Revisiting Weak-to-Strong Generalization in Theory and Practice: Reverse KL vs. Forward KL

Anonymous ACL submission

Abstract

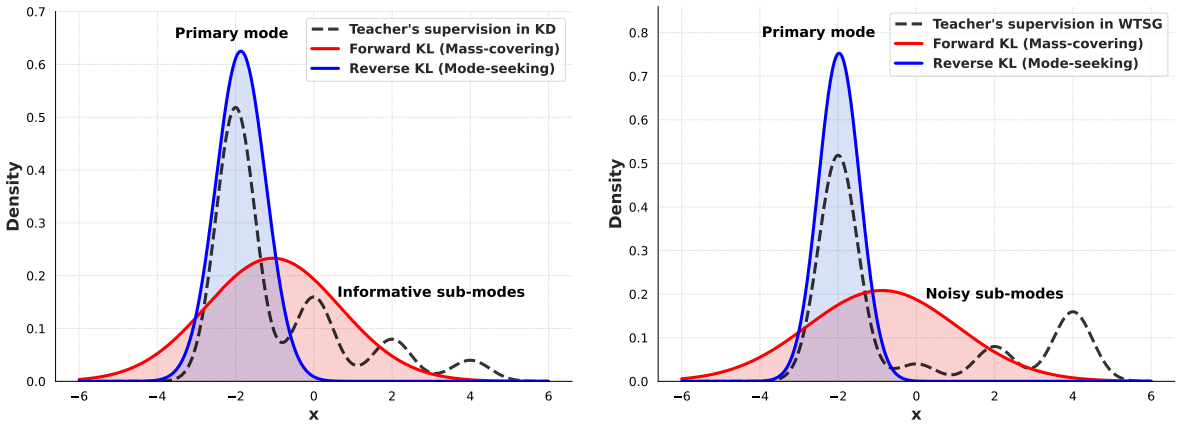
As large language models advance toward superhuman performance, ensuring their alignment with human values and abilities grows increasingly complex. Weak-to-strong generalization offers a promising approach by leveraging predictions from weaker models to guide stronger systems, but its effectiveness could be constrained by the inherent noise and inaccuracies in these weak predictions. To address this, we propose a theoretically grounded approach that replaces forward KL divergence—whose mass-covering behavior risks overfitting to imperfect weak signals—with reverse KL divergence. Reverse KL divergence's zero-forcing effect prioritizes high-confidence predictions, effectively mitigating the influence of unreliable weak supervision. Theoretically, we extend existing bounds and derive tighter lower bounds for both forward and reverse KL divergence, establishing that reverse KL achieves at least comparable guarantees to forward KL. Notably, when a sufficiently pre-trained strong model is fine-tuned on the last layer, reverse KL uniquely guarantees that it outperforms its weak supervisor by the magnitude of their disagreement—a guarantee that forward KL cannot provide. Empirically, we demonstrate that reverse KL and reverse cross-entropy enable strong models to successfully outperform those trained with forward KL and standard cross-entropy across most settings, highlighting the practical advantages of these reverse losses.

1 Introduction

Human supervision is indispensable to align Large Language Models (LLMs) with human values (Bai et al., 2022a; OpenAI, 2023a). However, as LLMs approach superhuman capabilities, their behaviors may exceed human ability to reliably manage (OpenAI, 2023b). To address this challenge, Weak-to-Strong Generalization (WTSG) (Burns et al., 2023) emerges as a promising approach, leveraging weaker models to guide and control more advanced systems, thereby bridging the gap between human oversight and superhuman AI capabilities.

In particular, WTSG demonstrates that strong pre-trained LLMs, when fine-tuned under weak model supervision, can achieve performance surpassing that of their weak supervisors. However, this approach is fundamentally constrained by the inherent imperfections of weak model supervision, which may introduce inaccuracies and noise (Burns et al., 2023). Blindly fitting the strong model to these imperfect signals can lead to a significant discrepancy between the ground truth and the model's predictions, ultimately undermining the effectiveness of WTSG (Yao et al., 2025). This raises a critical question: *How to effectively leverage weak supervision to guide strong models while mitigating the impact of noisy or inaccurate signals?*

To answer this question, we propose a theoretically principled approach, supported by fine-grained analysis and a simple yet effective solution. Our motivation stems from an insightful comparison with Knowledge Distillation (KD) (Hinton, 2015) in classification, where strong teachers provide informative soft labels to guide weak students. In KD, the forward KL divergence loss plays a crucial role as it encourages students to learn not only the target class probabilities but also the relative relationships among non-target classes encoded in the teacher's soft labels. For instance, in the image classification scenario, a strong teacher might assign higher probabilities to "tiger" than to "dog" when the input image is a "cat", reflecting the semantic similarity between cats and tigers in the feature space. However, this advantageous property of forward KL in KD becomes a limitation in the WTSG paradigm. The fundamental distinction lies in the quality of supervision: while strong teachers in KD provide reliable and informative soft labels, weak teachers in WTSG often generate noisy and potentially misleading signals for non-target classes (Burns et al., 2023). Thus, the mass-



(a) Knowledge Distillation (b) Weak-to-Strong Generalization

Figure 1: Illustration of the mass-covering behavior of forward KL divergence and the mode-seeking behavior of reverse KL divergence, highlighting their roles in KD and WTSG. A Gaussian mixture distribution, representing the teacher's supervision in KD and WTSG, is approximated by fitting a single Gaussian distribution using both forward and reverse KL divergence as loss functions.

covering nature of forward KL (Jerfel et al., 2021; Sun and van der Schaar, 2024), which forces the student to match the entire probability distribution of the teacher's predictions, becomes detrimental in WTSG as it may lead the strong model to overfit to the weak teacher's unreliable supervision. This observation motivates our investigation of reverse KL divergence as a more suitable alternative for WTSG. As shown in Figure 1, the key advantage of reverse KL lies in its mode-seeking behavior (Minka et al., 2005; Ji et al., 2024a), which enables the strong model to focus on the weak teacher's high-confidence predictions while being less sensitive to potentially noisy low-probability regions. This property aligns better with the WTSG setting, as it allows the strong model to extract reliable patterns from weak supervision without being overly constrained by its imperfections.

Building on the intuitive motivation above, we first conduct a theoretical analysis to compare forward losses and reverse losses in the context of WTSG. Inspired by the lower and upper bounds established for the strong model in WTSG (Yao et al., 2025), we extend these results and derive tighter lower bounds for both forward and reverse losses, demonstrating that reverse losses achieves at least equivalent theoretical guarantees to forward losses. Furthermore, we identify a unique advantage of reverse KL: when an adequately pre-trained strong model undergoes last-layer fine-tuning, reverse KL guarantees that the strong student will outperform its weak teacher by at least the magnitude of their disagreement. Notably, this performance guarantee fails to hold for forward KL without additional assumptions, underscoring the theoretical advantage of reverse losses. In our experiments, we empirically demonstrate that employing reverse KL divergence and reverse Cross-Entropy (CE) as loss functions enables the strong model to achieve superior performance compared to using forward KL divergence and standard CE. We also extend the analysis to an improved algorithm discussed in Burns et al. (2023), where the optimization objective incorporates an additional regularization term. It further demonstrates the practical advantages of reverse CE over standard CE in the context of WTSG.

2 Related Work

Weak-to-Strong Generalization. The weak-to-strong paradigm (Burns et al., 2023) emerges as a promising framework to address the challenges of AI alignment, particularly in the context of superalignment (OpenAI, 2023b)—where future AI systems may surpass human capabilities, rendering human supervision weak or insufficient. It leverages weaker models to guide stronger models, potentially unlocking their full capabilities while maintaining alignment with human values. It has been extensively studied through algorithms (Zhu et al., 2024; Agrawal et al., 2024; Sang et al., 2024; Guo and Yang, 2024), empirical analyses (Yang et al., 2024; Ye et al., 2024), and theoretical frameworks (Lang et al., 2024; Somerstep et al., 2024; Wu and Sahai, 2024; Charikar et al., 2024; Yao et al., 2025), these works primarily focus on WTSG with forward KL divergence and CE losses. How-

ever, to the best of our knowledge, the potential of reverse KL and reverse CE losses in classification under the WTSG framework remains unexplored.

Forward KL and Reverse KL. Forward KL and Reverse KL are employed in distinct applications, each offering unique advantages. *Forward KL* is widely utilized in standard classification tasks (Goodfellow, 2016), often appearing in the form of CE loss to align predicted and true label distributions. Its mass-covering behavior (Jerfel et al., 2021; Sun and van der Schaar, 2024) ensures that the model comprehensively captures all high-probability regions of the target distribution, making it particularly effective in knowledge distillation (Hinton, 2015) for classification tasks. In such tasks, the teacher model's soft labels provide informative guidance, enabling the student model to learn a representative distribution (Yang et al., 2025). In contrast, *reverse KL* is frequently adopted in variational inference (Kingma and Welling, 2014; Pinheiro Cinelli et al., 2021), where it exhibits zero-forcing behavior (Minka et al., 2005). By focusing on high-confidence predictions while disregarding low-probability regions, reverse KL prioritizes precision over diversity. In the context of WTSG, the choice of divergence is especially significant. Weak teachers in WTSG provide imperfect supervision signals (Burns et al., 2023; Yang et al., 2024; Yao et al., 2025), and using forward KL divergence as the loss function may lead to overfitting to these noisy or incomplete guidance. Reverse KL, on the other hand, allows the strong model to extract reliable patterns from weak supervision without being overly constrained by its imperfections. This property aligns well with the goal of WTSG, where the focus is on leveraging weak supervision while avoiding its pitfalls.

Furthermore, reverse KL divergence has recently gained increasing attention in related fields such as domain adaptation (Nguyen et al., 2022) and KL-regularized reinforcement learning (Rafailov et al., 2024; Wang et al., 2024; Ji et al., 2024b). These applications share a conceptual similarity with WTSG, as they all involve transferring knowledge across domains or models under imperfect or constrained conditions. Moreover, beyond classification tasks, reverse KL divergence has been increasingly utilized in generation tasks within knowledge distillation (Gu et al., 2024; Agarwal et al., 2024; Wu et al., 2025), owing to its mode-seeking properties. Given these developments, it is natural to investigate the role of reverse KL within the WTSG framework. To the best of our knowledge, no prior work has systematically explored this direction, leaving a significant gap in understanding its potential applications and implications.

3 Preliminaries

3.1 Classification

We consider k-classification tasks. Given the data domain $\mathcal{X} \subseteq \mathbb{R}^d$ and output domain $\mathcal{Y} \subseteq \mathbb{R}^k$, let the model space be $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$. Consider the model equipped with a softmax module, which ensures that its outputs form a valid probability distribution, i.e., $\forall y = (y_1, \cdots, y_k)^T \in \mathcal{Y}$, there holds $\sum_{i=1}^k y_i = 1$ and $0 < y_i \leq 1$. The forward and reverse KL divergence losses are defined below.

Definition 1 (KL divergence losses)**.** *Given the data distribution $\mathcal{P}$ and two models $g, h \in \mathcal{F}$, the **forward** KL divergence loss is defined as:*

$$\begin{aligned}\mathrm{KL}(g, h) &\triangleq \mathbb{E}_{x\sim\mathcal{P}}\left[\mathrm{D}_{\mathrm{KL}}(g(x)\|h(x))\right], \\ &= \mathbb{E}_{x\sim\mathcal{P}}\left[\sum_{i=1}^{k}[g(x)]_i \log\frac{[g(x)]_i}{[h(x)]_i}\right],\end{aligned}$$

*where $[g(x)]_i, [h(x)]_i$ represent the i-th elements of $g(x), h(x)$, respectively. Thus, the **reverse** KL divergence loss is* $\mathrm{KL}(h, g)$.

As illustrated in Figure 1, forward KL promotes full coverage of the target distribution, whereas reverse KL focuses on capturing the dominant mode. Additionally, the difference between KL divergence and CE is an entropy term:

Definition 2 (Cross-entropy losses)**.** *Given the data distribution $\mathcal{P}$ and two models $g, h \in \mathcal{F}$, define the **forward** cross-entropy divergence loss:*

$$\begin{aligned}\mathrm{CE}(g, h) &\triangleq -\mathbb{E}_{x\sim\mathcal{P}}\left[\sum_{i=1}^{k}[g(x)]_i \log[h(x)]_i\right] \\ &= \mathrm{KL}(g, h) + \mathbb{E}_{x\sim\mathcal{P}}\, H(g(x)),\end{aligned}$$

*where $H(\cdot)$ is the Shannon entropy. Thus, the **reverse** cross-entropy loss is* $\mathrm{CE}(h, g)$.

Consequently, note that when minimizing forward losses, the model g is fixed to provide supervision signals. Thus, minimizing forward KL divergence loss is equivalent to minimizing standard CE loss as $\mathbb{E}_{x\sim\mathcal{P}}\, H(g(x))$ is a constant.

3.2 Weak-to-Strong Generalization

Consider WTSG in the context of k-classification tasks. We focus on the fine-tuning phase after pre-training. The labeling function $F^\star$ maps data x to its label $F^\star(x)$. The strong model aims to learn $F_{sw} = f \circ h_s$, where h_s is a fixed strong model representation and $f \in \mathcal{F}_s$ is a task-specific function from a hypothesis class $\mathcal{F}_s$. In the convention setting of AI alignment (Ouyang et al., 2022), the model is fine-tuned through ground truth data:

$$f_s = \text{argmin}_{f \in \mathcal{F}_s} L(F^\star, f \circ h_s), \quad (1)$$

where the loss $L(\cdot, \cdot)$ can be $\text{KL}(\cdot, \cdot)$ or $\text{CE}(\cdot, \cdot)$. However, it is humans who provide weak supervision in the super-alignment scenario (OpenAI, 2023b). To explore this, the WTSG framework (Burns et al., 2023) leverages a weak model's predictions to supervise the strong model:

$$f_{sw} = \text{argmin}_{f \in \mathcal{F}_s} L(F_w, f \circ h_s), \quad (2)$$

where F_w is a given weak model, and $L(\cdot, \cdot)$ is originally the standard CE loss. If we employ reverse losses, the objective transforms into

$$f^r_{sw} = \text{argmin}_{f \in \mathcal{F}_s} L(f \circ h_s, F_w). \quad (3)$$

Regardless of the choice of loss function, the core objective is replacing ground truth data with weak supervision. Thus, while minimizing forward losses $L(F_w, F_{sw})$ or reverse losses $L(F_{sw}, F_w)$, we simultaneously strive to achieve an F_{sw} with a small generalization error $L(F^\star, F_{sw})$.

4 Theoretical Analysis: Justifying Reverse KL in WTSG

In Sections 4.1, we establish that both reverse and forward losses offer comparable generalization guarantees for the strong model, indicating that *reverse losses is at least as favorable as forward losses in terms of theoretical properties*. However, our analysis in Section 4.2 uncovers a key distinction: *with reverse KL divergence loss employed in WTSG, the strong model is theoretically guaranteed to outperform the weak model* by at least the magnitude of their disagreement under reasonable assumptions. Notably, this performance guarantee does not hold for forward KL, highlighting the theoretical advantage of reverse losses in WTSG.

4.1 Generalization Analysis of Both Losses

We establish that both reverse and forward losses yield comparable generalization guarantees by deriving upper and lower bounds for their respective generalization errors. We begin with a universal result for both forward and reverse losses.

Upper and lower bounds. We extend Yao et al. (2025) and establish bounds of strong model's performance. Unlike previous work that focuses only on forward KL loss, we comprehensively examine all four loss variants: forward KL, reverse KL, forward CE, and reverse CE.

Lemma 1 (Proved in Appendix A.1)**.** *Let* $L(\cdot, \cdot)$ *be* $\text{KL}(\cdot, \cdot)$ *or* $\text{CE}(\cdot, \cdot)$*. Given the data domain* $\mathcal{X}$*, output domain* $\mathcal{Y}$ *and models* $F_w, F^\star$ *defined above. For any strong model* F_{sw}*, there holds*

$$\begin{aligned} L(F^\star, F_w) - C_1 d(F_w, F_{sw}) \\ \leq L(F^\star, F_{sw}) \leq \\ L(F^\star, F_w) + C_1 d(F_w, F_{sw}), \end{aligned}$$

where C_1 *is a positive constant,* $d(F_w, F_{sw})$ *can be* $\sqrt{\text{KL}(F_w, F_{sw})}$ *or* $\sqrt{\text{KL}(F_{sw}, F_w)}$*, and* $L(F^\star, F_{sw})$ *and* $L(F^\star, F_w)$ *represent the error of strong model and weak model, respectively.*

Note that $d(F_w, F_{sw})$ captures the disagreement between the strong and weak models, which serves as the minimization objective in WTSG. Lemma 1 quantifies the difference between the weak and strong models' performance from two perspectives: a lower bound and an upper bound, which is similar to Yao et al. (2025). The **lower bound** indicates that strong model's performance cannot be arbitrarily improved using weak supervision. Improving the strong model depends critically on ensuring $L(F^\star, F_w)$ is small, underscoring *the importance of weak model's performance*. Also, whether we choose forward or reverse loss, the student-supervisor disagreement $d(F_w, F_{sw})$ is minimized. While reducing $L(F^\star, F_{sw})$ requires increasing $d(F_w, F_{sw})$, the lower bound also implies that strong model's performance gain may be inherently constrained by WTSG's own optimization objective (Yao et al., 2025). In other words, *achieving the minimal optimization objective limits the strong model's ability* to significantly outperform its weak supervisor. The **upper bound** ensures that strong model's error $L(F^\star, F_{sw})$ remains bounded and do not be arbitrarily large. It shows that a better weak model is also crucial to improve strong model's

performance. Building on these universal results for both forward and reverse losses, we further conduct a fine-grained analysis to investigate how to achieve tighter lower and upper bounds.

Tighter lower bound. Consider the lower bound in Lemma 1, we employ alternative proof techniques rooted in information-theoretic inequalities to derive a tighter lower bound.

Theorem 1 (Proved in Appendix A.2)**.** *Let* $L(\cdot,\cdot)$ *be* $\mathrm{KL}(\cdot,\cdot)$ *or* $\mathrm{CE}(\cdot,\cdot)$*. Given* $F_{sw}, F_w, F^\star$*, then*

$$L(F^\star, F_{sw}) \geq L\left(F^\star, F_w\right) - C_2 d(F_w, F_{sw}),$$

where C_2 *is a positive constant, and* $d(F_w, F_{sw})$ *can be* $\sqrt{\mathrm{KL}(F_w, F_{sw})}$ *or* $\sqrt{\mathrm{KL}(F_{sw}, F_w)}$*.*

Remark. *C_2 is generally smaller than C_1, leading to a tighter lower bound than Lemma 1.*

Similar to Lemma 1, it also highlights the importance of selecting a well-generalizing weak model and cautious optimization of the strong model to prevent overfitting to weak supervision. Note that Theorem 1 applies to both forward and reverse losses, which share the same theoretical properties.

Tighter upper bound. In Lemma 1, there is no theoretical guarantee that the strong model will necessarily surpass the performance of its weak supervisor in WTSG, such as $L(F^\star, F_{sw}) \leq L\left(F^\star, F_w\right)$. This raises the question of whether a tighter upper bound can be derived. Therefore, we first explore how to achieve this goal.

Proposition 1 (Proved in Appendix A.3)**.** *Let* $L(\cdot,\cdot)$ *be* $\mathrm{KL}(\cdot,\cdot)$ *or* $\mathrm{CE}(\cdot,\cdot)$*. Given* $F_{sw}, F_w, F^\star$*, then there holds*

$$L(F^\star, F_{sw}) = L(F^\star, F_w) - \underbrace{\left\langle F^\star, \log \frac{F_{sw}}{F_w} \right\rangle_E}_{R},$$

where the expectation inner product is defined as $\langle f, g\rangle_E \triangleq \mathbb{E}_{x\sim\mathcal{P}}[f(x)^T g(x)]$.

Remark. *It can also be extended to reverse KL and squared loss, as detailed in Appendix A.3.*

Therefore, $L(F^\star, F_{sw}) \leq L\left(F^\star, F_w\right)$ satisfies *if and only if* $R \geq 0$. To achieve it, we aim to establish a clear relationship between model capacity and model confidence across all data points and all k classes. Specifically, for any $x \in \mathcal{X}$ and $i \in \{1, \cdots, k\}$, a positive $[F^\star(x)]_i \log \frac{[F_{sw}(x)]_i}{[F_w(x)]_i}$ ensures a positive R. Therefore, we expect the model predictions to satisfy either of the two inequalities:

$$[F^\star(x)]_i \geq [F_{sw}(x)]_i \geq [F_w(x)]_i, \tag{4}$$

$$[F^\star(x)]_i \leq [F_{sw}(x)]_i \leq [F_w(x)]_i. \tag{5}$$

In other words, the predicted probabilities of F_{sw} reflect the true outcome better than F_w. Intuitively, because the weak model is pre-trained and fine-tuned on ground truth data, we can trust its decisions for major classes. As shown in Figure 1, reverse KL's mode-seeking behavior encourages the strong model to focus on the weak model's high-confidence predictions, while disregarding low-probability, potentially noisy regions. This behavior facilitates the fulfillment of Inequality (4)-(5). In contrast, forward KL, with its mass-covering nature, forces the strong model to match the entire probability distribution, including unreliable signals from the weak model's lower-probability classes, thereby hindering the fulfillment of the above inequalities. In the context of WTSG, where weak supervision is imperfect, reverse KL's focus on high-confidence decisions provides stronger guarantees for strong model's performance. In particular, the theoretical analysis in the following section further supports this, demonstrating that reverse KL can theoretically ensure superior performance for the strong model in certain settings.

4.2 Unique Advantage of Reverse Losses

To achieve a tighter upper bound, our theoretical analysis below yields an intriguing insight: *when using reverse KL in WTSG, an adequate pre-training and subsequent last layer fine-tuning guarantees that the strong student can outperform its weak teacher* by at least the magnitude of their disagreement (i.e., $R \geq 0$ in Proposition 1).

Theorem 2 (Proved in Appendix A.4)**.** *Consider WTSG using reverse KL divergence loss:*

$$f_{sw} = \mathit{argmin}_{f\in\mathcal{F}_s} \mathrm{KL}(f \circ h_s, F_w).$$

Denote $F_{sw} = f_{sw} \circ h_s$*. Assume that the function class* $\mathcal{F}_s$ *is a convex set and* $\exists f_s \in \mathcal{F}_s$ *such that* $f_s \circ h_s = F^\star$*. Then:*

$$\mathrm{KL}\left(F^\star, F_{sw}\right) \leq \mathrm{KL}\left(F^\star, F_w\right) - \mathrm{KL}\left(F_{sw}, F_w\right).$$

Remark. *Similar result can be naturally extended to reverse CE loss. Furthermore, note that our proof leverages Bregman divergence, a generalization of both squared loss and KL divergence. This approach not only broadens the applicability of our results but also demonstrates how our framework naturally recovers the regression analysis of Charikar et al. (2024). On the contrary,*

under this proof framework, employing forward KL or CE losses does not inherently offer such theoretical guarantees unless we introduce additional assumptions. The above extension and discussion are detailed in Appendix A.4.

The assumptions are consistent with previous theory (Charikar et al., 2024; Yao et al., 2025). Firstly, the convex set assumption includes the case that $\mathcal{F}_s$ is the class of all linear functions. This aligns with previous insights (Howard and Ruder, 2018; Kumar et al., 2022; Mao et al., 2023; Kirichenko et al., 2023) on fine-tuning the last linear layer. Secondly, we consider the case where $\exists f_s \in \mathcal{F}_s$ such that $f_s \circ h_s = F^\star$. It shows the remarkable capability of pre-training. It assumes that the representation h_s has already captured a wealth of information during pre-training, a phenomenon well-demonstrated by modern pre-trained LLMs (Touvron et al., 2023; OpenAI, 2023a).

Theorem 2 establishes that in WTSG, using the reverse KL divergence loss guarantees that the strong model, trained with weak supervision, surpasses the weak model by at least their disagreement, $\mathrm{KL}(F_{sw}, F_w)$. This upper bound is tighter than Lemma 1, as Lemma 1 does not ensure that the strong model surpasses the weak model. Theorem 2 highlights the superior theoretical benefits of reverse losses compared to forward losses.

Now we draw n i.i.d. samples to perform WTSG and relax the assumption, where for any $f_s \in \mathcal{F}_s$, $\exists f_s \circ h_s = F^\star$ may not be satisfied. The unique result for reverse KL below further emphasizes its advantageous theoretical properties in WTSG.

Theorem 3 (Proved in Appendix A.5)**.** *Given F_{sw} defined in Theorem 2. Assume that $\mathcal{F}_s$ is a convex set. Consider WTSG using reverse KL divergence loss with n i.i.d. samples:*

$$\hat{f}_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s} \widehat{\mathrm{KL}}(f \circ h_s, F_w),$$

where $\widehat{\mathrm{KL}}(\cdot,\cdot)$ is the empirical version of $\mathrm{KL}(\cdot,\cdot)$. Denote $\hat{F}_{sw} = \hat{f}_{sw} \circ h_s$ and strong ceiling model's generalization error $\varepsilon = \mathrm{KL}(F^\star, F_s)$. With probability at least $1-\delta$, there holds

$$\mathrm{KL}(F^\star, \hat{F}_{sw}) \le \mathrm{KL}(F^\star, F_w) - \mathrm{KL}(\hat{F}_{sw}, F_w) + \mathcal{O}(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$. The asymptotic notation is for $\varepsilon \to 0, n \to \infty$.

Compared to Theorem 2, this bound introduces two additional error terms: the first term $\mathcal{O}(\sqrt{\varepsilon})$ is small due to the capability of the strong ceiling model F_s. The remaining two error terms, which are of the order $\mathcal{O}\left(1/\sqrt{n}\right)$, stem from the strong model $\hat{F}_{sw}$ being trained on a finite weakly-labeled dataset. These terms also diminish asymptotically as the sample size n increases. Overall, by using a sufficiently large dataset and a strong model with enough capacity, we achieve a large n and a very small ε, rendering the remainders in Theorem 3 negligible and increasing the likelihood that the theoretical guarantee in Theorem 2 holds. Theorem 3 aligns with previous wisdom (Charikar et al., 2024; Yao et al., 2025). However, whereas their corresponding bounds are specifically designed for regression tasks, our result offers new insights into applying reverse KL loss in classification tasks.

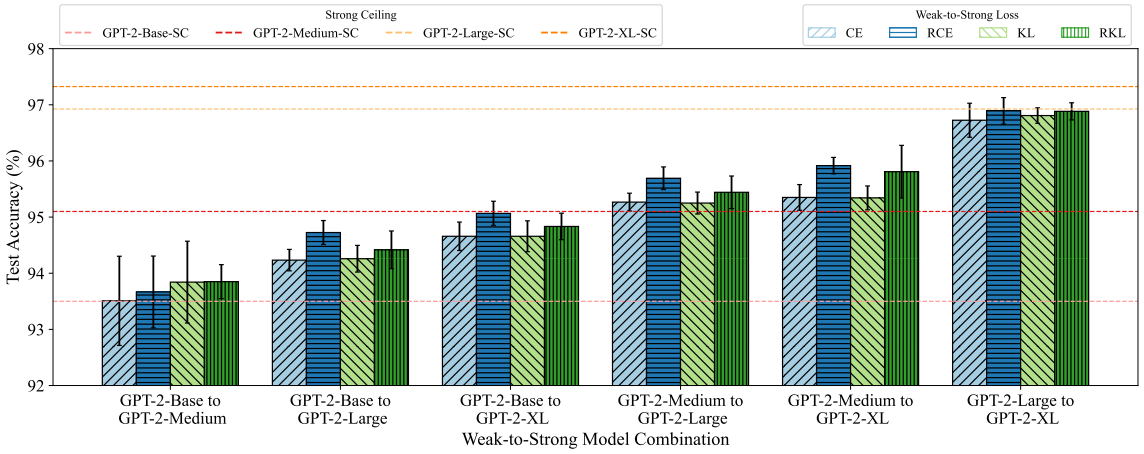
5 Empirical Validation

In this section, we empirically compare reverse KL, forward KL, reverse CE, and standard CE losses in the context of WTSG. Our experiments directly support the claim that reverse losses outperform forward losses in most experimental settings.

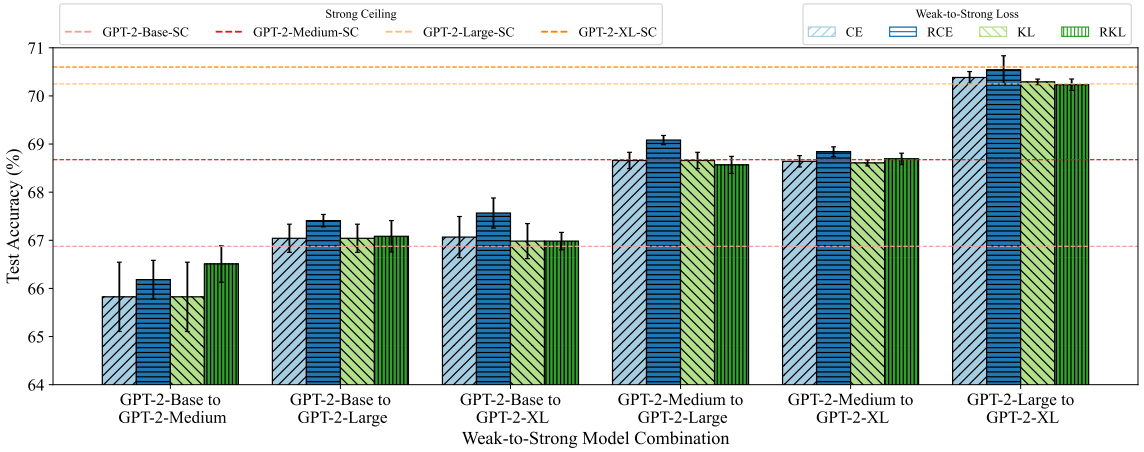
5.1 Experimental Settings

Datasets. We follow previous studies (Burns et al., 2023; Yang et al., 2024) to conduct experiments mainly in the reward modeling task in two settings: enabling a weak model to effectively guide a strong model in achieving either harmlessness or helpfulness. To achieve **harmlessness**, we follow (Yang et al., 2024) to leverage CAI-Harmless (Bai et al., 2022b), a widely adopted benchmark for single-turn harmless dialogue tasks. To achieve **helpfulness**, we utilize HH-RLHF (Bai et al., 2022a), a benchmark designed to guide models toward producing responses that are helpful, informative, and contextually relevant. We use a subset of single-turn helpful data of HH-RLHF.

Each dataset includes three subsets: **(1) Ground truth set**: 4K samples with ground truth labels, used to fine-tune the base models to create strong ceiling models. **(2) Weak supervision set**: 4K held-out samples, where the weak model generates predicted labels to guide the training of the strong model. **(3) Test set**: The extra 4K samples, reserved for evaluating the generalization performance of all strong ceiling and weak-to-strong models. Each sample is formatted as



(a) Results of GPT-2-series on CAI-Harmless



(b) Results of GPT-2-series on helpful set of HH-RLHF

Figure 2: Results of GPT-2-series. "SC" denotes the strong ceiling model, and "A to B" indicates the use of weak teacher "A" to supervise strong student "B". The terms CE, RCE, KL, and RKL refer to CE loss, reverse CE loss, forward KL divergence loss, and reverse KL divergence loss, respectively. Error bars represent the standard deviation across three runs of the experiment.

$\tilde{x} = (x; y_c, y_r)$, where x is the user input, y_c and y_r represent human-chosen and human-rejected responses separately.

Models. We conduct experiments on two types of model families: (1) GPT-2-series (Radford et al., 2019), including GPT-2-Base, GPT-2-Medium, GPT-2-Large, and GPT-2-XL; (2) Pythia-series (Biderman et al., 2023), specifically, Pythia-70M, Pythia-160M, Pythia-410M, and Pythia-1B. Each model is trained to generate a soft value between 0 to 1 for each sample:

$$F(\tilde{x}) = \text{Sigmoid}(F(y_c) - F(y_r)).$$

When implementing forward and reverse losses, the single predicted logit is transformed into a logits distribution represented as $(1 - F(\tilde{x}), F(\tilde{x}))$.

Training and Evaluation. The strong ceiling models are trained using the standard CE loss. We apply four loss functions in WTSG: forward KL, reverse KL, CE and reverse CE. To ensure the reliability and consistency of our results, each experiment is repeated across three random seeds. We set the training batch size to 16, learning rate to 10^{-5}, and `max_seq_len` to 512. Following the approach of Burns et al. (2023), we train each model for a single epoch to reduce overfitting. Finally, we report the average accuracy on the test set across the three random seeds for each model for comparison.

5.2 Main Results

The experimental results of the GPT-2 series on the CAI-Harmless and HH-RLHF datasets are presented in Figure 2. Due to space limitation, we put the detailed results for the Pythia series in Ap-

pendix B.1, but the similar trends can be observed.

We can draw several conclusions from the results in Figure 2: (1) The accuracy demonstrates a consistent upward trend from left to right. It indicates that the generalization capability of the strong model improves when a more capable weak model is employed as the supervisor. This finding is in line with Lemma 1 and aligns with prior research (Burns et al., 2023; Yao et al., 2025), which suggests that utilizing a higher-capacity weak model enhances the strong model's performance. Furthermore, with a fixed weak model, leveraging a stronger model also yield improved strong model's performance, consistent with established research (Burns et al., 2023; Yang et al., 2024). (2) We observe that **reverse KL and reverse CE losses enable strong models to outperform those trained with forward KL and CE losses across most experimental settings**. In particular, in all settings (12 out of 12), the use of reverse KL yields a stronger model compared to standard KL. Similarly, reverse CE outperforms or parallels forward CE in nearly all experimental settings (10 out of 12). These empirical results, supported by our theoretical framework, underscore the superiority of reverse losses over forward losses. (3) In the majority of settings (10 out of 12), the strong model surpasses or meets the performance of its weak supervisor when trained with reverse KL or reverse CE loss. This observation supports Theorem 2 and Theorem 3. However, the theoretical guarantees may not always hold in practice, particularly in scenarios involving extremely complex LLMs with limited training set in WTSG, where the underlying assumptions may be violated.

5.3 Ablation Study

We notice that Burns et al. (2023) investigates an improved strategy: incorporating an additional regularization term aimed at boosting the strong model's confidence in its predictions, while utilizing the standard CE loss as the primary objective. This naturally raises the question of whether combining reverse CE loss with such regularization can further improve the strong model's performance compared to standard CE loss with regularization. To explore this question, we conduct experiments using the GPT-2 series on the CAI-Harmless dataset as a representative case. Due to space limitation, we only put the results when GPT-2-Base acts as the weak model to supervise GPT-2-Medium, GPT-2-Large, and GPT-2-XL here in Figure 5, while put the full results in Appendix B.2.

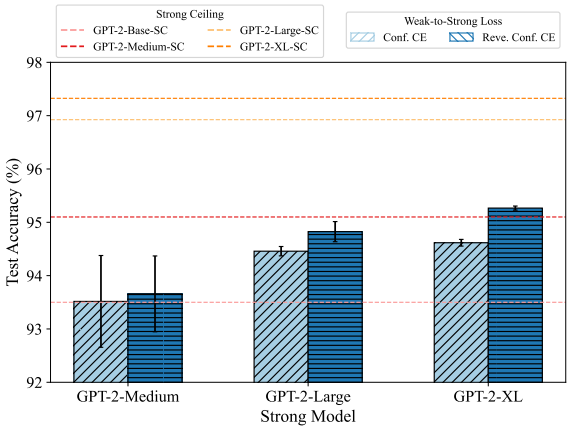


Figure 3: Results of GPT-2 series on CAI-Harmless. "SC" denotes the strong ceiling model. The terms "Conf. CE" and "Reve. Conf. CE" refer to the auxiliary confidence loss with vanilla CE loss and reverse CE loss, respectively. Error bars represent the standard deviation across three runs of the experiment.

First, by integrating the insights from Figure 2 and Figure 5, we can see that incorporating the confidence regularization leads to a modest improvement in the strong model's performance, aligning with the observations of Burns et al. (2023). Second, the strong model trained using reverse CE loss with regularization consistently surpasses its counterpart trained with standard CE loss. This result, together with our previous results in Section 5.2, underscores the clear advantage of reverse losses over forward losses in enhancing model performance in diverse settings.

6 Conclusion

In this work, we propose a theoretically principled approach by rethinking the loss function in WTSG. Unlike the mass-covering nature of forward KL, reverse KL exhibits a mode-seeking behavior that focuses on high-confidence predictions from the weak supervisor, thereby reducing the influence of noisy signals. Theoretically, we derive both upper and lower bounds for forward and reverse losses, demonstrating that reverse losses provide at least comparable guarantees to forward losses. Notably, when fine-tuning a pre-trained strong model on its last layer, reverse KL theoretically ensures that the strong model outperforms its weak supervisor by the magnitude of their disagreement—a guarantee forward KL cannot provide. Empirically, we show that reverse losses successfully outperform forward losses in most settings, highlighting the practical benefits of reverse KL and CE losses in WTSG.

Limitations

While our study provides theoretical insights and empirical validation for the advantages of reverse losses in WTSG, several limitations remain. First, our analysis mainly assumes relatively reliable weak supervision from pre-trained and fine-tuned models. However, real-world applications often involve noisy weak supervision, and reverse KL's mode-seeking nature may amplify extreme noise. Further research is needed to assess its suitability in such cases. Second, while the theoretical results in Section 4.1 provide broad insights, the assumptions in Section 4.2 may not hold for complex LLMs. This limitation is shared by all related work on theoretical understanding of WTSG, which relies on simplifying assumptions. Nonetheless, these foundations offer valuable guidance and a starting point for future research on advancing WTSG theory in LLMs. Third, our experiments are conducted on two well-known alignment-focused binary classification datasets with relatively smaller model sizes. While these results offer valuable insights, it remains an open question whether they can be generalized to more diverse datasets and larger-scale models. Exploring this aspect in future work will help further validate the broader applicability of our approach.

Ethics Statement

Our intention is to highlight the positive impact of reverse losses in improving weak-to-strong generalization, ensuring more robust and reliable model performance while minimizing the influence of potentially imperfect weak supervision. However, the potential amplification of biases from weak models remains a concern, particularly in sensitive applications where fairness is a critical issue. While reverse KL mitigates overfitting to unreliable supervision, its mode-seeking nature may amplify the biases present in the weak model's predictions. Additionally, stronger AI models trained using WTSG could be misused if deployed without appropriate safeguards, emphasizing the need for responsible development and oversight.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*.

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, John Langford, and Furong Huang. 2024. Ensemw2s: Can an ensemble of llms be leveraged to obtain a stronger llm? *arXiv preprint arXiv:2410.04571*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. 2024. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*.

Monroe D Donsker and SR Srinivasa Varadhan. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212.

Ian Goodfellow. 2016. *Deep learning*, volume 196. MIT press.

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

Yue Guo and Yi Yang. 2024. Improving weak-to-strong generalization with reliability-aware alignment. *arXiv preprint arXiv:2406.19032*.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A Heller, Yian Ma, and Michael I Jordan. 2021. Variational refinement for importance sampling using the forward kullback-leibler divergence. In *Uncertainty in Artificial Intelligence*, pages 1819–1829.

Haozhe Ji, Pei Ke, Hongning Wang, and Minlie Huang. 2024a. Language model decoding as direct metrics optimization. In *International Conference on Learning Representations*.

Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. 2024b. Towards efficient and exact optimization of language model alignment. In *International Conference on Machine Learning*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. 2024. Theoretical analysis of weak-to-strong generalization. *arXiv preprint arXiv:2405.16043*.

Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. 2023. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*.

Tom Minka et al. 2005. Divergence measures and message passing. Technical report, Microsoft Research.

A Tuan Nguyen, Toan Tran, Yarin Gal, Philip HS Torr, and Atılım Güneş Baydin. 2022. Kl guided domain adaptation. In *International Conference on Learning Representations*.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Introducing superalignment.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. 2021. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pages 111–149. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Jitao Sang, Yuhang Wang, Jing Zhang, Yanxu Zhu, Chao Kong, Junhong Ye, Shuyu Wei, and Jinlin Xiao. 2024. Improving weak-to-strong generalization with scalable oversight and ensemble learning. *arXiv preprint arXiv:2402.00667*.

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Ya'acov Ritov, Mikhail Yurochkin, and Yuekai Sun. 2024. A statistical framework for weak-to-strong generalization. *arXiv preprint arXiv:2405.16236*.

Hao Sun and Mihaela van der Schaar. 2024. Inverse-rlignment: Inverse reinforcement learning from demonstrations for llm alignment. *arXiv preprint arXiv:2405.15624*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *International Conference on Learning Representations*.

David X Wu and Anant Sahai. 2024. Provable weak-to-strong generalization via benign overfitting. *arXiv preprint arXiv:2410.04638*.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2025. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755.

Wenkai Yang, Yankai Lin, Jie Zhou, and Ji-Rong Wen. 2025. Distilling rule-based knowledge into large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 913–932.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Zhi Gong, Yankai Lin, and Ji-Rong Wen. 2024. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *arXiv preprint arXiv:2406.11431*.

Wei Yao, Wenkai Yang, Wang Ziqiao, Yankai Lin, and Yong Liu. 2025. Understanding the capabilities and limitations of weak-to-strong generalization. *arXiv preprint arXiv:2502.01458*.

Ruimeng Ye, Yang Xiao, and Bo Hui. 2024. Weak-to-strong generalization beyond accuracy: a pilot study in safety, toxicity, and legal reasoning. *arXiv preprint arXiv:2410.12621*.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. 2024. Weak-to-strong preference optimization: Stealing reward from weak aligned model. *arXiv preprint arXiv:2410.18640*.

Contents

Appendix

A Main Proof

A.1 Proof of Lemma 1

We first state some preliminaries for the proof.

Lemma 2 (Donsker and Varadhan's variational formula (Donsker and Varadhan, 1983))**.** *Let Q, P be probability measures on $\mathcal{X}$, for any bounded measurable function $f : \mathcal{X} \to \mathbb{R}$, we have*

$$\mathrm{D}_{\mathrm{KL}}(Q\|P) = \sup_f \mathbb{E}_{x\sim Q}[f(x)] - \log \mathbb{E}_{x\sim P}[\exp f(x)].$$

Lemma 3 (Hoeffding's lemma)**.** *Let $X \in \mathbb{R}$ such that $a \le X \le b$. Then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Definition 3 (Subgaussian random variable)**.** *A random variable $X \in \mathbb{R}$ is σ-subgaussian if for any ρ,*

$$\log \mathbb{E}\exp(\rho(X - \mathbb{E}X)) \le \rho^2\sigma^2/2.$$

We define the corresponding probability distributions for prediction of F_{sw} and F_w. $\forall x \in \mathcal{X}$, we know that $\sum_{j=1}^{k}[F_w(x)]_j = 1$. Therefore, given the class space $C_k = \{1, \cdots, k\}$, we define a probability distribution $\mathcal{P}_w(x)$ with the probability density function p_w, where $j \in C_k$ and

$$p_w(j) = [F_w(x)]_j. \tag{6}$$

Using this method, we also define the probability distribution $\mathcal{P}_{sw}(x)$ for $F_w(x)$.

Lemma 4 (Yao et al. (2025))**.** *Given the probability distributions $\mathcal{P}_w(x)$ and $\mathcal{P}_{sw}(x)$ above. For any $x \in \mathcal{X}$, $j \in C_k$, $g : C_k \to \mathbb{R}$ and assume that g is σ-subgaussian. Let $f = t \cdot g$ for any $t \in \mathbb{R}$, then*

$$\mathrm{D}_{\mathrm{KL}}\left(F_w(x)\|F_{sw}(x)\right) \ge \sup_t t\left(\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right) - t^2\sigma^2/2.$$

Now we start the proof.

Proof. By taking expectations of x on both sides of the inequality in Lemma 4, we obtain

$$\begin{aligned}\mathrm{KL}(F_w, F_{sw}) = \mathbb{E}_x\mathrm{D}_{\mathrm{KL}}\left(F_w(x)\|F_{sw}(x)\right)& \\ \ge \sup_t \underbrace{t\left(\mathbb{E}_x\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right) - t^2\sigma^2/2}_{\phi(t)}&.\end{aligned}$$

Note that $\phi(t)$ is a quadratic function of t. Therefore, by AM–GM inequality, we find the maximum of this quadratic function:

$$\phi(t) \le \frac{1}{2\sigma^2}\left(\mathbb{E}_x\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right)^2 = \sup_t \phi(t) \le \mathrm{KL}(F_w, F_{sw}).$$

Subsequently, there holds

$$\left|\mathbb{E}_x\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right| \le \sqrt{2\sigma^2\mathrm{KL}(F_w, F_{sw})}. \tag{7}$$

Likewise, according to Lemma 4, we have

$$\mathrm{D}_{\mathrm{KL}}\left(F_{sw}(x)\|F_w(x)\right) \ge \sup_t t\left(\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}\left[g\left(j'\right)\right] - \mathbb{E}_{j'\sim\mathcal{P}_w(x)}[g(j)]\right) - t^2\sigma^2/2. \tag{8}$$

We apply the same proof technique to (8) and obtain:

$$\left|\mathbb{E}_x\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right| \leq \sqrt{2\sigma^2\mathrm{KL}(F_{sw}, F_w)}. \tag{9}$$

Now we construct g to associate the above results with $L(F^\star, F_{sw})$ and $L\left(F^\star, F_w\right)$. Specifically, given a probability distribution $\mathcal{P}_g$ with the density function p_g, we define function $g : C_k \to (0, 1]$ associated with $\mathcal{P}_g$:

$$g(j) \triangleq \frac{[F^\star(x)]_j}{p_g(j)} \log \frac{[F^\star(x)]_j}{p_g(j)}, \quad \text{for } j \in C_k.$$

We have

$$\begin{aligned}
\mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_g}\left[g(j)\right] &= \mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_g}\left[\frac{[F^\star(x)]_j}{p_g(j)} \log \frac{[F^\star(x)]_j}{p_g(j)}\right] \\
&= \mathbb{E}_x\left[\sum_{j\in C_k} p_g(j) \cdot \frac{[F^\star(x)]_j}{p_g(j)} \cdot \log \frac{[F^\star(x)]_j}{p_g(j)}\right] \\
&= \mathbb{E}_x\left[\sum_{j\in C_k} [F^\star(x)]_j \cdot \log \frac{[F^\star(x)]_j}{p_g(j)}\right]
\end{aligned}$$

Recall the definition of $\mathcal{P}_{sw}$ and $\mathcal{P}_w$ in (6), we replace $\mathcal{P}_g$ with $\mathcal{P}_{sw}$ and $\mathcal{P}_w$ in the above equation:

$$\mathbb{E}_x\mathbb{E}_{j'\sim\mathcal{P}_{sw}}\left[g\left(j'\right)\right] = \mathbb{E}_x\left[\sum_{j=1}[F^\star(x)]_j \log \frac{[F^\star(x)]_j}{[F_{sw}(x)]_j}\right] = \mathrm{KL}(F^\star, F_{sw}),$$

$$\mathbb{E}_x\mathbb{E}_{j\sim\mathcal{P}_w}[g(j)] = \mathbb{E}_x\left[\sum_{j=1}[F^\star(x)]_j \log \frac{[F^\star(x)]_j}{[F_w(x)]_j}\right] = \mathrm{KL}(F^\star, F_w).$$

Substitute the above into (7):

$$|L(F^\star, F_{sw}) - L(F^\star, F_w)| \leq \sqrt{2\sigma^2\mathrm{KL}(F_w, F_{sw})}, \tag{10}$$

The above inequality is because whether L is KL or CE, we have

$$L(F^\star, F_{sw}) - L(F^\star, F_w) = \mathrm{KL}(F^\star, F_{sw}) - \mathrm{KL}(F^\star, F_w).$$

Likewise, we apply the same proof technique to (9) and obtain:

$$|L(F^\star, F_{sw}) - L(F^\star, F_w)| \leq \sqrt{2\sigma^2\mathrm{KL}(F_{sw}, F_w)}. \tag{11}$$

Finally, we obtain the subgaussian factor σ of function g by using the fact that g is bounded. Recall that the output domain $\mathcal{Y} \subseteq \mathbb{R}^k$, where $\forall y = (y_1, \cdots, y_k)^T \in \mathcal{Y}$, there holds $\sum_{i=1}^k y_i = 1$ and $0 < y_i \leq 1$. In other words, $\exists\gamma > 0$ such that $0 < \gamma \leq y_i \leq 1$. It means that $g(j) \in [-\frac{1}{\gamma}\log\frac{1}{\gamma}, \frac{1}{\gamma}\log\frac{1}{\gamma}]$. According to Lemma 3, $\forall\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(g(j)-\mathbb{E}[g(j)])}\right] \leq \exp\left(\frac{\lambda^2\left(\frac{1}{\gamma}\log\frac{1}{\gamma}\right)^2}{2}\right).$$

In other words, $g(j)$ is σ-subgaussian, where $\sigma = \frac{1}{\gamma}\log\frac{1}{\gamma}$. Substitute it into (10) and (11), we prove the final results:

$$|L(F^\star, F_{sw}) - L\left(F^\star, F_w\right)| \leq C_1\sqrt{\mathrm{KL}(F_w, F_{sw})},$$
$$|L(F^\star, F_{sw}) - L\left(F^\star, F_w\right)| \leq C_1\sqrt{\mathrm{KL}(F_{sw}, F_w)},$$

where the constant $C_1 = \frac{\sqrt{2}}{\gamma}\log\frac{1}{\gamma}$.

□

A.2 Proof of Theorem 1

Total variation distance is introduced for our proof.

Definition 4 (Total Variation Distance)**.** *Given two probability distributions P and Q, the Total Variation (TV) distance between P and Q is*

$$\mathrm{D}_{\mathrm{TV}}(P\|Q) = \frac{1}{2}\int_{x\in\mathcal{X}} |P(x) - Q(x)|\, dx.$$

Note that $\mathrm{D}_{\mathrm{TV}}(P\|Q) \in [0,1]$. Also, $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 0$ if and only if P and Q coincides, and $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 1$ if and only if P and Q are disjoint.

Proof. We have

$$\begin{aligned}
L(F^\star, F_w) &= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F^\star(x)]_i}{[F_w(x)]_i}\right] \\
&= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\left(\frac{[F^\star(x)]_i}{[F_{sw}(x)]_i}\cdot\frac{[F_{sw}(x)]_i}{[F_w(x)]_i}\right)\right] \\
&= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F^\star(x)]_i}{[F_{sw}(x)]_i}\right] + \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F_{sw}(x)]_i}{[F_w(x)]_i}\right] \\
&= L(F^\star, F_{sw}) + \left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E. \qquad (12)
\end{aligned}$$

Rearranging terms and we know that:

$$L(F^\star, F_{sw}) = L(F^\star, F_w) - \left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E. \qquad (13)$$

Recall that the output domain $\mathcal{Y} \subseteq \mathbb{R}^k$, where $\forall y = (y_1, \cdots, y_k)^T \in \mathcal{Y}$, there holds $\sum_{i=1}^k y_i = 1$ and $0 < y_i \leq 1$. In other words, $\exists\gamma > 0$ such that $0 < \gamma \leq y_i \leq 1$. Firstly, we know that $F^\star(x)$ is element-wise non-negative. Denote $\vec{1} = (1, 1, \cdots, 1)^T$. We know that there is a positive constant $\frac{1}{\gamma} \geq (\min_i [F_w(x)]_i)^{-1}$. We use element-wise addition, subtraction, multiplication, division and absolute value in the proof. Note that

$$\begin{aligned}
\left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E &\leq \left\langle F^\star, \frac{F_{sw}}{F_w} - \vec{1}\right\rangle_E && (\log x \leq x - 1) \\
&\leq \left\langle F^\star, \frac{1}{\gamma}\cdot F_w\left|\frac{F_{sw}}{F_w} - \vec{1}\right|\right\rangle_E && (\tfrac{1}{\gamma}\cdot F_w \geq \vec{1} \text{ (element-wise)}) \\
&= \frac{1}{\gamma}\cdot\langle F^\star, |F_{sw} - F_w|\rangle_E,
\end{aligned}$$

and

$$\begin{aligned}
\langle F^\star, |F_{sw} - F_w|\rangle_E &= \mathbb{E}_x\left[\left(F^\star(x)\right)^T\left(|F_{sw}(x) - F_w(x)|\right)\right] \\
&\leq \mathbb{E}_x\left[\|F^\star(x)\|_\infty \cdot \|F_{sw}(x) - F_w(x)\|_1\right] \\
&\qquad \text{(Holder's inequality for vector-valued functions)} \\
&\leq \mathbb{E}_x\left[\|F_{sw}(x) - F_w(x)\|_1\right] && ([F^\star(x)]_i \leq 1) \\
&= 2\mathbb{E}_x \mathrm{D}_{\mathrm{TV}}\left(F_w(x), F_{sw}(x)\right) && \text{(Definition of TV distance)} \\
&\leq 2\sqrt{\mathbb{E}_x \mathrm{D}^2_{\mathrm{TV}}\left(F_w(x), F_{sw}(x)\right)} && \text{(Jensen's inequality)} \\
&\leq 2\sqrt{\frac{1}{2}\mathbb{E}_x \mathrm{D}_{\mathrm{KL}}\left(F_w(x), F_{sw}(x)\right)} && \text{(Pinsker's inequality)}
\end{aligned}$$

$$= \sqrt{2\mathrm{KL}(F_w, F_{sw})}. \qquad \text{(Definition of KL}(\cdot,\cdot)\text{)}$$

Therefore,

$$\left\langle F^\star, \log \frac{F_{sw}}{F_w} \right\rangle_E \leq \frac{\sqrt{2}}{\gamma} \cdot \sqrt{\mathrm{KL}(F_w, F_{sw})}.$$

Since the TV distance is symmetric, we also have

$$\left\langle F^\star, \log \frac{F_{sw}}{F_w} \right\rangle_E \leq \frac{\sqrt{2}}{\gamma} \cdot \sqrt{\mathrm{KL}(F_{sw}, F_w)}.$$

Substitute them into Equation (13) and we can prove that:

$$L(F^\star, F_{sw}) \geq L\left(F^\star, F_w\right) - \underbrace{\frac{\sqrt{2}}{\gamma}}_{C_2} \sqrt{\mathrm{KL}(F_w, F_{sw})},$$

$$L(F^\star, F_{sw}) \geq L\left(F^\star, F_w\right) - \underbrace{\frac{\sqrt{2}}{\gamma}}_{C_2} \sqrt{\mathrm{KL}(F_{sw}, F_w)}.$$

The above inequalities also applies to $L(\cdot,\cdot) = \mathrm{CE}(\cdot,\cdot)$ because whether L is KL or CE, we have

$$L(F^\star, F_{sw}) - \mathrm{KL}(F^\star, F_{sw}) = L(F^\star, F_w) - \mathrm{KL}(F^\star, F_w).$$

□

Discussion of the constant. Recall that $C_1 = \frac{\sqrt{2}}{\gamma} \log \frac{1}{\gamma}$ and $C_2 = \frac{\sqrt{2}}{\gamma}$. In other words, $\gamma < \frac{1}{e}$ leads to $C_2 \leq C_1$. While γ is the minimal value of softmax output, it is generally a very small value ($\gamma = 10^{-3} \ll \frac{1}{e}$ in our experiments), i.e., $C_2 \leq C_1$. Therefore, the lower bound in Theorem 1 is tighter than that in Lemma 1.

Further Discussion. We show that adding an additional assumption leads to $L(F^\star, F_{sw}) \geq L\left(F^\star, F_w\right) - L(F_w, F_{sw})$. Particularly, if $L(F_w, F_{sw})$ can be improved to some extent, the constant C and square root in Theorem 1 can be eliminated, contributing to a more elegant version:

Corollary 1. *Let L to be* KL *or* CE. *Let $R \geq 0$ and consider the same constant C in Theorem 1. If $L(F_w, F_{sw}) \geq \sqrt{2}C$ is satisfied, then:*

$$L(F^\star, F_{sw}) \geq L\left(F^\star, F_w\right) - \mathrm{KL}(F_w, F_{sw}).$$

Corollary 1 removes the constant coefficient and square root from Theorem 1. Notably, if $R \geq 0$, the results above reinforce that the key bottleneck for performance improvement over F_w arises from the optimization objective's inherent nature (Yao et al., 2025): If $L(F_w, F_{sw})$ can be large, the performance improvement cannot exceed $L(F_w, F_{sw})$, which is exactly the minimum of Equation (2).

Proof. We adopt an alternative proof technique in the proof of Theorem 1:

$$\begin{aligned}
|\langle F^\star, |F_{sw} - F_w| \rangle_E| &\leq 2\mathbb{E}_x \mathrm{D_{TV}}\left(F_w(x), F_{sw}(x)\right) && \text{(The derivations in Appendix A.2)} \\
&\leq 2\mathbb{E}_x \sqrt{1 - \exp\left[-\mathrm{D_{KL}}\left(F_w(x), F_{sw}(x)\right)\right]} && \text{(Bretagnolle–Huber inequality)} \\
&\leq 2\sqrt{1 - \exp\left[-\mathbb{E}_x \mathrm{D_{KL}}\left(F_w(x), F_{sw}(x)\right)\right]} && \text{(Jensen's inequality)} \\
&= 2\sqrt{1 - \exp\left(-\mathrm{KL}(F_w, F_{sw})\right)}. && \text{(Definition of KL)}
\end{aligned}$$

Let $u(t) = e^{-t} + \frac{\gamma^2}{4}t^2 - 1, t \geq 0$. Taking the first-order and second-order derivative: $u'(t) = -e^{-t} + \frac{\gamma^2}{2}t$, and $u''(t) = e^{-t} + \frac{\gamma^2}{2} > 0$. While $u'(0) = -1 < 0$, $u'(\frac{2}{\gamma^2}) > 0$, we know that there only exists a $t_0 \in (0, \frac{2}{\gamma^2})$ such that $u'(t_0) = 0$. And $u(t)$ decreases at $[0, t_0]$, increases at $(t_0, +\infty)$ and $u(0) = 0$.

Denote $u(t^\star)=0$. Notice that $u(\frac{2}{\gamma})=e^{-\frac{2}{\gamma}}>0$, which means that $t^\star<\frac{2}{\gamma}$. In other words, $t>\frac{2}{\gamma}$ leads to $u(t)>0$, i.e., $\sqrt{1-e^{-t}}\leq\frac{\gamma}{2}t$.

Using the above results, if $\left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E \geq 0$ and $\mathrm{KL}(F_w,F_{sw})\geq\frac{2}{\gamma}$, then

$$\begin{aligned}
\left|\left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E\right| &\leq \left|\frac{1}{\gamma}\cdot\langle F^\star, |F_{sw}-F_w|\rangle_E\right| \qquad \text{(The derivations in Appendix A.2)}\\
&\leq \frac{2}{\gamma}\sqrt{1-\exp\left(-\mathrm{KL}(F_w,F_{sw})\right)}\\
&\leq \frac{2}{\gamma}\cdot\frac{\gamma}{2}\mathrm{KL}(F_w,F_{sw})\\
&= \mathrm{KL}(F_w,F_{sw}).
\end{aligned}$$

The proof is complete.

□

A.3 Proof of Proposition 1

Proof. We have

$$\begin{aligned}
L(F^\star,F_w) &= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F^\star(x)]_i}{[F_w(x)]_i}\right]\\
&= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\left(\frac{[F^\star(x)]_i}{[F_{sw}(x)]_i}\cdot\frac{[F_{sw}(x)]_i}{[F_w(x)]_i}\right)\right]\\
&= \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F^\star(x)]_i}{[F_{sw}(x)]_i}\right] + \mathbb{E}_x\left[\sum_{i=1}^k [F^\star(x)]_i \log\frac{[F_{sw}(x)]_i}{[F_w(x)]_i}\right]\\
&= L(F^\star,F_{sw}) + \left\langle F^\star, \log\frac{F_{sw}}{F_w}\right\rangle_E.
\end{aligned}$$

Rearranging terms and we can prove the result.

The above also applies to $L(\cdot,\cdot)=\mathrm{CE}(\cdot,\cdot)$ because whether L is KL or CE, we have

$$L(F^\star,F_{sw})-\mathrm{KL}(F^\star,F_{sw}) = L(F^\star,F_w)-\mathrm{KL}(F^\star,F_w).$$

□

Insights for reverse KL loss. Using similar decomposition technique, we obtain

$$L(F_w,F^\star) = L(F_{sw},F^\star) + \underbrace{\left\langle F_w-F_{sw}, \log\frac{F_w}{F^\star}\right\rangle_E - L(F_{sw},F_w)}_{R_1}.$$

Therefore, $L(F_{sw},F^\star)\leq L(F_w,F^\star)$ satisfies *if and only if* $R_1\geq 0$. While the teacher-student disagreement is minimized in WTSG, we expect a small value of $L(F_{sw},F_w)$. Therefore, we want to obtain a large $\left\langle F_w-F_{sw}, \log\frac{F_w}{F^\star}\right\rangle_E$. Intuitively, for any $x\in\mathcal{X}$ and $i\in\{1,\cdots,k\}$, we expect the model predictions to satisfy either of the two inequalities:

$$[F_w(x)]_i \geq \max([F_{sw}(x)]_i, [F^\star(x)]_i), \tag{14}$$
$$[F_w(x)]_i \leq \min([F_{sw}(x)]_i, [F^\star(x)]_i). \tag{15}$$

In other words, the predicted probabilities of F_{sw} reflect the true outcome better than F_w. The confidence level of F_{sw} should be better aligned with $F^\star$ than that of F_w.

Insights for squared loss. Charikar et al. (2024) consider the squared loss:

$$L(f,g) = \mathbb{E}_{x\sim\mathcal{P}}(f(x)-g(x))^2.$$

In this setting, $L(f,g) = L(g,f)$ and we have

$$\begin{aligned}
L\left(F_w, F^\star\right) &= \mathbb{E}_{x\sim\mathcal{P}}\left(F^\star(x) - F_w(x)\right)^2 \\
&= \mathbb{E}_{x\sim\mathcal{P}}\left(F^\star(x) - F_{sw}(x) + F_{sw}(x) - F_w(x)\right)^2 \\
&= \mathbb{E}_{x\sim\mathcal{P}}\left(F^\star(x) - F_{sw}(x)\right)^2 + \mathbb{E}_{x\sim\mathcal{P}}\left(F_{sw}(x) - F_w(x)\right)^2 \\
&\qquad + 2\cdot\mathbb{E}_{x\sim\mathcal{P}}\left[\left(F^\star(x) - F_{sw}(x)\right)\left(F_{sw}(x) - F_w(x)\right)\right] \\
&= L\left(F_{sw}, F^\star\right) + L\left(F_{sw}, F_w\right) + 2\cdot\mathbb{E}_{x\sim\mathcal{P}}\left[\left(F^\star(x) - F_{sw}(x)\right)\left(F_{sw}(x) - F_w(x)\right)\right].
\end{aligned}$$

If we define

$$\langle f, g\rangle_S = 2\cdot\mathbb{E}_{x\sim\mathcal{P}}\left[f(x)\cdot g(x)\right],$$

then we have

$$L\left(F_w, F^\star\right) = L\left(F_{sw}, F^\star\right) + L\left(F_{sw}, F_w\right) + \langle F^\star - F_{sw}, F_{sw} - F_w\rangle_S.$$

Rearranging terms and we have

$$L\left(F_{sw}, F^\star\right) = L\left(F_w, F^\star\right) - L\left(F_{sw}, F_w\right) - \langle F^\star - F_{sw}, F_{sw} - F_w\rangle_S. \tag{16}$$

Therefore, $\langle F^\star - F_{sw}, F_{sw} - F_w\rangle_S > 0$ is the sufficient and necessary condition for the inequality

$$L(F_{sw}, F^\star) \le L(F_w, F^\star) - L(F_w, F_{sw}),$$

when L is the squared loss. Therefore, we should make the confidence level of F_{sw} better aligned with $F^\star$. Despite the difficulty to attain this objective, Charikar et al. (2024) demonstrate that, within an elegant proof framework using convexity assumption, this condition is guaranteed to hold.

A.4 Proof of Theorem 2

Proof sketch. We define $\mathrm{KL}(\cdot,\cdot)$ in a Bregman-divergence manner. To derive the desired properties, we construct a convex combination of the form $F_{sw}(x) + t(F^\star(x) - F_{sw}(x))$, where $t \to 0^+$. By analyzing this construction, we show that the sum of the first-order term $\mathcal{O}(t)$ and the second-order term $\mathcal{O}(t^2)$ is non-negative. This implies that the first-order term itself must also be non-negative. Leveraging this principle and the associated derivations, we establish the proof of our results.

Our proof technique is general and unifying, covering *both squared loss and KL divergence loss*. While Theorem 1-2 from Charikar et al. (2024) focus exclusively on squared loss in regression, and Theorem 3-4 from Yao et al. (2025) restrict the analysis to KL divergence-like loss in regression, our work extends the scope to classification problems, encompassing both squared loss and KL divergence loss in a single framework. This broader applicability highlights the versatility of our proof and its potential to bridge gaps between regression and classification settings.

We first restate a lemma for our proof. Let the strong model learns from $\mathcal{F}_s : \mathbb{R}^{d_s} \to \mathbb{R}$ (which is a convex set) of fine-tuning tasks. Recall that we denote the strong model representation map by $h_s : \mathbb{R}^d \to \mathbb{R}^{d_s}$. Let $V_s = \{f \circ h_s : f \in \mathcal{F}_s\}$ be the set of all tasks in $\mathcal{F}_s$ composed with the strong model representation. Then V_s is also a convex set.

Lemma 5 (Charikar et al. (2024))**.** *V_s is a convex set.*

Proof. Fix $f, g \in \mathcal{F}_s$, and consider $f \circ h_s, g \circ h_s \in V_s$. Fix any $\lambda \in [0,1]$. Since $\mathcal{F}_s$ is the linear function class so that it is a convex set, there exists $p \in \mathcal{F}_s$ such that for all $y \in \mathbb{R}^{d_s}$, $p(y) = \lambda f(y) + (1-\lambda)g(y)$. Now, fix any $x \in \mathbb{R}^d$. Then, we have that

$$\lambda(f\circ h_s)(x) + (1-\lambda)(g\circ h_s)(x) = \lambda f(h_s(x)) + (1-\lambda)g(h_s(x)) = p(h_s(x)) = (p\circ h_s)(x),$$

and hence $\lambda(f\circ h_s) + (1-\lambda)(g\circ h_s) = p\circ h_s \in V_s$, which means that V_s is also a convex set. □

We then present our theoretical results.

Proof. Fix $f, g \in \mathcal{F}_s$, and consider $f \circ h_s, g \circ h_s \in V_s$. Fix any $\lambda \in [0,1]$. Since $\mathcal{F}_s$ is the linear function class so that it is a convex set, there exists $p \in \mathcal{F}_s$ such that for all $y \in \mathbb{R}^{d_s}$, $p(y) = \lambda f(y) + (1-\lambda)g(y)$. Now, fix any $x \in \mathbb{R}^d$. Then, we have that

$$\lambda(f \circ h_s)(x) + (1-\lambda)(g \circ h_s)(x) = \lambda f(h_s(x)) + (1-\lambda)g(h_s(x)) = p(h_s(x)) = (p \circ h_s)(x),$$

and hence $\lambda(f \circ h_s) + (1-\lambda)(g \circ h_s) = p \circ h_s \in V_s$. □

Motivated by the definition of Bregman divergence, we consider L as:

$$L(F_1, F_2) = \mathbb{E}_x \left[\phi(F_1(x)) - \phi(F_2(x)) - \langle \nabla\phi(F_2(x)), F_1(x) - F_2(x)\rangle\right], \tag{17}$$

where $F_1, F_2 \in \mathcal{F}$, and $\phi : \mathbb{R}^k \to \mathbb{R}$ is a strictly convex and differentiable function. Note that squared loss and KL divergence loss are special cases of the definition of L above:

$$\textbf{Squared loss:} \quad L(F_1, F_2) = \mathbb{E}_x \|F_1(x) - F_2(x)\|_2^2, \quad \phi(x) = x^T x,$$

$$\textbf{KL divergence loss:} \quad L(F_1, F_2) = \mathbb{E}_x \sum_{i=1}^{k} [F_1(x)]_i \log \frac{[F_1(x)]_i}{[F_2(x)]_i}, \quad \phi(x) = \sum_{i=1}^{k} x_i \log x_i.$$

Now we start our proof of Theorem 2.

Proof. We observe that

$$\begin{aligned}
L(g, F_w) &= \mathbb{E}_x \left[\phi(g) - \phi(F_w) - \langle \nabla\phi(F_w), g - F_w\rangle\right], \\
L(g, F_{sw}) &= \mathbb{E}_x \left[\phi(g) - \phi(F_{sw}) - \langle \nabla\phi(F_{sw}), g - F_{sw}\rangle\right], \\
L(F_{sw}, F_w) &= \mathbb{E}_x \left[\phi(F_{sw}) - \phi(F_w) - \langle \nabla\phi(F_w), F_{sw} - F_w\rangle\right],
\end{aligned}$$

which means that

$$L(g, F_w) = L(g, F_{sw}) + L(F_{sw}, F_w) + \underbrace{\mathbb{E}_x \left\langle g(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F_w(x))\right\rangle}_{R_1}. \tag{18}$$

Now our goal is to prove that $R_1 \geq 0$. We use reverse KL as the loss function in WTSG: $f_{sw} = \operatorname{argmin}_{f \in \mathcal{F}} L(f \circ h_s, F_w)$. In other words, F_{sw} is the *projection* of F_w onto the convex set V_s, i.e., $L(g, F_w) \geq L(F_{sw}, F_w)$. Substitute it into Equation (18) and we have

$$R_1 + L(g, F_{sw}) \geq 0. \tag{19}$$

Case 1: squared loss. Let $g = F_{sw} + t(F^\star - F_{sw})$, $t \in (0,1)$, $t \to 0^+$. Consider $\phi(x) = x^T x$, so $\nabla\phi(x) = 2x$. There holds

$$\begin{aligned}
R_1 &= 2t \cdot \mathbb{E}_x \left\langle F_{sw}(x) - F_w(x), F^\star(x) - F_{sw}(x)\right\rangle = \mathcal{O}(t), \\
L(g, F_{sw}) &= t^2 \mathbb{E}_x \|F^\star(x) - F_{sw}(x)\|_2^2 = \mathcal{O}(t^2).
\end{aligned}$$

Recall Equation (18) that $\mathcal{O}(t) + \mathcal{O}(t^2) \geq 0$, which means that $\mathcal{O}(t) \geq 0$. Therefore, there holds $R_1 \geq 0$, which means

$$\mathbb{E}_x \left\langle F^\star(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F_w(x))\right\rangle \geq 0.$$

Let $g = F^\star$ in Equation (18) and we can prove the result $L(F^\star, F_{sw}) \leq L(F^\star, F_w) - L(F_{sw}, F_w)$. While our proof is different from Charikar et al. (2024), we obtain the same conclusion for squared loss.

Case 2: reverse KL divergence. We consider $L(\cdot,\cdot)=\mathrm{KL}(\cdot,\cdot)$. Let $g=F_{sw}+t(F^\star-F_{sw}), t\in(0,1)$, $t\to 0^+$. Consider $\phi(x)=\sum_{i=1}^k x_i\log x_i$, so $\nabla\phi(x)=[\log x_1+1,\cdots,\log x_k+1]^T$. Firstly,

$$R_1=t\cdot\mathbb{E}_x(F^\star(x)-F_w(x))^T\begin{bmatrix}\log\frac{[F_{sw}(x)]_1}{[F_w(x)]_1}\\ \vdots\\ \log\frac{[F_{sw}(x)]_k}{[F_w(x)]_k}\end{bmatrix}=\mathcal{O}(t).$$

Secondly,

$$\begin{aligned}L(g,F_{sw})&=\mathbb{E}_x\sum_{i=1}^k[g(x)]_i\log\frac{[g(x)]_i}{[F_{sw}(x)]_i}\\&=\mathbb{E}_x\sum_{i=1}^k[F_{sw}(x)+t(F^\star(x)-F_{sw}(x))]_i\log\left(1+t\cdot\frac{[F^\star(x)-F_{sw}(x)]_i}{[F_{sw}(x)]_i}\right)\\&=\mathbb{E}_x\sum_{i=1}^k[F_{sw}(x)+t(F^\star(x)-F_{sw}(x))]_i\left(t\cdot\frac{[F^\star(x)-F_{sw}(x)]_i}{[F_{sw}(x)]_i}+\mathcal{O}(t^2)\right)\\&\qquad\qquad\text{(Taylor expansion)}\\&=\mathbb{E}_x\sum_{i=1}^k[F_{sw}(x)]_i\left(t\cdot\frac{[F^\star(x)-F_{sw}(x)]_i}{[F_{sw}(x)]_i}+\mathcal{O}(t^2)\right)+\mathcal{O}(t^2)\\&=t\cdot\mathbb{E}_x\sum_{i=1}^k[F^\star(x)-F_{sw}(x)]_i+\mathcal{O}(t^2)\\&=\mathcal{O}(t^2),\end{aligned}$$

where the last equation is because $\mathbb{E}_x\sum_{i=1}^k[F^\star(x)]_i=\mathbb{E}_x\sum_{i=1}^k[F_{sw}(x)]_i=1$. Therefore,

$$\underbrace{R_1}_{\mathcal{O}(t)}+\underbrace{L(g,F_{sw})}_{\mathcal{O}(t^2)}\geq 0,$$

which means $R_1\geq 0$, i.e.,

$$\mathbb{E}_x\left\langle F^\star(x)-F_{sw}(x),\nabla\phi(F_{sw}(x))-\nabla\phi(F_w(x))\right\rangle\geq 0.$$

Let $g=F^\star$ in Equation (18) and we can prove the result $L(F^\star,F_{sw})\leq L(F^\star,F_w)-L(F_{sw},F_w)$. □

Discussion of forward KL divergence. It is natural to ask, whether can the above proof technique be extended to forward KL? Our answer is that, we may need an additional assumption. In our proof, since reverse KL yields *a linear term*, the proof can be carried through. However, forward KL introduces *a logarithmic term*. While the Taylor expansions of the log function and a linear term differ only by a remainder term, proving the result requires assuming this remainder is non-negative, and that is why we need an additional assumption like Theorem 3 in (Yao et al., 2025). Here are the detailed explanations.

Note that

$$L(F_w,g)=L(F_{sw},g)+L(F_w,F_{sw})+\underbrace{\mathbb{E}_x\left\langle F_w(x)-F_{sw}(x),\nabla\phi(F_{sw}(x))-\nabla\phi(g(x))\right\rangle}_{R_2}.\tag{20}$$

Our goal is to prove that $R_2\geq 0$. Now we use forward KL as the loss function in WTSG: $f_{sw}=\operatorname{argmin}_{f\in\mathcal{F}}\ L(F_w,f\circ h_s)$. In other words, F_{sw} is the *projection* of F_w onto the convex set V_s, i.e., $L(F_w,g)\geq L(F_w,F_{sw})$. Substitute it into Equation (20) and we have

$$R_2+L(F_{sw},g)\geq 0.\tag{21}$$

Again, let $g = F_{sw} + t(F^\star - F_{sw})$, $t \in (0,1)$, $t \to 0^+$. Consider $\phi(x) = \sum_{i=1}^k x_i \log x_i$, so $\nabla\phi(x) = [\log x_1 + 1, \cdots, \log x_k + 1]^T$. Using a similar proof technique, we can obtain $R_2 = \mathcal{O}(t)$ and $L(F_{sw}, g) = \mathcal{O}(t^2)$. Therefore, we know that $R_2 \geq 0$, i.e.,

$$R_2 = \mathbb{E}_x \left\langle F_w(x) - F_{sw}(x), \underbrace{\nabla\phi(F_{sw}(x)) - \nabla\phi(F_{sw} + t(F^\star - F_{sw})(x))}_{\neq \nabla\phi(F_{sw}(x)) - \nabla\phi(F^\star(x))} \right\rangle \geq 0.$$

Consequently, even if we select $g = F^\star$ in Equation (20) and obtain

$$L(F_w, g) = L(F_{sw}, g) + L(F_w, F_{sw}) + \underbrace{\mathbb{E}_x \left\langle F_w(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F^\star(x)) \right\rangle}_{R_3 \neq R_2}.$$

Since we do not know whether $R_3 \geq 0$ is satisfied, we cannot directly prove the desired result. Since the difference between R_2 and R_3 can be quantified using exhaustive Taylor expansion, the nature of proof is similar to the regression analysis of WTSG (Proof of Theorem 3 from Yao et al. (2025), which introduces an additional assumption for the remainder of Taylor expansion). However, we do not know whether the remainder is larger than zero. In other words, to prove similar results for forward KL, we may introduce other assumptions like Theorem 3 in Yao et al. (2025). In contrast, the success of reverse KL and squared loss is because $R_3 = t \cdot R_2$. In the proof for these reverse losses, if $R_2 \geq 0$, then there also holds $R_3 \geq 0$.

Extension to reverse cross entropy loss. To extend the proof to reverse cross entropy, consider the following theoretical result.

Corollary 2. *Consider WTSG using reverse cross entropy loss:*

$$f_{sw} = \textit{argmin}_{f \in \mathcal{F}_s} \mathrm{CE}(f \circ h_s, f_w \circ h_w).$$

Assume that the function class $\mathcal{F}_s$ is a convex set and $\exists f_s \in \mathcal{F}_s$ such that $F_s = F^\star$. Then:

$$\mathrm{CE}(F^\star, F_{sw}) \leq \frac{1}{2}\left(\mathrm{CE}(F^\star, F_w) - \mathrm{KL}(F_{sw}, F_w)\right) + \log k.$$

If we consider binary classification (such as two famous datasets in AI safety: HH-RLHF (Bai et al., 2022a) and CAI-Harmless (Bai et al., 2022b)), then $k = 2$, making $\log k$ negligible due to the nature of KL divergence $\mathrm{KL}(\cdot,\cdot) \in [0, +\infty)$ and cross-entropy $\mathrm{CE}(\cdot,\cdot) \in [0, +\infty)$. It shows that if we use reverse cross-entropy loss in WTSG, the strong model's performance is also probably better than weak model's performance, which is also validated in our experiments.

Remark. *The proof also demonstrates that*

$$\mathrm{CE}(F^\star, F_{sw}) \leq \mathrm{CE}(F^\star, F_w) - \mathrm{KL}(F_{sw}, F_w) - \epsilon,$$

where $\epsilon = \mathrm{CE}(F^\star, F_{sw}) - \log k$. Due to the same reason, we expect $\epsilon \geq 0$, which comes to the same conclusion.

Proof. Rewrite Equation (18) and we have

$$\begin{aligned}\mathrm{CE}(g, F_w) = \mathrm{CE}(g, F_{sw}) + \mathrm{CE}(F_{sw}, F_w) \\ + \underbrace{\mathbb{E}_x\left(-H(F_{sw}(x)) + \langle g(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F_w(x))\rangle\right)}_{R_1'}. \quad (22)\end{aligned}$$

If we use reverse cross-entropy as the loss function in WTSG: $f_{sw} = \mathrm{argmin}_{f \in \mathcal{F}} \mathrm{CE}(f \circ h_s, F_w)$. In other words, $\mathrm{CE}(g, F_w) \geq \mathrm{CE}(F_{sw}, F_w)$. Let $g = F_{sw} + t(F^\star - F_{sw})$, $t \in (0,1)$, $t \to 0^+$. Substitute it into Equation (18) and we have

$$R_1' + \mathrm{CE}(g, F_{sw}) \geq 0,$$

$$\Rightarrow \underbrace{R_1}_{\mathcal{O}(t)} + \underbrace{L(g, F_{sw})}_{\mathcal{O}(t^2)} + \mathbb{E}_x \left(H(g(x)) - H(F_{sw}(x)) \right) \geq 0. \tag{23}$$

Note that

$$\begin{aligned}
&\mathbb{E}_x \left(H(g(x)) - H(F_{sw}(x)) \right) \\
=&\mathbb{E}_x \sum_{i=1}^{k} [g(x)]_i \log [g(x)]_i - [F_{sw}(x)]_i \log [F_{sw}(x)]_i \\
=&\mathbb{E}_x \sum_{i=1}^{k} [F_{sw}(x)]_i \log [g(x)]_i + t[F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i - [F_{sw}(x)]_i \log [F_{sw}(x)]_i \\
=&\mathbb{E}_x \sum_{i=1}^{k} [F_{sw}(x)]_i \log \frac{[g(x)]_i}{[F_{sw}(x)]_i} + t[F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i \\
=&\mathbb{E}_x \sum_{i=1}^{k} [F_{sw}(x)]_i \log \left(1 + t \cdot \frac{[F^\star(x) - F_{sw}(x)]_i}{[F_{sw}(x)]_i} \right) + t[F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i \\
=&\mathbb{E}_x \sum_{i=1}^{k} [F_{sw}(x)]_i \left(t \cdot \frac{[F^\star(x) - F_{sw}(x)]_i}{[F_{sw}(x)]_i} + \mathcal{O}(t^2) \right) + t[F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i \\
=&\mathbb{E}_x \sum_{i=1}^{k} t \cdot [F^\star(x) - F_{sw}(x)]_i + \mathcal{O}(t^2) + t[F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i \\
=&\mathcal{O}(t^2) + t \cdot \mathbb{E}_x \sum_{i=1}^{k} [F^\star(x) - F_{sw}(x)]_i \log [g(x)]_i && (\mathbb{E}_x \textstyle\sum_{i=1}^{k} [F^\star(x) - F_{sw}(x)]_i = 0) \\
=&\mathcal{O}(t^2) + t \cdot [\mathbb{E}_x H(F_{sw}(x)) - \mathrm{CE}(F^\star, F_{sw})], && \text{(Definition of entropy and cross entropy)}
\end{aligned}$$

where the last inequality is because as $t \to 0^+$, $g \to F_{sw}$. Consequently, recall Equation (23), we know that the sum of first-order terms $\mathcal{O}(t)$ is non-negative, i.e.,

$$t \cdot [\mathbb{E}_x H(F_{sw}(x)) - \mathrm{CE}(F^\star, F_{sw})] + R_1 \geq 0,$$

which means that

$$\mathbb{E}_x H(F_{sw}(x)) - \mathrm{CE}(F^\star, F_{sw}) + \mathbb{E}_x \left\langle F^\star(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F_w(x)) \right\rangle \geq 0.$$

Let $g = F^\star$ in Equation (22) and we obtain

$$\begin{aligned}
&\mathrm{CE}(F^\star, F_w) = \mathrm{CE}(F^\star, F_{sw}) + \mathrm{CE}(F_{sw}, F_w) - \mathbb{E}_x H(F_{sw}(x)) \\
&\qquad\qquad + \mathbb{E}_x \left\langle F^\star(x) - F_{sw}(x), \nabla\phi(F_{sw}(x)) - \nabla\phi(F_w(x)) \right\rangle \\
\Rightarrow&\mathrm{CE}(F^\star, F_w) \geq \mathrm{CE}(F^\star, F_{sw}) + \mathrm{CE}(F_{sw}, F_w) - \mathbb{E}_x H(F_{sw}(x)) \\
&\qquad\qquad + \mathrm{CE}(F^\star, F_{sw}) - \mathbb{E}_x H(F_{sw}(x)) \\
\Rightarrow&\mathrm{CE}(F^\star, F_w) \geq \mathrm{CE}(F^\star, F_{sw}) + \mathrm{KL}(F_{sw}, F_w) + \mathrm{CE}(F^\star, F_{sw}) - \mathbb{E}_x H(F_{sw}(x)) \\
\Rightarrow&\mathrm{CE}(F^\star, F_w) \geq \mathrm{CE}(F^\star, F_{sw}) + \mathrm{KL}(F_{sw}, F_w) + \mathrm{CE}(F^\star, F_{sw}) - \log k \qquad (H(F_{sw}(x)) \leq \log k)
\end{aligned}$$

Therefore, we prove the result

$$\mathrm{CE}(F^\star, F_{sw}) \leq \frac{1}{2} \left(\mathrm{CE}(F^\star, F_w) - \mathrm{KL}(F_{sw}, F_w) \right) + \log k.$$

□

A.5 Proof of Theorem 3

Proof sketch. By defining nine variables associated with given models, we substitute key components in the proof of Theorem 2 to derive a set of inequalities among these variables. Through a series of carefully designed transformations, we reformulate the triangle-like inequalities involving three remainder terms. Ultimately, leveraging tools from statistical learning theory, several inequalities in information-theoretic analysis, and the properties of specific functions, we sequentially demonstrate that these three remainder terms become infinitesimal as $n \to \infty$ and $\epsilon \to 0$.

Let $L(\cdot, \cdot)$ be $\mathrm{KL}(\cdot, \cdot)$. For a clear presentation, denote

$$
\begin{aligned}
A &= L(F_s, F_{sw}) \\
B &= L(F_{sw}, F_w) \\
C &= L(F_s, F_w) \\
D &= L(F^\star, F_s) = \varepsilon \\
E &= L(F^\star, F_{sw}) \\
F &= L(F^\star, F_w) \\
G &= L(F^\star, \hat{F}_{sw}) \\
H &= L(\hat{F}_{sw}, F_{sw}) \\
I &= L(\hat{F}_{sw}, F_w).
\end{aligned}
$$

Now we start the proof of Theorem 3. A uniform convergence result and two claims used in the proof are provided at the end of the proof. The proof is strongly motivated by Theorem 4 in Yao et al. (2025). While our work primarily focuses on classification, their Theorem 4 is specifically centered on regression.

Proof. Note that by virtue of the range of $f^\star$, f_w and all functions in $\mathcal{F}$ being absolutely bounded, and L is also bounded.

Due to $F^\star \notin V_s$, we replace $F^\star$ with F_s in the final step of proof of Theorem 2, we obtain

$$
C \geq A + B. \tag{24}
$$

Recall that $\langle f, g \rangle_E \triangleq \mathbb{E}_{x \sim \mathcal{P}}[f(x)^T g(x)]$, which is used here for a clear presentation. So we have

$$
\begin{aligned}
E &= A + D - \mathbb{E}_x \sum_{i=1}^{k} ([F^\star(x)]_i - [F_s(x)]_i) \log \frac{[F_{sw}(x)]_i}{[F_s(x)]_i} \\
&= A + D - \underbrace{\left\langle F^\star - F_s, \log \frac{F_{sw}}{F_s} \right\rangle_E}_{t_1}.
\end{aligned}
$$

The $\log$ here is element-wise. Using the similar notation, we have the following

$$
E = A + D - \underbrace{\left\langle F^\star - F_s, \log \frac{F_{sw}}{F_s} \right\rangle_E}_{t_1}, \tag{25}
$$

$$
F = C + D - \underbrace{\left\langle F^\star - F_s, \log \frac{F_w}{F_s} \right\rangle_E}_{t_2}, \tag{26}
$$

$$
G = E - H - \underbrace{\left\langle \hat{F}_{sw} - F^\star, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E}_{t_3}. \tag{27}
$$

Combining (24) and (25), we get

$$E \leq C + D - B - t_1. \tag{28}$$

By a uniform convergence argument (Lemma 6), we have that with probability at least $1 - \delta$ over the draw of $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ that were used to construct $\hat{F}_{sw}$,

$$I \leq B + \underbrace{\mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right)}_{t_4} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)}_{t_5}. \tag{29}$$

Combining (28) with (29) and we have

$$E \leq C + D - I - t_1 + t_4 + t_5. \tag{30}$$

Combining (26) with (30) and we have

$$E \leq F - I - t_1 + t_2 + t_4 + t_5. \tag{31}$$

Combining (27) with (31) and we have

$$G \leq F - I - H - t_1 + t_2 - t_3 + t_4 + t_5. \tag{32}$$

We replace $F^\star$ with $\hat{F}_{sw}$ in the final step of proof of Theorem 2 and obtain:

$$I \geq H + B. \tag{33}$$

Combining (33) with (29) and we have

$$0 \leq H \leq t_4 + t_5 = \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{34}$$

Combining (34) with (32) and we have

$$G \leq F - I - t_1 + t_2 - t_3 + t_4 + t_5. \tag{35}$$

While t_4 and t_5 are known in (29), we analyze t_1, t_2 and t_3 one by one.

Deal with t_1. We know that

$$t_1 = \left\langle F^\star - F_s, \log \frac{F_{sw}}{F_s} \right\rangle_E.$$

Using the fact that $\frac{F_{sw}(x)}{F_s(x)} \leq \frac{1}{\gamma}$, we have

$$\begin{aligned}
|t_1| &\leq \frac{1}{\gamma} \mathbb{E}_x \sum_{i=1}^{k} |[F^\star(x)]_i - [F_s(x)]_i| \\
&= \frac{2}{\gamma} \mathbb{E}_x \mathrm{D}_{\mathrm{TV}}(F^\star(x), F_s(x)) && \text{(Definition of TV distance)} \\
&\leq \frac{2}{\gamma} \mathbb{E}_x \sqrt{\frac{1}{2} \mathrm{D}_{\mathrm{KL}}(F^\star(x) \| F_s(x))} && \text{(Pinsker's inequality)} \\
&\leq \frac{2}{\gamma} \sqrt{\frac{1}{2} \mathbb{E}_x \mathrm{D}_{\mathrm{KL}}(F^\star(x) \| F_s(x))} && \text{(Jensen's inequality)} \\
&= \frac{2}{\gamma} \sqrt{\frac{1}{2} L(F^\star, F_s)} && \text{(Definition of } L\text{)} \\
&= \frac{1}{\gamma} \sqrt{2\varepsilon} && (36)
\end{aligned}$$

Therefore,

$$|t_1| = \mathcal{O}(\sqrt{\varepsilon}). \tag{37}$$

Deal with t_2. The proof for t_2 is similar for t_1. In particular, replacing F_{sw} with F_w in the above and we can get

$$|t_2| = O(\sqrt{\varepsilon}). \tag{38}$$

Deal with t_3. We know that

$$t_3 = \left\langle \hat{F}_{sw} - F^\star, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E = \mathbb{E}_x \sum_{i=1}^{k} ([\hat{F}_{sw}(x)]_i - [F^\star(x)]_i) \log \frac{[F_{sw}(x)]_i}{[\hat{F}_{sw}(x)]_i}.$$

According to Lemma 6, with probability at least $1-\delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left| L(\hat{F}_{sw}, F_w) - L(F_{sw}, F_w) \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{39}$$

Notice that

$$\begin{aligned} H &= L(F_{sw}, \hat{F}_{sw}) \\ &= L(F_w, F_{sw}) - L(F_w, \hat{F}_{sw}) + \left\langle F_w + F_{sw}, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E. \end{aligned} \tag{40}$$

Substitute (34) and (39) into Equation (40) with the triangle inequality for absolute values, we get

$$\left| \left\langle F_w + F_{sw}, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Since $|F_w(x) + F_{sw}(x)|$ is lower bounded, we have

$$\left| \left\langle \vec{1}, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Since $|\hat{F}_{sw}(x) - F^\star(x)|$ is upper bounded, there holds

$$|t_3| = \left| \left\langle \hat{F}_{sw} - F^\star, \log \frac{F_{sw}}{\hat{F}_{sw}} \right\rangle_E \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{41}$$

Therefore, combing (37), (38) and (41), we have

$$|t_1| + |t_2| + |t_3| \leq O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{42}$$

Finally, combing (29) and (35) with (34) and (42), we get the result:

$$L(F^\star, \hat{F}_{sw}) \leq L(F^\star, F_w) - L(\hat{F}_{sw}, F_w) + O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_\mathcal{F}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where in the last inequality, we instantiate asymptotics with respect to $\varepsilon \to 0$ and $n \to \infty$.

□

Here are some tools used in the above proof.

Lemma 6 (Uniform convergence)**.** *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be an i.i.d. training sample, where each $x_i \sim \mathcal{P}$ and $y_i = F_w(x_i)$ for a target function F_w. For a fixed strong model representation h_s, we employ reverse KL loss in WTSG:*

$$f_{sw} = \textit{argmin}_{f \in \mathcal{F}_s} L(f \circ h_s, F_w) = \textit{argmin}_{f \in \mathcal{F}_s} \mathbb{E}_{x \sim \mathcal{P}} \left[\sum_{i=1}^{k} [f \circ h_s(x)]_i \log \frac{[f \circ h_s(x)]_i}{[F_w(x)]_i} \right],$$

$$\hat{f}_{sw} = \textit{argmin}_{f \in \mathcal{F}_s} \hat{L}(f \circ h_s, F_w) = \textit{argmin}_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{j=1}^{n} \left[\sum_{i=1}^{k} [f \circ h_s(x_j)]_i \log \frac{[f \circ h_s(x_j)]_i}{[F_w(x_j)]_i} \right].$$

Assume that the range of F_w and functions in $\mathcal{F}_s$ is absolutely bounded. Then, with probability at least $1 - \delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left| L(\hat{F}_{sw}, F_w) - L(F_{sw}, F_w) \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}} \right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} \right),$$

where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$.

Proof. The proof follows lemma 4 in Yao et al. (2025). Swap the order of the two elements in $L(\cdot, \cdot)$ and $\hat{L}_{\mathcal{P}}(\cdot, \cdot)$ in their proof and we can prove the result. □

Claim 1 (Yao et al. (2025))**.** *Let $f(x), g(x) \in [\gamma, 1]$ where $\gamma > 0$. If there exists $\xi > 0$ such that $\int_{\mathcal{X}} |f(x) - g(x)| \, dx \leq \xi$, then there holds*

$$\int_{\mathcal{X}} |\log f(x) - \log g(x)| \, dx \leq \frac{1}{\gamma} \xi.$$

Claim 2 (Yao et al. (2025))**.** *Let $f(x), g(x) \in [\gamma, 1]$ where $\gamma > 0$. If there exists $\xi > 0$ such that $\int_{\mathcal{X}} |\log f(x) - \log g(x)| \, dx \leq \xi$, then there holds*

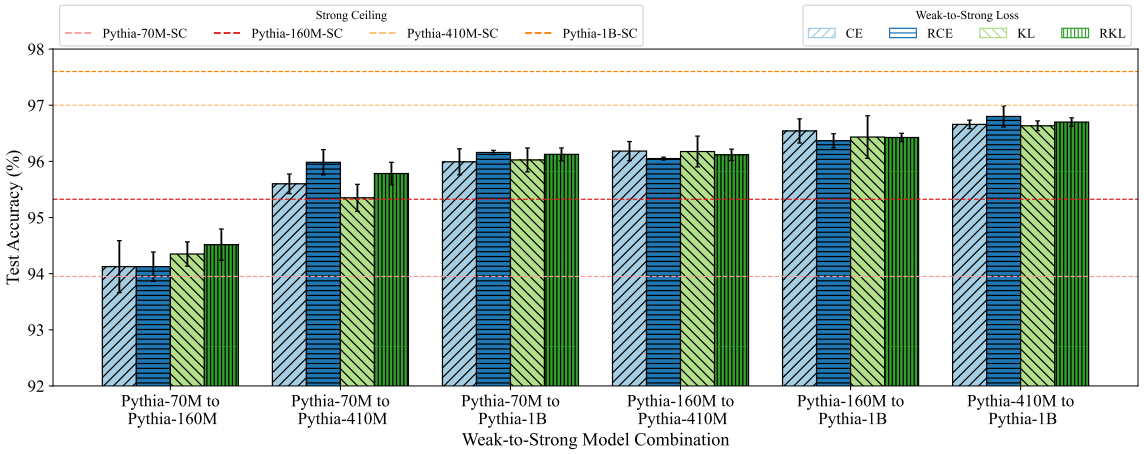
$$\int_{\mathcal{X}} |f(x) - g(x)| \, dx \leq \xi.$$

B Additional Experimental Details and Results

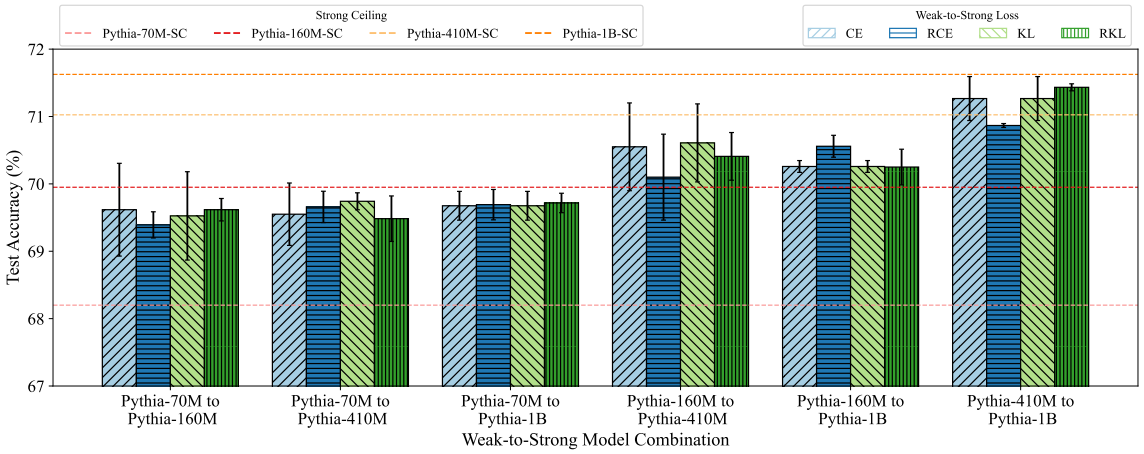
We first provide a detailed explanation of the evaluation metric. To determine the effectiveness of a model F in distinguishing between the selected and rejected completions (y_c and y_r) for a given prompt x, we require that F ranks the chosen completion higher than the rejected one. This condition is formulated as $F(y_c) - F(y_r) > 0$ for each pair $\tilde{x} = (x; y_c, y_r)$, implying that $F(\tilde{x})$ should exceed 0.5. Consequently, the test accuracy is defined as the fraction of instances where $F(\tilde{x}) > 0.5$.

B.1 Results of Pythia

The overall trends observed in Figure 4 are similar with those in Figure 2. Our analysis of the results in Figure 4 further highlights those insights: First, the accuracy exhibits a consistent upward trend from left to right, reinforcing the finding that the generalization capability of the strong model improves when a more capable weak model is utilized as the supervisor. Second, the results demonstrate that in the majority of experimental settings (7 out of 12), reverse losses outperform forward losses, leading to stronger model performance. Given the superior capabilities of the Pythia series compared to the GPT-2 series (Biderman et al., 2023), as well as the fact that Pythia's strong ceiling model outperforms GPT-2, a key implication emerges. When the Pythia series serves as a weak model, it may generate less noise on non-target labels. As a result, the potential advantages of reverse losses are diminished, leading to only a slight improvement of reverse losses over forward losses. Finally, across almost all of the settings (10 out of 12), the strong model trained with reverse KL and CE losses achieves superior performance compared to its weak supervisor. This observation is in full agreement with our theoretical predictions, further validating the effectiveness of reverse losses in enhancing model performance.



(a) Results of Pythia-series on CAI-Harmless



(b) Results of Pythia-series on helpful set of HH-RLHF

Figure 4: Results of Pythia-series. "SC" denotes the strong ceiling model, and "A to B" indicates the use of weak teacher "A" to supervise strong student "B". The terms CE, RCE, KL, and RKL refer to cross-entropy loss, reverse cross-entropy loss, forward KL divergence loss, and reverse KL divergence loss, respectively. Error bars represent the standard deviation across three runs of the experiment.

B.2 Auxiliary Confidence Loss

As highlighted by Burns et al. (2023), we explore an alternative approach: introducing an additional regularization term designed to enhance the strong model's confidence in its predictions using standard cross-entropy loss, which is called "Auxiliary Confidence Loss" in Burns et al. (2023):

$$L_{\text{conf}}(f) = (1-\alpha)\cdot \underbrace{\text{CE}\left(F_w, f\circ h_s\right)}_{\text{vanilla cross-entropy loss}} + \alpha\cdot \underbrace{\text{CE}\left(\hat{f}_t\circ h_s, f\circ h_s\right)}_{R(f)}, \tag{43}$$

where α is the weight constant, $R(f)$ is the regularization term, and $\hat{f}_t$ corresponds to hardened strong model predictions using a threshold t, i.e., for any x:

$$\hat{f}_t\circ h_s(x) = \mathbb{I}(f\circ h_s(x) > t) \in \{0,1\},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Rewrite Equation (43) as the minimization objective in WTSG:

$$f_{sw} = \text{argmin}_{f\in\mathcal{F}_s}\, L_{\text{conf}}(f). \tag{44}$$

As advocated by Burns et al. (2023), this regularization serves to mitigate overfitting to weak supervision, thereby improving the overall performance of the strong model. Therefore, to further explore the advantage

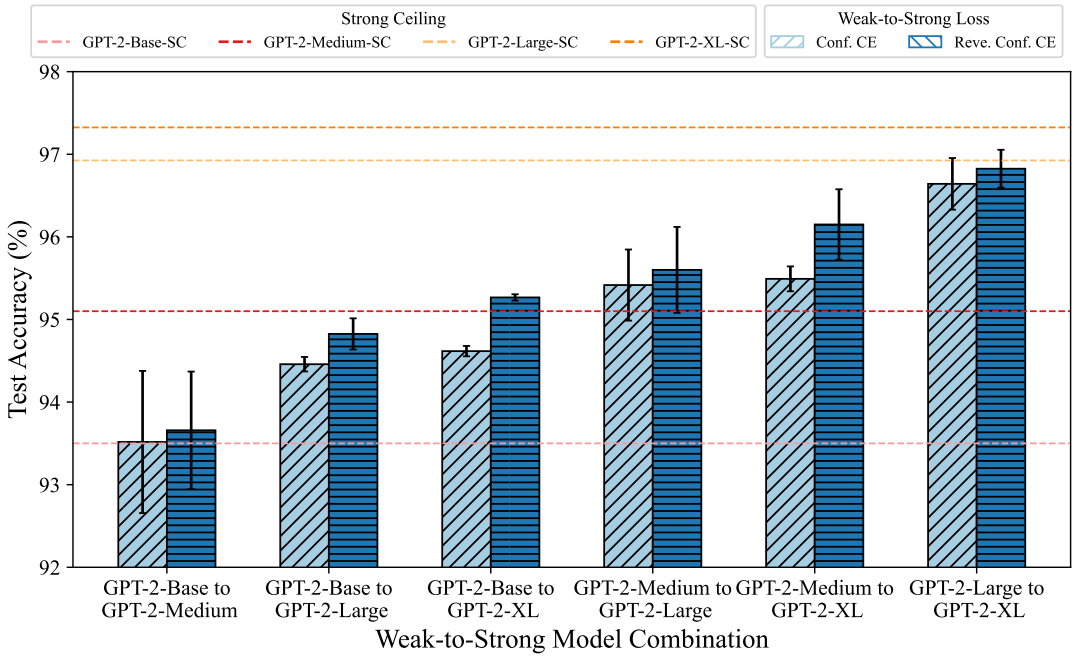


Figure 5: Results of GPT-2 series on CAI-Harmless. "SC" denotes the strong ceiling model, and "A to B" indicates the use of weak teacher "A" to supervise strong student "B". The terms "Conf. CE" and "Reve. Conf. CE" refer to the auxiliary confidence loss with vanilla cross-entropy loss (Equation (43)) and reverse cross-entropy loss (Equation (45)), respectively. Error bars represent the standard deviation across three runs of the experiment.

of reverse cross-entropy loss, we replace the vanilla cross-entropy with reverse cross-entropy in $L_{\text{conf}}(f)$ and conduct WTSG using the following objective:

$$\begin{aligned} f_{sw}^r &= \text{argmin}_{f \in \mathcal{F}_s} \, L_{\text{conf}}^r(f) \\ &= \text{argmin}_{f \in \mathcal{F}_s} \, (1-\alpha) \cdot \underbrace{\text{CE}\left(f \circ h_s, F_w\right)}_{\text{reverse cross-entropy loss}} + \alpha \cdot R(f). \end{aligned} \tag{45}$$

We set $\alpha = 0.2$ to ensure that the reverse/forward CE loss dominates the regularization, because we use a small batch size here and we want to reduce the negative impact of the randomness and instability brought by the auxiliary confidence loss within a single batch. The experimental comparison between f_{sw} and f_{sw}^r is presented in Figure 5. First, by combining the observations from Figure 2 and Figure 5, we observe that the application of auxiliary confidence loss slightly enhances the performance of the strong model, consistent with the findings of Burns et al. (2023). Second, the use of reverse cross-entropy loss consistently enables the strong model to outperform its counterpart trained with standard cross-entropy loss. This finding, combined with previous experimental results in this work, highlights the superior effectiveness of reverse losses compared to forward losses.