 21–27 Jul

2024. URL https://proceedings.mlr.press/v235/rame24a.html.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Roberts, A., Chung, H. W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., Hawthorne, C., Lewkowycz, A., Salcianu, A., van Zee, M., Austin, J., Goodman, S., Soares, L. B., Hu, H., Tsvyashchenko, S., Chowdhery, A., Bastings, J., Bulian, J., Garcia, X., Ni, J., Chen, A., Kenealy, K., Clark, J. H., Lee, S., Garrette, D., Lee-Thorp, J., Raffel, C., Shazeer, N., Ritter, M., Bosma, M., Passos, A., Maitin-Shepard, J., Fiedel, N., Omernick, M., Saeta, B., Sepassi, R., Spiridonov, A., Newlan, J., and Gesmundo, A. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*, 2022. URL https://arxiv.org/abs/2203.17189.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Song, F., Yu, B., Lang, H., Yu, H., Huang, F., Wang, H., and Li, Y. Scaling data diversity for fine-tuning language models in human alignment. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14358–14369, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1251/.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Strathern, M. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33: 5776–5788, 2020.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Supernaturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., and Feng, S. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xJljiPE6dg.

Weng, L. Reward hacking in reinforcement learning. *lilianweng.github.io*, Nov 2024. URL https://lilianweng.github.io/posts/2024-11-28-reward-hacking/.

Zhang, R., Shen, J., Liu, T., Liu, J., Bendersky, M., Najork, M., and Zhang, C. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*, 2023.

Zhang, S., Zhang, X., Sun, Z., Chen, Y., and Xu, J. Dual-space knowledge distillation for large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

WMT-14 en-de, Forward KL loss, T5-large to T5-base

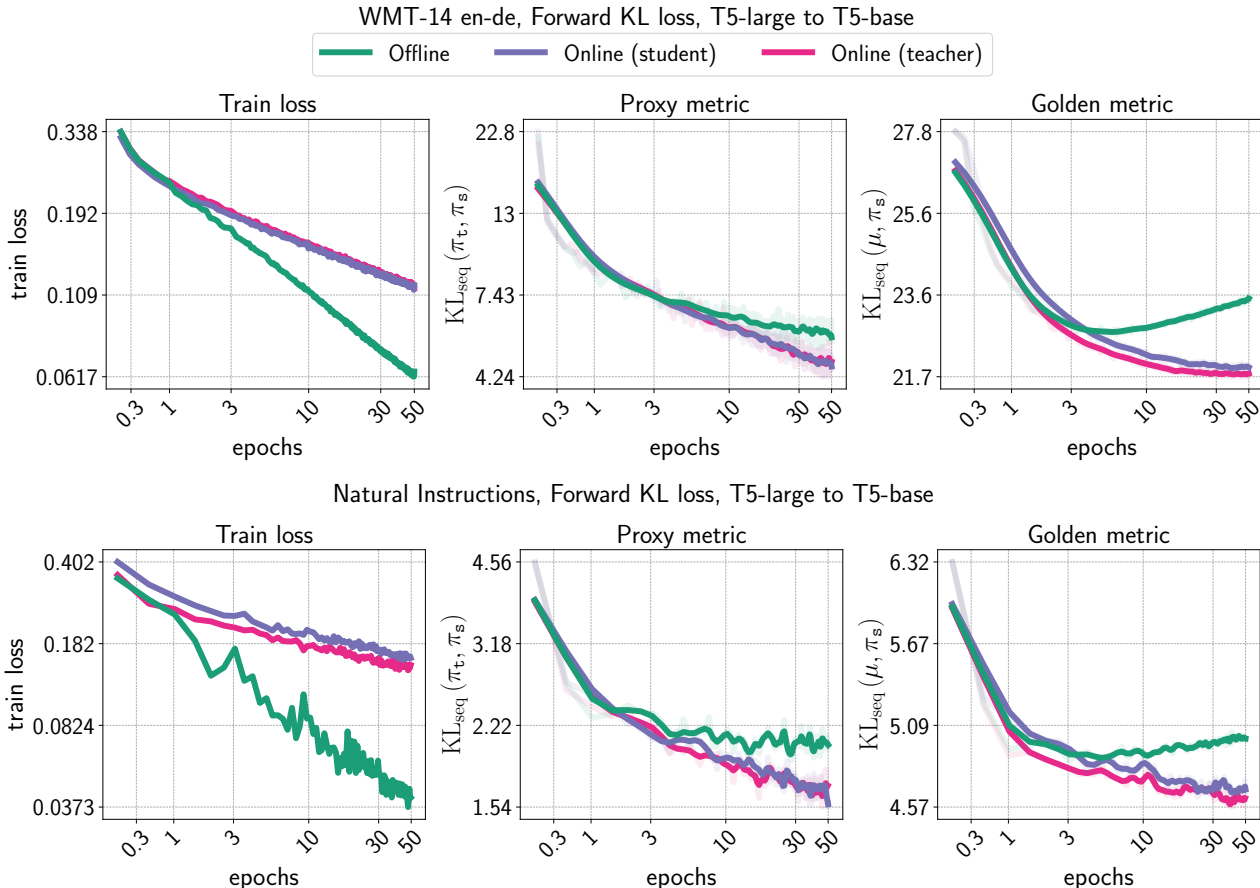Natural Instructions, Forward KL loss, T5-large to T5-base

Figure 8. **Impact of the dataset choice: offline vs. online data sources.** We verify our claims on the presence of teacher hacking in the case of offline data sources for two different tasks: the translation task on WMT-14 en-de (top row) and the instruction following task on Natural Instruction (bottom row). In general, the behavior of the curves is the same across all the datasets: for online data sources, both proxy and golden metrics are decreasing. At the same time, for offline data sources, the proxy metric is decreasing or stagnating, whereas the golden metric is clearly increasing.

A. Additional Experiments

This section presents additional experiments on teacher hacking across various datasets, model sizes, and loss types, as well as an additional experiment on the effect of mixing offline and online data.

A.1. The Impact of the Dataset

In this subsection, we validate our claims across various tasks, such as the translation task on the WMT-14 en-de dataset (Bojar et al., 2014) and the instruction-following task on the Natural Instructions dataset (Mishra et al., 2022; Wang et al., 2022). We consider the same pair of models: T5-large as the teacher model and T5-base as the student model, utilizing forward KL token-level loss along with the corresponding proxy and golden metrics.

Online vs. offline data sources. The results of the comparison of online and offline data sources on the WMT-14 en-de and Natural Instruction datasets are presented in Figure 9.

Overall, we observe the same phenomenon as noted in the summarization task in Section 4. The general patterns are consistent: for all data sources, the training loss decreases slowly for online data sources, as expected, while the proxy metric decreases or stagnates, showing no signs of classical overfitting. However, the golden metric continues to decline in the case of offline data sources, and it starts declining in the case of offline data source, indicating the presence of teacher hacking.
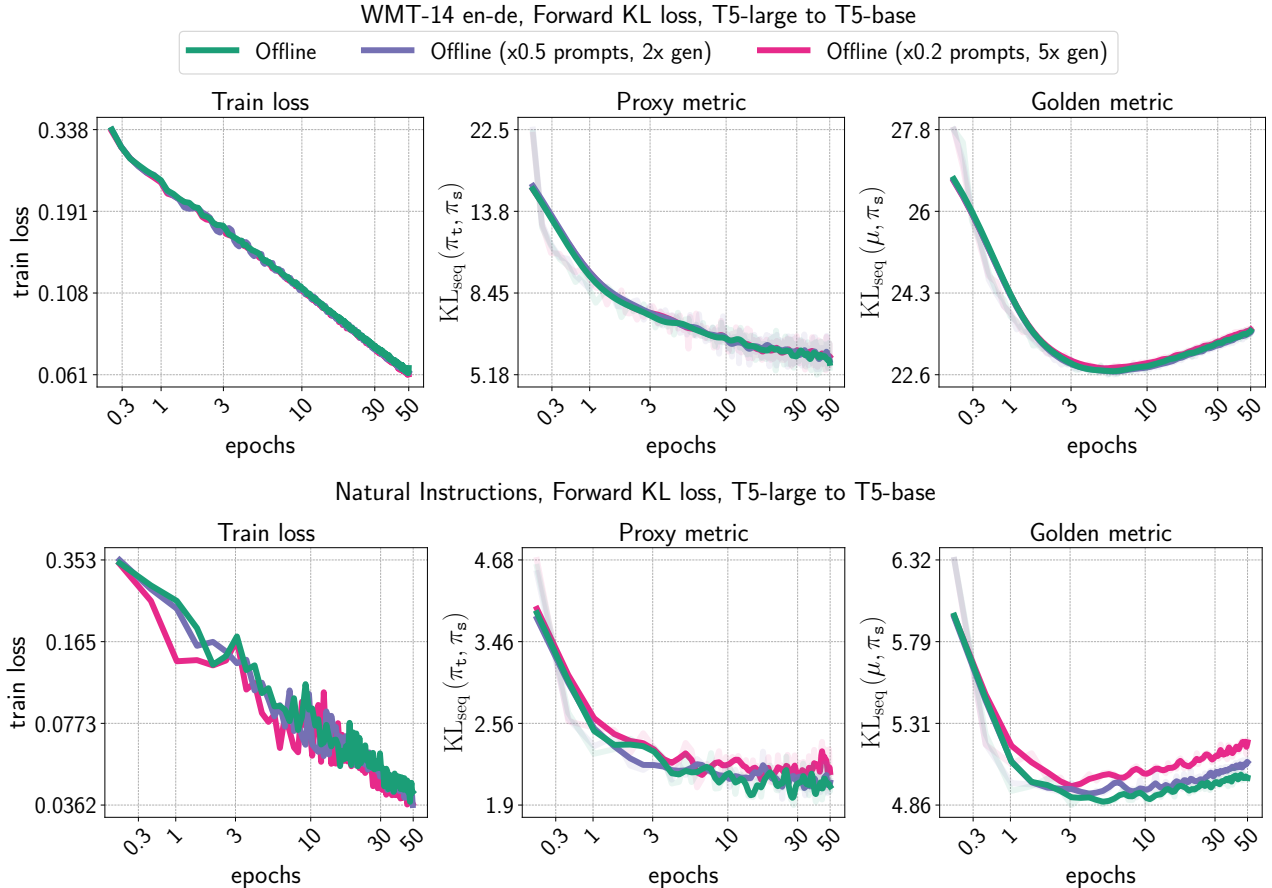
Figure 9. **Impact of the dataset choice: dataset diversity.** Across other datasets, we can notice that the impact of diversity is still present for the instruction-following task but not for the translation task. It can be explained by the initially small diversity of the WMT-14 dataset.

Diversity of offline data sources. Next, we study the impact of the diversity of the dataset on teacher hacking, following the setup described in Section 4.3. The results are presented in Figure 9. We observe the same detrimental effect of the decreasing of the dataset diversity in the case of the instruction following task. However, in the case of the translation task, it almost has no effect. We connect this effect to the small change in diversity of the initial dataset since it contains only relatively short sentences. To validate this claim, we compute the quantitative measurement of diversity, used by Song et al. (2024), that is equal to a ratio of unique bigrams to all bigrams in the tokenized dataset, multiplied by the size of the dataset, see Table 1. In particular, we can see that for WMT-14 dataset, the decay of quantitative diversity with reducing number of unique prompts is less dramatic, explaining a little change from point of view of proxy/golden metrics.

Generation budget. Finally, we verify the claims on increasing the generation budget for the offline data sources. The results are presented in Figure 10. In this case, for both tasks, we observe an improvement in the golden metric, especially for the translation task, and we observe a marginal improvement in the proxy metric. As in the case of the summarization task, it signals the decreasing teacher hacking effect.

A.2. The impact of student and teacher model sizes

In this subsection, we validate our findings across different student and teacher model sizes using the XSum dataset and forward KL token-level loss.

Online vs. offline data sources. We conduct distillation experiments from T5-base to T5-small and from T5-large to T5-small, evaluating performance across offline and online data sources. The results are shown in Figure 11.
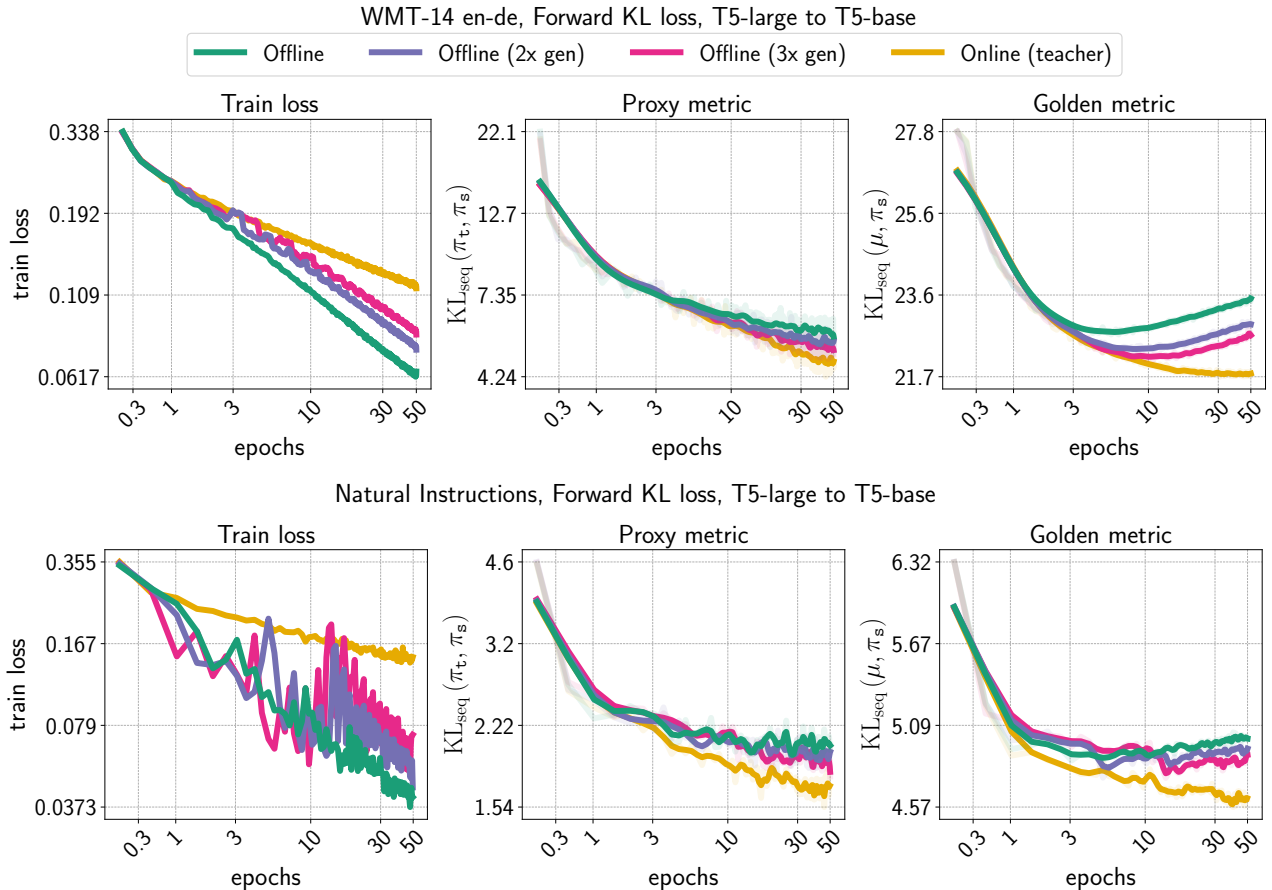
Figure 10. **Impact of the dataset choice: generation budget.** We additionally confirm the claim on the positive impact of a larger number of generations per prompt across two other datasets.

For online data sources, both the proxy and golden metrics decrease monotonically, regardless of the model sizes, confirming our earlier observations. In contrast, for offline data sources, the golden metric consistently increases. Notably, in the case of distilling T5-large to T5-small, the proxy metric also shows a slight increase, indicating standard overfitting rather than teacher hacking. Meanwhile, when distilling T5-base to T5-small, the proxy metric stagnates, suggesting the presence of teacher hacking.

A.3. The Impact of Token-Level Loss Functions

In this subsection, we examine how the choice of token-level loss function and proxy/golden metrics influences the teacher hacking phenomenon. The experiments are conducted using the XSum summarization dataset, with T5-base as the student model and T5-large as the teacher model.

Online vs. offline data sources. This experiment compares different data sources using two token-level loss functions: Jensen-Shannon divergence and reverse KL divergence, alongside their corresponding proxy and golden metrics. The results are shown in Figure 12.

The observed behavior across all plots closely resembles that of the forward KL loss. Specifically, when using online data sources, both proxy and golden metrics decrease monotonically. In contrast, for offline data sources, proxy metrics continue to decrease, but the golden metric increases, signaling the presence of teacher hacking.

Additional experiments were conducted using generalized Jensen-Shannon divergences with coefficients of 0.1 and 0.9. However, their performance was either comparable to or worse than reverse or forward KL divergence for the respective

Table 1. **Diversity measurement.** Following an approach of Song et al. (2024), we report dataset diversity as the ratio of unique bigrams in tokenized prompt-completion pairs to the total number of tokenized bigrams, multiplied by the dataset size. Additionally, we report the dataset diversity ratio to the dataset's diversity with a 1:1 prompt-generation ratio. These measurements explain why the influence of diversity reduction on WMT is much smaller: diversity changes in the prompt-completion dataset are not as dramatic as in the summarization or instruction-following datasets.

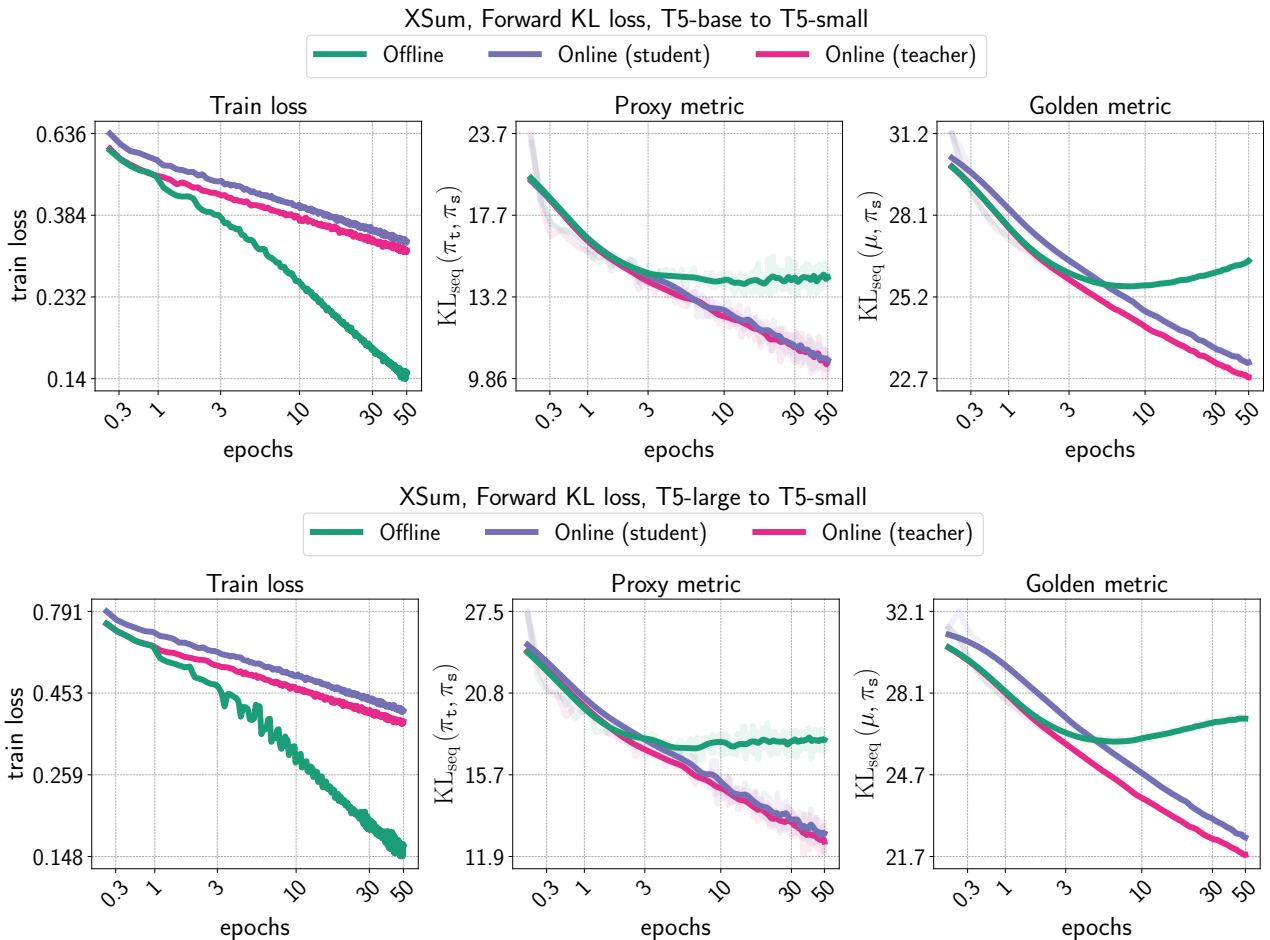| Dataset	Prompt-Generation Ratio				
	0.2:5	0.5:2	1:1	1:2	1:3
XSum	4813.3 (0.47)	7292.1 (0.72)	10140.1 (1.00)	10645.8 (1.05)	11112.9 (1.10)
WMT-14	48079.4 (0.77)	54511.5 (0.87)	62652.0 (1.00)	85855.5 (1.37)	105242.1 (1.68)
Natural Instructions	8464.1 (0.45)	13311.6 (0.71)	18751.5 (1.00)	19631.7 (1.05)	20430.0 (1.08)

Figure 11. **Impact of the model sizes: offline vs. online data sources.** We examine the train loss and proxy/golden metrics for two model size pairs: T5-base as the teacher and T5-small as the student (top), and T5-large as the teacher and T5-small as the student (bottom). For both pairs, no teacher hacking occurs with online data generation. However, teacher hacking is observed during distillation from T5-base to T5-small, as the proxy metric stagnates while the golden metric decreases. In contrast, distillation from T5-large to T5-small shows behavior consistent with standard overfitting as the proxy metric slightly increases. This may be attributed to the larger difference in model sizes.

proxy and golden metrics. As a result, they are excluded from the comparison.
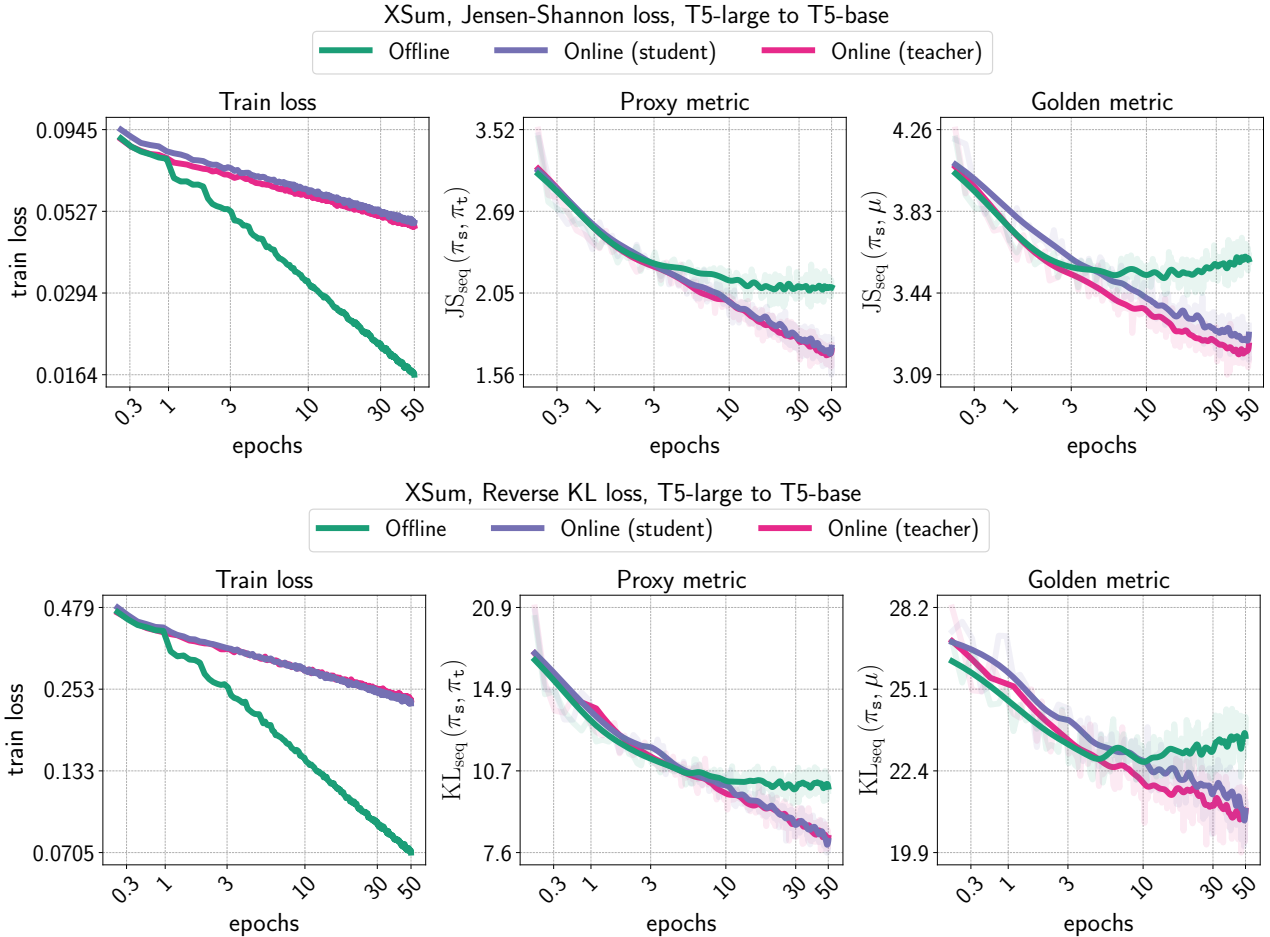
Figure 12. **Impact of the loss type: offline vs. online data sources.** We can observe that the effect of teacher hacking appears regardless of the choice of loss function.

A.4. Offline-Online Data Mixtures

In this subsection, we examine how varying mixtures of offline and online data influence the occurrence of teacher hacking. The experiment uses the XSum dataset, T5-base as the student model, T5-large as the teacher model, and forward KL token-level loss, following Section 4.

We use the following procedure to generate each training batch: with probability α, the batch is sampled from a fixed offline dataset, and with probability $1 - \alpha$, it is generated by the current student model. This process is repeated at each training step. We evaluate three values for α: 0.1, 0.5, and 0.9, corresponding to 10%, 50%, and 90% offline data proportion, respectively. The results are shown in Figure 13.

Our results show that increasing the proportion of *online* data in the distillation process (that is equivalent to decreasing the proportion of offline data) significantly improves the golden metric. Even with 10% online student data (corresponding to 90% offline data), the golden metric plateaus, thus effectively reducing teacher hacking. Higher proportions of online data further mitigate the effect, with 90% online student data resulting in training dynamics nearly identical to those observed when using only student-generated data.

Discussion. These results suggest that exclusively using online-generated data is not required to avoid teacher hacking. Instead, incorporating a fraction of online-generated data during the distillation process is sufficient. In particular, as little as 10% online data can substantially reduce the impact of teacher hacking.
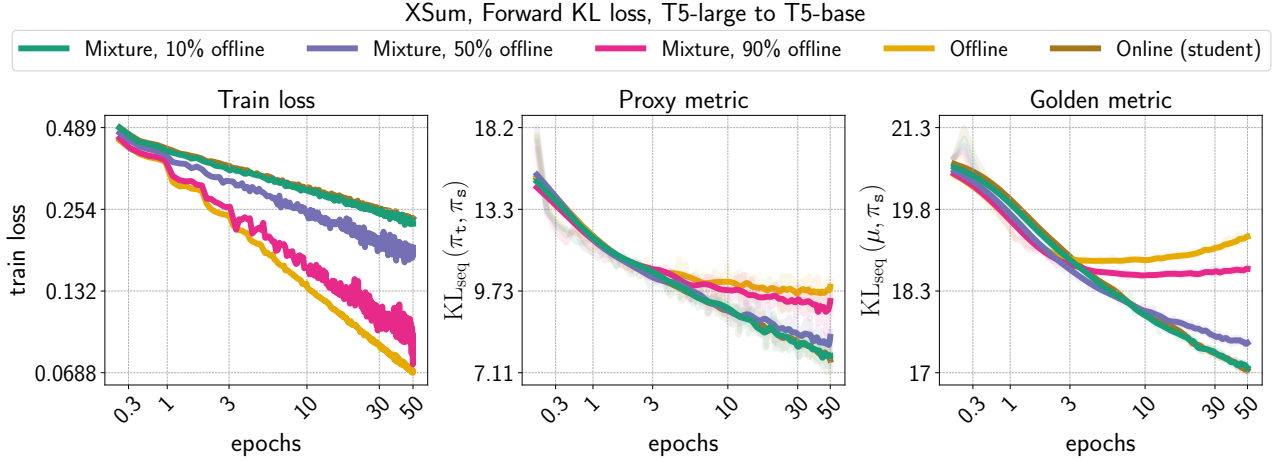
Figure 13. **Mixture of offline and online data.** This plot compares strategies for combining offline and online data during the distillation process. The results show that incorporating just 10% online student data significantly reduces the effect of teacher hacking, causing the golden metric to stabilize rather than increase. At the same time, the usage of at least 50% of the online generated data allows to avoid the effect of teacher hacking completely.

B. Additional Details

In this section, we provide an exact formulation for sequence-level divergences

$$\mathrm{KL}_{\mathrm{seq}}(\pi, \pi') \triangleq \mathbb{E}_{x \sim d(\cdot), y \sim p_\pi(\cdot|x)} \left[\sum_{i=1}^{|y|} \mathrm{KL}(\pi(\cdot|x, y_{:i}), \pi'(\cdot|x, y_{:i})) \right]. \tag{3}$$

In particular, to estimate this KL-divergence, we need to sample prompts x and generate responses using a language model π. Additionally, we introduce a sequence-based Jensen-Shannon divergence, defined as

$$\mathrm{JS}_{\mathrm{seq}}(\pi, \pi') \triangleq \mathbb{E}_{x \sim d(\cdot), y \sim p_\pi(\cdot|x), y' \sim p_{\pi'}(\cdot|x)} \left[\frac{1}{2} \sum_{i=1}^{|y|} \mathrm{KL}(\pi(\cdot|x, y_{:i}), m(\cdot|x, y_{:i})) + \frac{1}{2} \sum_{i=1}^{|y'|} \mathrm{KL}(\pi'(\cdot|x, y'_{:i}), m(\cdot|x, y'_{:i})) \right], \tag{4}$$

where $m(\omega|x, y) \triangleq 0.5 \cdot \pi(\omega|x, y) + 0.5 \cdot \pi'(\omega|x, y)$ is a mixture of two language models. To estimate this divergence, we need to use samples from both models π and π' and compute an average of two token-level KL-divergences. Notice that computation of a true Jensen-Shannon divergence between p_π and $p_{\pi'}$ is computationally infeasible since the log-probabilities of the mixture $0.5 \cdot p_\pi(y|x) + 0.5 \cdot p_{\pi'}(y|x)$ does not satisfy a chain rule in terms of π and π'.

C. Hyperparameters

In this section, we provide detailed information on the hyperparameters used for our experiments; see Table 2 and Table 3.

Table 2. Hyperparameter details for summarization & translation tasks.

Hyperparameter	Value	Hyperparameter	Value
Oracle Dataset Size	100,000	Oracle Dataset Size	50,000
Distillation Dataset Size	200,000	Distillation Dataset Size	450,000
Training Steps	390,625	Training Steps	703,125
Batch Size	32	Batch Size	32
Task	XSum	Task	WMT-14 en-de
Dropout	0.0	Dropout	0.0
Warmup Schedule	100 steps	Warmup Schedule	100 steps
Optimal Learning Rate (LR)	0.0003	Optimal Learning Rate (LR)	0.0003
Input Length (Tokenized)	1024	Input Length (Tokenized)	80
Output Length (Tokenized)	128	Output Length (Tokenized)	80
Softmax Temperature	1.0	Softmax Temperature	1.0

Table 3. Hyperparameter details for instruction following task.

Hyperparameter	Value
Oracle Dataset Size	100,000
Distillation Dataset Size	500,000
Training Steps	390,625
Batch Size	64
Task	Natural Instructions
Dropout	0.0
Warmup Schedule	100 steps
Optimal Learning Rate (LR)	0.0003
Input Length (Tokenized)	2048
Output Length (Tokenized)	256
Softmax Temperature	1.0