

Do Student LLMs Inherit OOD Robustness? Invariance-Weighted Distillation for Reliable Knowledge Transfer

Dileesha Kannangara, Sanghamitra Dutta

University of Maryland, College Park
dileesha@umd.edu, sanghamd@umd.edu

Abstract

Knowledge distillation (KD) aims to compress high-performance teacher LLMs into lightweight students. However, distilled students often exhibit substantial performance degradation in out-of-distribution (OOD) settings, a critical gap that remains underexplored. We identify two compounding mechanisms causing OOD performance degradation: (1) data spuriousness: students can learn spurious correlations in the distillation dataset over genuine causal relationships; and (2) teacher capability: standard KD treats all samples uniformly, ignoring whether the teacher is guided by causal features or misled by spurious shortcuts on a given sample. To address these challenges, we propose Invariance-Weighted Distillation (IWD), a theoretically grounded framework that dynamically reweights training samples using an estimate of the teacher’s causal reliance derived from prediction invariance across multiple synthetic environments. IWD perturbs spurious cues while preserving core semantics, assigning higher distillation weights to samples whose teacher predictions remain invariant, indicating greater reliance on causal rather than spurious features. We theoretically show that IWD reduces the student’s Spurious-to-Causal (S2C) gradient ratio compared to standard uniformly weighted KD, driving the student toward more invariant representations. Experiments on four NLP benchmarks (MNLI, SQuAD-v2, CoNLL-2003 NER, and SST-2) across two model families (DeBERTa-v3 and Qwen-2.5) demonstrate that IWD consistently outperforms strong KD baselines on OOD evaluations while maintaining competitive in-distribution (ID) performance. Specifically, IWD achieves the highest accuracy in 15 out of 16 OOD benchmarks and improves average OOD performance over standard KD by 4.34 percentage points on NLI and 14.94 percentage points on QA.

Introduction

Large Language Models (LLMs) are central to modern applications due to their remarkable ability to solve complex tasks (OpenAI 2023; Yang et al. 2024; DeepSeek-AI 2025). However, high computational and memory demands hinder training, fine-tuning, and deployment in resource-constrained environments (Girija et al. 2025). Knowledge distillation (KD) addresses these challenges by training a compact student model to mimic the predictive behavior of a high-capacity teacher model across a designated distillation dataset (Hinton, Vinyals, and Dean 2015; Xu et al. 2024). Distilled students deliver competitive performance with drastically reduced resource overhead, enabling efficient real-world deployment.

However, student models often demonstrate significant performance degradation on out-of-distribution (OOD) data compared to the in-distribution (ID) data used during training (Du et al. 2023; Stanton et al. 2021). OOD data can be formally defined as data originating from a distribution that differs from that encountered during any stage of training prior to deployment (Shen et al. 2021). Although teacher models themselves are not fully robust to OOD data, we observe that the performance gap between teacher and student models becomes substantially larger under OOD settings than under ID settings. This suggests that current KD techniques do not effectively transfer the teacher model’s existing ability to handle OOD data and directly undermine this capability in the resulting student models (Stacey and Rei 2024). This leads to two central questions: *To what extent do student LLMs inherit OOD robustness from their teachers, and how can we transform distillation to enable reliable knowledge transfer under distribution shifts?*

We identify two main limitations of current KD techniques that cause OOD performance degradation. First, **under existing distillation methods, students learn spurious correlations from the distillation dataset instead of genuine causal relationships**. Standard KD approaches train the student model to mimic the teacher’s outputs (Hinton, Vinyals, and Dean 2015; Gu et al. 2024) or intermediate representations (Romero et al. 2015; Sun et al. 2019, 2020; Jiao et al. 2020) by minimizing discrepancies over the distillation dataset. Additionally, distillation minimizes the standard cross-entropy loss over these training samples. In doing so, the student learns dataset-specific spurious correlations, reducing its ability to generalize when encountering OOD data. Second, **standard distillation treats all training samples uniformly, ignoring the teacher’s ability to isolate causal signals from spurious correlations across individual samples**. In practice, a teacher model may successfully resist spurious shortcuts on samples with strong causal semantics, yet succumb to spurious correlations on others. Standard KD applies a uniform weight across all samples, inadvertently forcing the student to imitate the teacher even when the teacher relies on brittle, spurious associations. For reliable knowledge transfer, distillation should prioritize samples where the teacher exhibits high fidelity to invariant causal features while suppressing spurious influences.

To address these limitations, we propose **Invariance-**

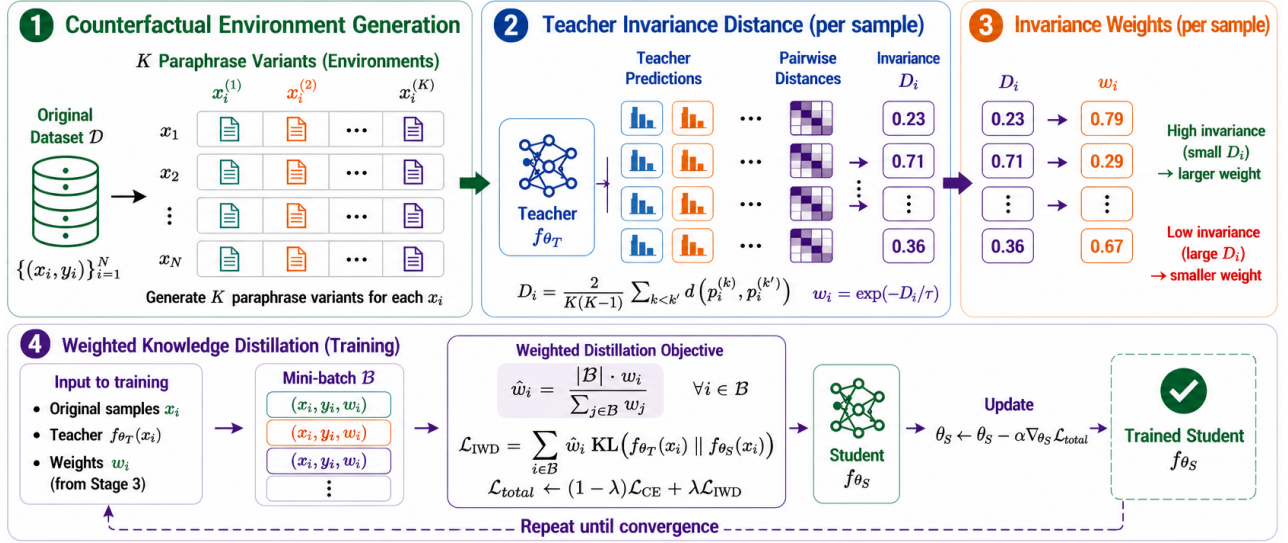


Figure 1: **Overview of IWD.** The pipeline operates in four stages: (1) **Counterfactual Environment Generation:** Generating semantically equivalent paraphrase variants; (2) **Teacher Invariance Distance:** Measuring prediction consistency across variants using the teacher predictions; (3) **Weight Calculation:** Assigning higher reliability weights to samples where teacher predictions remain invariant; and (4) **Weighted Knowledge Distillation:** Training the student model using the weighted loss objective.

Weighted Distillation (IWD), a novel strategy that uses a weighted objective function to prioritize samples where the teacher relies on causal relationships over spurious correlations (see Fig. 1). We first generate multiple synthetic environments for each training sample by altering its spurious features while preserving its original semantic meaning and label. This environment generation step resolves the first limitation by exposing the student to diverse spurious perturbations across counterfactual variants with fixed core semantics. Distilling across these environments prevents the student from learning dataset-specific spurious correlations and forces it to learn invariant representations that generalize across distribution shifts. To address the second limitation, we measure the discrepancy among the teacher’s output distributions across these environments. This distance quantifies the teacher’s prediction invariance, i.e., how strongly the teacher relies on invariant features versus environment-specific cues. We incorporate this distance into the distillation loss as a weight, increasing emphasis on samples where the teacher exhibits high prediction invariance and, consequently, high fidelity to invariant causal features.

Contributions Our main contributions are as follows:

- **Conceptualizing OOD Vulnerabilities in KD.** We identify two key failure modes of standard KD under distribution shifts: (1) student models learn spurious correlations from the distillation dataset, and (2) current KD methods treat all training samples equally, failing to recognize the teacher’s ability to distinguish between causally robust and spurious-dominated samples.
- **Invariance-Weighted Distillation (IWD).** We propose IWD, a novel distillation strategy that assigns sample-specific weights based on the teacher’s prediction invari-

ance across synthetically perturbed environments (see Fig. 1). To overcome Limitation 1, IWD varies spurious features while preserving core causal semantics, exposing the student to diverse counterfactual variations of each sample. To address Limitation 2, IWD prioritizes samples where the teacher exhibits high invariance, and thus high fidelity to causal features. Our strategy equips the student with improved robustness to spurious features over standard KD and conventional data augmentation.

- **Theoretical Guarantee.** We provide a theoretical guarantee for IWD by leveraging a metric called the Spurious-to-Causal (S2C) gradient ratio (see Theorem 1). We prove that IWD yields a lower S2C ratio than standard distillation, guaranteeing that the student’s expected gradient aligns more strongly with invariant signals while reducing spurious feature transfer.
- **Empirical Validation.** We evaluate IWD across four distinct NLP tasks (NLI, extractive QA, NER, and sentiment classification) and two model families (DeBERTa-v3 and Qwen-2.5), using several OOD datasets per task under various subsampling regimes. IWD consistently improves OOD robustness while maintaining ID performance. For instance, across model families, IWD improves average OOD performance over standard KD by 4.34 and 14.94 percentage points on NLI and QA, respectively.

Related Works Existing KD methods (Hinton, Vinyals, and Dean 2015; Zhu et al. 2024; Xu et al. 2024; Romero et al. 2015; Sun et al. 2019; Zagoruyko and Komodakis 2017; Tian, Krishnan, and Isola 2020; Kim and Rush 2016; Hamman et al. 2025a; Dissanayake et al. 2025) focuses on matching teacher and student outputs, intermediate features, or attention patterns. However, the robustness of distilled models under distribution shifts remains underexplored. While

ShiftKD (Zhang et al. 2025) shows that distillation suffers substantial performance degradation under covariate and correlation shifts, other approaches attempt to enhance robustness via confidence-guided data augmentation (Popp et al. 2025), adversarial distillation (Goldblum et al. 2019), or distillation from robust foundation models (Zhou et al. 2023). Yet, these methods primarily target conventional deep learning models, leaving OOD robustness in distilled LLMs underexplored.

Prior work on OOD robustness in distilled LLMs seeks to improve performance through data augmentation, sample selection, or adaptive distillation strategies. (Stacey and Rei 2024) use domain-targeted data augmentation and sampling strategies to improve NLI distillation under distribution shift. Prior works Chen et al. (2023) and Hamman et al. (2025a) leverage LLMs to generate label-flipping counterfactuals, prompting students to learn causal features over spurious correlations. Other works introduce difficulty-aware distillation using model pruning (Du et al. 2023) or OOD-guided data generation to expose students to more diverse samples (Gholami et al. 2024). Notably, these works do not explicitly enforce invariance across semantically equivalent perturbations and rely on uniform distillation objectives that treat all training samples equally. *In contrast, our approach evaluates the teacher’s prediction invariance across synthetically perturbed environments with varying spurious features, leveraging the teacher-invariance signal to weight the distillation loss toward more reliable samples.*

Preliminaries

Knowledge distillation (Xu et al. 2024; Hamman et al. 2025a) trains a smaller student f_{θ_S} to imitate a larger teacher f_{θ_T} . For an input x with target label y , the student’s loss is:

$$\mathcal{L}_{KD} = (1-\lambda) \mathcal{L}_{CE}(y, p_S(x)) + \lambda \mathcal{L}_{KL}(p_T(x) \parallel p_S(x)), \quad (1)$$

Here, $\mathcal{L}_{CE}$ is the cross-entropy (CE) loss, $\mathcal{L}_{KL}$ is the KL divergence (Kullback and Leibler 1951) at each generation step over the vocabulary, $p_S(x) = \text{softmax}(f_{\theta_S}(x))$ and $p_T(x) = \text{softmax}(f_{\theta_T}(x))$ are the output distributions, and $\lambda \in [0, 1]$ balances hard-label CE and soft target imitation.

Out-of-distribution (OOD) refers to the setting in which a deployed model make predictions on inputs that come from a different distribution from the one it was trained on. A model is said to be OOD-robust if its performance does not degrade substantially under such a shift. Formally, let P_{train} denote the training distribution and $P_{\text{OOD}} \neq P_{\text{train}}$ denote a shifted evaluation distribution. Standard supervised learning offers no formal guarantee in this regime, and trained models often degrade sharply (Hendrycks and Dietterich 2019; Geirhos et al. 2020), particularly when they exploit features that are spuriously correlated with the label in P_{train} but uncorrelated in P_{OOD} . OOD robustness is measured via the metric:

$$\phi(\theta) = \mathbb{E}_{(x,y) \sim P_{\text{OOD}}} [\mathbf{1}[f_{\theta}(x) = y]], \quad (2)$$

where θ denotes the parameters of the model under evaluation, $f_{\theta}(x)$ is its predicted label for input x , and $\mathbf{1}[\cdot]$ is the indicator function returning 1 when its argument is true and 0 otherwise. The expectation $\mathbb{E}_{(x,y) \sim P_{\text{OOD}}} [\cdot]$ averages over input-label pairs drawn from the shifted distribution, so $\phi(\theta)$ is the expected accuracy of f_{θ} on P_{OOD} .

Problem Setup. Our goal is to study how KD in LLMs affects the OOD robustness of the student model. Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a labeled training set drawn from an in-distribution (ID) source P_{train} , and let P_{OOD} represent a shifted evaluation distribution. Given a teacher model f_{θ_T} pre-trained on P_{train} , our objective is to learn a smaller student model f_{θ_S} that maximizes the OOD evaluation metric $\phi(\theta_S)$.

Specifically, our goal is to distill student parameters θ_S that learn genuine causal relationships while suppressing spurious correlations, and achieve higher OOD performance than standard distillation, i.e., $\phi(\theta_S) > \phi(\theta_S^{\text{KD}})$, where θ_S^{KD} denotes parameters learned via standard KD.

Main Contributions

We introduce **Invariance-Weighted Distillation (IWD)**, a framework enhancing OOD robustness in distilled LLMs. Our strategy operates in four stages: (i) generation of counterfactual environment variants that vary spurious features while preserving the semantic meaning; (ii) computation of the teacher-invariance distance that quantifies the teacher’s prediction variation across those environments; (iii) mapping distance to sample weights; and (iv) distilling using our proposed weighted distillation objective.

Stage 1: Counterfactual Environment Generation. For each training sample $(x_i, y_i) \in \mathcal{D}$, we construct a set of K counterfactual variants:

$$\mathcal{E}_i = \{x_i^{(0)}, x_i^{(1)}, \dots, x_i^{(K-1)}\}, \quad (3)$$

where $x_i^{(0)} \triangleq x_i$ is the original sample, and $x_i^{(1)}, \dots, x_i^{(K-1)}$ are paraphrases generated across different environments $e^{(k)}$. These environments are designed to perturb surface-level spurious features such as register, syntactic structure, domain vocabulary, capitalization, and sentence length, while holding the core semantic features determining y_i fixed.

Formally, each input can be decomposed into an invariant *causal* component $\mathbf{c}_i$ that determines the label and an environment-specific *spurious* component $\mathbf{s}_i^{(k)}$. The generator ensures these variants satisfy counterfactual label invariance:

$$P(y_i \mid \mathbf{c}_i, \mathbf{s}_i^{(k)}) = P(y_i \mid \mathbf{c}_i) \quad \forall k \in \{0, 1, \dots, K-1\}. \quad (4)$$

Equation 4 guarantees that the label distribution conditioned on causal features remains identical across all environments, isolating variation strictly to the spurious components. We construct $\mathcal{E}_i$ offline by prompting an auxiliary language model (GPT-4.1-mini) to generate specific spurious perturbations. To enforce label invariance, any paraphrase that would alter the target label y_i is rejected. The union of these environment sets forms the augmented training dataset $\mathcal{D}_{\text{aug}} = \bigcup_{i=1}^N \mathcal{E}_i$, which is precomputed once and shared across all experiments.

Our semantics-preserving counterfactual environment generation strategy directly addresses the first limitation by preventing the student from learning dataset-specific spurious correlations. Exposing the model to these variants influences both components of the training objective: (i) distillation loss ($\mathcal{L}_{KL}$), which forces the student to match teacher predictions across varied spurious contexts rather than context-dependent noise; and (ii) task loss ($\mathcal{L}_{CE}$), which trains the student directly

on perturbed inputs to ensure ground-truth predictions rely on invariant semantics rather than surface shortcuts.

Stage 2: Teacher Invariance Distance. Using the environment set $\mathcal{E}_i$, we directly quantify the invariance of the teacher’s predictions across counterfactual perturbations.

Definition 1 (Teacher Invariance Distance). *Let $\hat{y}_i^{(k)} = f_{\theta_T}(x_i^{(k)})$ denote the teacher’s prediction on the k -th variant in a task-appropriate output space. The teacher invariance distance D_i is defined as the mean pairwise disagreement across all environment pairs:*

$$D_i \triangleq \binom{K}{2}^{-1} \sum_{0 \leq u < v \leq K-1} d(\hat{y}_i^{(u)}, \hat{y}_i^{(v)}). \quad (5)$$

Here, $d(\cdot, \cdot)$ is a non-negative, symmetric distance metric.

Since causal features $\mathbf{c}_i$ remain fixed across variants, D_i directly isolates the teacher’s sensitivity to spurious perturbations $\mathbf{s}_i^{(k)}$. A small D_i indicates that predictions are invariant and grounded in causal semantics, making the sample a reliable source for distillation. Conversely, a large D_i reveals that predictions fluctuate with surface cues, signaling the teacher’s reliance on spurious features that will not generalize in OOD settings. The distance metric $d(\cdot, \cdot)$ is tailored to the output space of the task. For classification (e.g., SST-2, MNLI), d measures distributional divergence over softmax outputs; for structured symbolic outputs like QA text spans or NER entities (SQuAD-v2, CoNLL-2003), d computes per-sample F_1 distance ($1 - F_1$) on decoded string predictions.

Stage 3: Weight Calculation. To convert the invariance distance D_i into a per-sample reliability weight, we apply a monotonically decreasing exponential function: $w_i = \exp(-D_i/\tau)$. Here, $\tau > 0$ is a temperature hyperparameter. A small τ heavily penalizes teacher variance, while a large τ approaches uniform weighting, recovering environment-augmented distillation as $\tau \rightarrow \infty$. Furthermore, to prevent implicit learning rate shifts caused by varying D_i distributions across mini-batches, we normalize weights over each mini-batch $\mathcal{B}$: $\tilde{w}_i = |\mathcal{B}| \cdot \frac{w_i}{\sum_{j \in \mathcal{B}} w_j}$, ensuring $\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \tilde{w}_i = 1$. This normalization preserves a unit average weight per batch while adjusting relative per-sample importance.

Stage 4: Weighted Knowledge Distillation. During training, we sample mini-batches $\mathcal{B} \subseteq \mathcal{D}_{\text{aug}}$, where each item corresponds to an environment variant $x_i^{(k)}$ derived from parent sample i . Given the batch-normalized weights $\tilde{w}_i$, the weighted distillation objective over the mini-batch is:

$$\mathcal{L}_{\text{IWD}}(\theta_S) = \sum_{(i,k) \in \mathcal{B}} \tilde{w}_i \cdot \text{KL}(f_{\theta_T}(x_i^{(k)}) \parallel f_{\theta_S}(x_i^{(k)})). \quad (6)$$

Combining $\mathcal{L}_{\text{IWD}}$ with the student task loss across variants, $\mathcal{L}_{\text{CE}} = \sum_{(i,k) \in \mathcal{B}} \text{CE}(y_i, f_{\theta_S}(x_i^{(k)}))$, yields our total training loss: $\mathcal{L}_{\text{total}} = (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{IWD}}$ (Eq. 1). By dynamically scaling the loss with $\tilde{w}_i$, IWD prioritizes samples where the teacher demonstrates invariant predictions while aggressively down-weighting unstable targets ($w_i \rightarrow 0$). This suppresses

Algorithm 1: Invariance Weighted Distillation (IWD)

Input: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, teacher f_{θ_T} , student f_{θ_S}

Parameters: No. of environments K , temperature τ , learning rate α , distance metric $d(\cdot, \cdot)$

Output: Trained student f_{θ_S}

```

1: – Stage 1: Counterfactual Environment Generation –
2:  $\mathcal{D}_{\text{aug}} \leftarrow \emptyset$ 
3: for each  $(x_i, y_i) \in \mathcal{D}$  do
4:   Set  $x_i^{(0)} \leftarrow x_i$  {Original sample}
5:   Generate  $K - 1$  paraphrases  $\{x_i^{(1)}, \dots, x_i^{(K-1)}\}$ 
6:    $\mathcal{E}_i \leftarrow \{x_i^{(0)}, x_i^{(1)}, \dots, x_i^{(K-1)}\}$ 
7:    $\mathcal{D}_{\text{aug}} \leftarrow \mathcal{D}_{\text{aug}} \cup \mathcal{E}_i$ 
8: end for
9: – Stage 2: Teacher Invariance Distance –
10: for each  $i \in \{1, \dots, N\}$  do
11:    $\hat{y}_i^{(k)} \leftarrow f_{\theta_T}(x_i^{(k)}) \quad \forall k \in \{0, \dots, K-1\}$ 
12:    $D_i \leftarrow \binom{K}{2}^{-1} \sum_{0 \leq u < v \leq K-1} d(\hat{y}_i^{(u)}, \hat{y}_i^{(v)})$ 
13: end for
14: – Stage 3: Weight Calculation –
15: for each  $i \in \{1, \dots, N\}$  do
16:    $w_i \leftarrow \exp(-D_i/\tau)$  {Parent weight once per-sample}
17: end for
18: – Stage 4: Weighted Knowledge Distillation –
19: while not converged do
20:   Sample mini-batch  $\mathcal{B} \subseteq \mathcal{D}_{\text{aug}}$ , where each item is indexed by  $(i, k)$ 
21:    $\tilde{w}_i \leftarrow |\mathcal{B}| \cdot w_i / \sum_{(j,k') \in \mathcal{B}} w_j \quad \forall (i, k) \in \mathcal{B}$ 
22:    $\mathcal{L}_{\text{CE}} \leftarrow \sum_{(i,k) \in \mathcal{B}} \text{CE}(y_i, f_{\theta_S}(x_i^{(k)}))$ 
23:    $\mathcal{L}_{\text{IWD}} \leftarrow \sum_{(i,k) \in \mathcal{B}} \tilde{w}_i \cdot \text{KL}(f_{\theta_T}(x_i^{(k)}) \parallel f_{\theta_S}(x_i^{(k)}))$ 
24:    $\mathcal{L}_{\text{total}} \leftarrow (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{IWD}}$ 
25:    $\theta_S \leftarrow \theta_S - \alpha \nabla_{\theta_S} \mathcal{L}_{\text{total}}$ 
26: end while

```

teacher signals on brittle samples, forcing the student to rely on ground-truth supervision instead. The complete training procedure for IWD is detailed in Algorithm 1.

Theoretical Guarantee. We now provide a theoretical guarantee for weighting the distillation objective by the teacher invariance distance. We first introduce the *Spurious-to-Causal (S2C) gradient ratio* (Definition 2), which measures gradient contamination from spurious feature variations under fixed causal semantics. We show that IWD achieves a strictly lower S2C ratio than uniform distillation, aligning student updates more closely with invariant causal features.

Definition 2 (Spurious-to-Causal (S2C) Gradient Ratio). *Let $p_T(x) = \text{softmax}(z_T(x))$ and $p_S(x) = \text{softmax}(z_S(x))$ denote teacher and student output distributions for logits z_T, z_S . Decomposing the teacher’s output distribution into invariant and spurious components,*

$$p_T(x_i^{(e)}) = \pi_i + s_i^{(e)}, \quad \pi_i := \mathbb{E}_e[p_T(x_i^{(e)})], \quad (7)$$

where $\pi_i \in \Delta^{C-1}$ is the invariant component, and $s_i^{(e)}$ is the spurious component with $\mathbb{E}_e[s_i^{(e)}] = \mathbf{0}$ and variance $\mathbb{E}_e[\|s_i^{(e)}\|^2] = \sigma_i^2$. The S2C gradient ratio under sample weighting W is defined as:

$$\rho^2(W) = \mathbb{E}_{i \sim W}[\|g_i^S\|^2] / \mathbb{E}_{i \sim W}[\|g_i^C\|^2], \quad (8)$$

where $g_i^C = (\pi_i - p_S(x_i))^\top J_i$ is the causal gradient component, $g_i^S = s_i^{(e_i)\top} J_i$ is the spurious gradient component, and $J_i := \nabla_{\theta_S} z_S(x_i) \in \mathbb{R}^{C \times |\theta_S|}$ is the student logit Jacobian.

A lower S2C ratio indicates a causally dominated gradient, ensuring the student prioritizes invariant predictions over non transferable surface perturbations. We analyze the distillation objective on dataset $\mathcal{D}$ across environments $\{\mathcal{E}_i\}_{i=1}^N$, comparing uniform weighting ($W_i = 1/N$) against IWD weighting ($W_i^{\text{IWD}} \propto \exp(-D_i/\tau)$). As $K \rightarrow \infty$, D_i converges to a strictly increasing function of spurious variance σ_i^2 (Proposition 1 in Appendix A.4), allowing D_i to directly isolate teacher spurious reliance. See Appendix A for more details.

Our analysis relies on three standard assumptions:

(A1) Zero-Mean Spurious Noise: Spurious components $s_i^{(e_i)}$ are independent across samples with zero mean and variance σ_i^2 , representing unstructured prediction variation rather than systematic dataset bias.

(A2) Independence of Spurious Variance: Spurious variance is independent of both the causal residual and gradient scale: $\sigma_i^2 \perp (c_i^2, \Phi_i)$, where $c_i^2 := \|\pi_i - p_S(x_i)\|^2$. This decouples teacher prediction invariance from student task difficulty.

(A3) Isotropic Logit Jacobians: Student logit Jacobians satisfy $J_i J_i^\top = \Phi_i I_C$ with per sample scale $\Phi_i \perp \sigma_i^2$, ensuring parameter updates do not inherently favor spurious directions over causal ones.

Theorem 1 (IWD Reduces the S2C Gradient Ratio). *Under Assumptions (A1)–(A3), let $W_i^{\text{IWD}} = \exp(-\sigma_i^2/\tau)/Z$ and $W_i^{\text{unif}} = 1/N$. Then,*

$$\rho^2(W^{\text{IWD}}) \leq \rho^2(W^{\text{unif}}),$$

with equality if and only if σ_i^2 is constant across all samples.

Experiments

We evaluate **IWD** on four tasks spanning the major NLP output structures—paired classification, single-sentence classification, span extraction, and token-level labeling—across two model families of markedly different scale and architecture. Additional details are given in Appendix B.

Datasets and Models. We train on standard ID splits and evaluate OOD generalization across two robustness benchmarks per task (Table 4). To cover diverse architectures, we distill both encoder-only models (deberta-v3-xsmall, ~71M parameters from -base teachers) and decoder-only models (Qwen2.5-0.5B, ~494M from -1.5B teachers); full setup details are in Appendix B.2.

Baselines. We compare against five key KD baselines: (1) **Vanilla KD** (Hinton, Vinyals, and Dean 2015) (forward KL divergence); (2) **RevKL** (Gu et al. 2024) (reverse KL divergence); (3) **DKD** (Zhao et al. 2022) (decoupled target and non-target KL); (4) **LWD** (Romero et al. 2015; Sun

et al. 2019, 2020; Jiao et al. 2020) (intermediate-layer representation alignment); and (5) **AugKD** (Vanilla KD on counterfactually augmented data, isolating our invariance weighting scheme from pure data augmentation).

Training Setup. We train models using AdamW in mixed precision ($T = 2.0, \tau = 1.0, K = 6$); full implementation details and hyperparameters are provided in Appendix B.4.

Results and Analysis

IWD outperforms baselines in OOD settings. Across both architectures, IWD achieves the top performance in **15 of 16** OOD settings, with the largest gains in Extractive QA and NLI (see Table 1). On DeBERTa-v3 Extractive QA, IWD outperforms Vanilla KD by **+23.83 points** on Natural Questions (28.23% $\rightarrow$ 52.06%) and **+15.64 points** on NewsQA (39.47% $\rightarrow$ 55.11%). On Qwen2.5, Natural Questions F1 rises by **+16.36 points** (40.89% $\rightarrow$ 57.25%). On NLI (HANS), IWD outperforms Vanilla KD by **+9.12 points** on DeBERTa-v3 (61.45%) and **+3.26 points** on Qwen2.5 (54.95%). For NER and Sentiment Analysis, IWD consistently matches or exceeds top baselines across all OOD splits (e.g., 80.38% on OntoNotes 5.0 and 87.58% on CR).

IWD is stable across experiments. Non-augmented distillation baselines exhibit high performance variance across tasks and architectures (see Table 1). For instance, LWD achieves competitive DeBERTa-v3 NLI performance (52.99% on HANS) but collapses on Qwen2.5 QA (2.37% on Natural Questions). Similarly, DKD yields OOD gains on DeBERTa-v3 QA (44.26% on NewsQA) yet lags Vanilla KD on NLI and Sentiment Analysis, while RevKL degrades on DeBERTa-v3 QA despite strong Qwen2.5 results. This variance highlights a key advantage of IWD: while standard baselines depend heavily on the task and architecture, IWD remains stable and consistently top-performing across both encoder-only and decoder-only models. This instability of baselines stems from rigid assumptions embedded within these objectives, such as target logit geometry, specific head dynamics, or a particular teacher softmax shape, that hold only for select backbone-task pairs. IWD instead evaluates teacher invariance under counterfactual perturbations. Invariance to spurious shifts is a model-agnostic property; thus, IWD operates seamlessly across both encoder-only and decoder-only students and stays in the top-performing tier on every OOD split we tested.

IWD outperforms basic data augmentation. Compared to AugKD, which uses augmented inputs without variance-aware weighting, IWD achieves a 100% win rate across **16 of 16** OOD settings, improving by **+1.60 points** on average (see Table 1). Advantages are most pronounced on decoder-only architectures like Qwen2.5, where IWD outperforms AugKD on Yelp-Polarity (**+4.84 points**), CR (**+4.80 points**), and SNLI (**+3.58 points**). This confirms that applying invariant weight allocation (w_i) to environment variants provides a clear advantage over basic data augmentation alone.

Encoder vs. Decoder OOD Performance. Under IWD, encoder-only DeBERTa-v3 achieves higher OOD performance than decoder-only Qwen2.5, widening the gap observed under Vanilla KD (see Table 1). On HANS, DeBERTa’s lead over Qwen expands from **+0.65 points** under Vanilla KD

Task / Dataset	Vanilla KD	RevKL	DKD	LWD	AugKD	IWD (Ours)
DeBERTa-v3 Family (Encoder-Only)						
NLI (Accuracy)						
MNLI (ID)	82.02 ± 0.29	82.08 ± 0.21	79.91 ± 0.47	82.56 ± 0.12	85.31 ± 0.18	84.66 ± 0.27
HANS (OOD-1)	52.34 ± 0.57	51.45 ± 0.81	51.33 ± 0.79	52.99 ± 1.10	<u>60.95 ± 1.30</u>	61.45 ± 1.01
SNLI (OOD-2)	78.26 ± 0.58	77.49 ± 0.26	75.19 ± 0.85	78.97 ± 0.92	<u>80.21 ± 0.24</u>	80.91 ± 0.50
Extractive QA (F1)						
SQuAD-v2 (ID)	70.64 ± 0.20	72.40 ± 0.81	68.01 ± 1.81	70.97 ± 1.35	51.50 ± 0.55	51.40 ± 0.77
NewsQA (OOD-1)	39.47 ± 0.77	39.95 ± 0.87	44.26 ± 0.68	36.10 ± 2.48	<u>54.35 ± 0.66</u>	55.11 ± 0.45
Natural Questions (OOD-2)	28.23 ± 0.74	28.11 ± 2.67	30.52 ± 2.33	25.02 ± 5.17	<u>51.93 ± 0.94</u>	52.06 ± 0.33
NER (Entity F1)						
CoNLL-2003 (ID)	92.18 ± 0.11	90.59 ± 0.21	90.11 ± 0.21	90.71 ± 0.20	<u>92.63 ± 0.22</u>	92.64 ± 0.10
WNUT-17 (OOD-1)	47.84 ± 0.42	43.68 ± 0.51	46.88 ± 0.35	46.34 ± 0.52	<u>46.78 ± 0.48</u>	48.01 ± 0.31
OntoNotes 5.0 (OOD-2)	<u>78.99 ± 0.21</u>	76.26 ± 0.19	77.68 ± 0.32	78.12 ± 0.11	<u>80.27 ± 0.22</u>	80.38 ± 0.11
Sentiment Analysis (Accuracy)						
SST-2 (ID)	92.73 ± 0.19	<u>92.84 ± 0.39</u>	92.71 ± 0.22	92.00 ± 0.30	93.21 ± 0.15	92.78 ± 0.18
Yelp-Polarity (OOD-1)	<u>91.39 ± 0.19</u>	91.28 ± 0.24	91.36 ± 0.38	90.26 ± 0.65	90.98 ± 0.56	91.42 ± 0.12
CR (OOD-2)	<u>87.29 ± 1.17</u>	87.08 ± 0.75	87.18 ± 0.48	85.81 ± 0.25	85.32 ± 0.48	87.58 ± 0.98
Qwen-2.5 Family (Decoder-Only)						
NLI (Accuracy)						
MNLI (ID)	81.88 ± 0.50	82.13 ± 0.30	79.24 ± 0.53	80.27 ± 0.36	82.48 ± 0.20	83.21 ± 0.19
HANS (OOD-1)	51.69 ± 0.38	52.04 ± 0.85	51.61 ± 1.06	51.20 ± 0.63	<u>52.79 ± 1.21</u>	54.95 ± 1.68
SNLI (OOD-2)	78.25 ± 0.63	<u>79.03 ± 0.96</u>	75.28 ± 1.68	76.38 ± 0.86	76.99 ± 0.79	80.57 ± 0.67
Extractive QA (F1)						
SQuAD-v2 (ID)	75.48 ± 0.31	<u>75.02 ± 0.23</u>	74.68 ± 0.37	49.86 ± 0.29	47.96 ± 0.43	48.84 ± 0.21
NewsQA (OOD-1)	16.43 ± 0.73	<u>16.03 ± 1.30</u>	14.88 ± 2.11	5.11 ± 1.21	<u>19.53 ± 0.65</u>	20.34 ± 0.18
Natural Questions (OOD-2)	40.89 ± 0.81	43.73 ± 2.19	35.57 ± 1.02	2.37 ± 1.05	<u>57.12 ± 0.60</u>	57.25 ± 0.50
NER (Entity F1)						
CoNLL-2003 (ID)	68.30 ± 0.25	67.69 ± 0.30	65.00 ± 0.22	35.52 ± 0.36	69.83 ± 0.20	69.29 ± 0.10
WNUT-17 (OOD-1)	<u>27.48 ± 0.32</u>	24.88 ± 0.42	26.13 ± 0.52	12.81 ± 1.30	25.02 ± 0.43	27.77 ± 0.33
OntoNotes 5.0 (OOD-2)	<u>54.90 ± 0.40</u>	54.93 ± 1.00	53.57 ± 0.68	26.58 ± 0.90	54.67 ± 0.42	55.03 ± 0.67
Sentiment Analysis (Accuracy)						
SST-2 (ID)	93.46 ± 0.54	93.17 ± 0.58	93.43 ± 0.22	60.17 ± 3.97	92.04 ± 0.89	90.11 ± 0.66
Yelp-Polarity (OOD-1)	84.69 ± 3.11	85.99 ± 1.59	<u>84.99 ± 2.91</u>	51.17 ± 1.63	77.80 ± 0.31	83.64 ± 0.86
CR (OOD-2)	74.57 ± 10.66	<u>76.87 ± 7.02</u>	74.95 ± 9.50	62.07 ± 3.33	73.00 ± 1.09	77.80 ± 2.29

Table 1: Main performance comparison of distilled student models across four NLP tasks for both DeBERTa-v3 and Qwen-2.5 families, evaluated on one ID and two OOD benchmarks per task. Results report mean ± standard deviation across multiple random seeds. The best performance per benchmark is in bold, and the second-best is underlined.

to **+6.50 points** under IWD; on NewsQA, this lead reaches **+34.77 points**. We attribute this gap to how teacher distance is measured. Encoder-only models output probabilities over just a few task labels, making teacher disagreement clean and straightforward to measure. In contrast, decoder-only models generate text token-by-token across a vast vocabulary. Here, minor surface variations, such as alternative word choices or sentence structures, add noise to the distance metric, even when the underlying meaning remains identical. Thus, while generative students still benefit from IWD, fully unlocking their performance may require measuring invariance at the semantic level rather than the token level.

IWD Performance on ID. In-distribution (ID), IWD remains fully competitive with standard baselines across tasks, achieving high accuracy on MNLI (84.66%), CoNLL-2003 (92.64%), and SST-2 (92.78%) (see Table 1). However, a

limitation emerges in **Extractive QA**: on SQuAD-v2, both AugKD and IWD exhibit a sharp ID performance drop compared to Vanilla KD (51.40% vs. 70.64% on DeBERTa-v3, and 48.84% vs. 75.48% on Qwen2.5). This ID–OOD trade-off likely stems from environment generation shifting the training distribution away from the ID baseline, highlighting a direction for future research.

Task Structure and IWD Performance. IWD gains correlate with the baseline OOD degradation and shortcut susceptibility of a task. On DeBERTa-v3 Extractive QA, Vanilla KD drops over 42 percentage points from the ID SQuAD-v2 benchmark (70.64%) to OOD Natural Questions (28.23%). IWD dramatically recovers this gap, yielding +23.83 points on Natural Questions (28.23% → 52.06%) and +15.64 points on NewsQA (39.47% → 55.11%). Conversely, Sentiment Analysis shows high baseline stability under Vanilla KD—

Data Size (N)	MNLI-m (ID)	MNLI-mm (ID)	HANS (OOD1)	SNLI (OOD2)
DeBERTa-v3				
5k	83.27	83.79	51.98	79.73
8k	84.29	84.54	52.40	79.79
10k	84.84	85.11	60.83	80.56
30k	86.39	86.42	62.64	82.47
50k	87.37	87.17	68.19	82.72
Qwen-2.5				
5k	81.87	82.69	53.19	78.39
8k	82.70	83.82	54.09	79.34
10k	83.50	84.36	59.05	81.17
30k	84.98	86.04	60.27	82.15
50k	85.56	86.09	66.22	81.74

Table 2: Data scaling analysis of IWD on MNLI across varying sample sizes (N) for DeBERTa-v3/Qwen-2.5.

Environments	ID	OOD1	OOD2	OOD3	OOD4
MNLI Benchmark					
<i>Datasets</i>	<i>MNLI</i>	<i>HANS</i>	<i>SNLI</i>	—	—
$E = 3$	84.34	56.27	78.98	—	—
$E = 4$	84.54	57.02	79.68	—	—
$E = 5$	84.78	58.41	80.10	—	—
$E = 6$	84.80	56.28	80.60	—	—
SST-2 Benchmark					
<i>Datasets</i>	<i>SST-2</i>	<i>SemEval</i>	<i>CR</i>	<i>Yelp</i>	<i>FPB</i>
$E = 2$	92.09	88.72	87.23	91.39	90.75
$E = 3$	92.55	89.08	87.50	91.54	90.14
$E = 4$	92.89	89.02	87.23	91.72	90.29
$E = 5$	92.43	89.38	86.44	91.55	90.09

Table 3: Impact of varying the number of environments (K) in IWD across MNLI/SST-2 benchmarks on accuracy.

dropping just 1.34 points from SST-2 (92.73%) to Yelp-Polarity (91.39%)—resulting in modest IWD gains (+0.03 to +0.29 points). This aligns with task geometry: Extractive QA and NLI heavily exploit lexical and positional shortcuts exposed by counterfactuals, whereas Sentiment Analysis and NER are inherently context-invariant. Thus, while IWD reliably improves shortcut-heavy tasks, stable tasks require precise tuning to maximize OOD generalization.

Dataset Scaling Effect. Across datasets, performance improves as data grows from 5k to 50k samples (see Table 2). DeBERTa-v3 exhibits stronger OOD scaling than Qwen2.5: on HANS, its accuracy increases by **+16.21 points** (51.98% $\rightarrow$ 68.19%) versus a **+13.03-point** gain for Qwen2.5 (53.19% $\rightarrow$ 66.22%). On SNLI, both models scale consistently, each gaining roughly **+3.00 points**.

Impact of Environment Count (K). Table 3 evaluates performance across environment variants ($K \in \{3, 4, 5, 6\}$). Increasing K beyond 3 generally improves ID and OOD accuracy by exposing the model to broader shortcut perturbations. On MNLI, scaling from $K = 3$ to $K = 5$ yields a **+2.14-point** gain on HANS (56.27% $\rightarrow$ 58.41%), while SNLI accuracy rises to 80.60% at $K = 6$. On SST-2, $K = 5$

achieves peak ID performance (92.89%) and peak Yelp OOD accuracy (91.72%). Overall, $K = 5$ provides the optimal balance between environmental diversity and efficiency.

Ablations. We summarize the key takeaways of our ablations (complete setups in Appendix C): (1) **Sensitivity to Weight Temperature (τ):** We evaluate the weighting temperature $\tau \in [0.1, 30.0]$, which controls the sharpness of weighting (Table 6). While classification and NER remain highly robust across $\tau \in [0.3, 3.0]$, Extractive QA requires sharp down-weighting ($\tau = 0.1$) to suppress high-variance span predictions, yielding substantial OOD gains of up to +9.8 F1. (2) **Sensitivity to Distance Metric Selection:** We evaluate how distance metric formulations impact performance across task domains (Table 7). Aligning metric geometry with the output space maximizes performance: bounded Jensen–Shannon divergence (D_{JS}) stabilizes classification by bounding extreme softmax tails, while structural metrics like Levenshtein distance and F1 (QA answer string drift) and Entity Set F1 (span-level NER) eliminate per-token noise to achieve top OOD performance on NQ (38.19 F1) and WNUT17 (46.04 F1). This proves that choosing the teacher invariance distance metric is a key structural design decision following directly from the output space. (3) **Environment Variant Utilization:** We compare distilling with all counterfactual variants against restricting the student to the original clean text or a single random variant (Table 8). Distilling across *All Variants* consistently yields the highest OOD robustness—outperforming single-variant baselines by up to 4.2 points on HANS—proving that full exposure to environmental perturbations is necessary for invariant feature learning. (4) **Weight Normalization:** We evaluate default weight normalization against an unnormalized baseline across experiments (Table 9). Empirical performance between the two is virtually identical (< 0.2-point shift), but batch-mean normalization is essential as it stabilizes training by decoupling learning rates from teacher divergence scales. (5) **Weight Distribution Across Tasks:** We inspect per-sample weight statistics across benchmarks to analyze how IWD allocates sample importance to tasks of varying complexity (Table 10). Across all datasets, effective sample sizes remain high ($\geq 95.0\%$), confirming that IWD performs smooth, task-adaptive reweighting rather than aggressive data pruning.

Discussion and Limitations

In this work, we first ask: do distilled student LLMs inherit OOD robustness under existing methods? We present **Invariance-Weighted Distillation (IWD)**, a framework that enhances OOD robustness in student LLMs without requiring architectural modifications, auxiliary classification heads, or adversarial training. By consistently outperforming KD and data augmentation baselines across OOD tasks, our results demonstrate a clear advantage over current strategies.

Limitations include: (i) the computational overhead of synthetic environment generation, which can impact distillation performance; and (ii) a reliance on output-level distance measures, leaving fine-grained token-level invariance as a promising unexplored research direction. Future work would look into extensions to instruction tuning (Fu, Hamman, and

Dutta 2026), particularly leveraging token-level measures, as well as, broader paraphrasing techniques (Hamman et al. 2025b). It would also be interesting to study transferability of other kinds of spurious correlations (Halder et al. 2025; Egea, Halder, and Dutta 2025) beyond OOD robustness.

References

- Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv:1907.02893*.
- Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Association for Computational Linguistics.
- Chen, Z.; Gao, Q.; Bosselut, A.; Sabharwal, A.; and Richardson, K. 2023. DISCO: Distilling Counterfactuals with Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5514–5528. Toronto, Canada: Association for Computational Linguistics.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*.
- Derczynski, L.; Nichols, E.; van Erp, M.; and Limsopatham, N. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 140–147. Association for Computational Linguistics.
- Dissanayake, P.; Hamman, F.; Halder, B.; Sucholutsky, I.; Zhang, Q.; and Dutta, S. 2025. Quantifying Knowledge Distillation using Partial Information Decomposition. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, 4474–4482. PMLR.
- Du, M.; Mukherjee, S.; Cheng, Y.; Shokouhi, M.; Hu, X.; and Awadallah, A. H. 2023. Robustness Challenges in Model Distillation and Pruning for Natural Language Understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Egea, D.; Halder, B.; and Dutta, S. 2025. Vision: Robust and interpretable code vulnerability detection leveraging counterfactual augmentation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 812–823.
- Fu, Y.; Hamman, F.; and Dutta, S. 2026. T-shirt: Token-selective hierarchical data selection for instruction tuning. *Advances in Neural Information Processing Systems*, 38: 113932–113958.
- Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11): 665–673.
- Gholami, M.; Akbari, M.; Hu, C.; Masrani, V.; Wang, Z. J.; and Zhang, Y. 2024. GOLD: Generalized Knowledge Distillation via Out-of-Distribution-Guided Language Data Generation. In *Findings of the Association for Computational Linguistics: NAACL 2024*.
- Girija, S. S.; Kapoor, S.; Arora, L.; Pradhan, D.; Raj, A.; and Shetgaonkar, A. 2025. Optimizing LLMs for Resource-Constrained Environments: A Survey of Model Compression Techniques. In *Proceedings of the 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMP-SAC)*, 1272–1277. IEEE.
- Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2019. Adversarially Robust Distillation. *arXiv:1905.09747*.
- Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.
- Halder, B.; Hamman, F.; Dissanayake, P.; Zhang, Q.; Sucholutsky, I.; and Dutta, S. 2025. Towards Formalizing Spuriousness of Biased Datasets Using Partial Information Decomposition. *Transactions on Machine Learning Research*.
- Hamman, F.; Dissanayake, P.; Fu, Y.; and Dutta, S. 2025a. Few-Shot Knowledge Distillation of LLMs With Counterfactual Explanations. In *Advances in Neural Information Processing Systems*, volume 38.
- Hamman, F.; Zhu, C.; Kumar, A.; Peng, X.; Dutta, S.; Liu, D.; and Samuel, A. 2025b. Improving consistency in retrieval-augmented systems with group similarity rewards. *arXiv preprint arXiv:2510.04392*.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *International Conference on Learning Representations (ICLR)*.
- Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*.
- Hu, M.; and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. ACM.
- Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4163–4174. Stroudsburg, Penn.: Association for Computational Linguistics.
- Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1317–1327.
- Kullback, S.; and Leibler, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.
- Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural Questions: A Benchmark for Study of Corpus-Level Language Understanding. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.
- Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4): 782–796.
- McCoy, R. T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Association for Computational Linguistics.
- OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.
- Popp, N.; Hein, M.; Laube, K. A.; and Schott, L. 2025. Improving Knowledge Distillation Under Unknown Covariate Shift Through Confidence-Guided Data Augmentation. arXiv:2506.02294.
- Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 143–152. Association for Computational Linguistics.
- Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don’t Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2: 784–789.
- Romero, A.; Ballas, N.; Ebrahimi Kahou, S.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*. San Diego, California, USA.
- Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. Association for Computational Linguistics.
- Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *IEEE Transactions on Knowledge and Data Engineering*.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Association for Computational Linguistics.
- Stacey, J.; and Rei, M. 2024. Distilling Robustness into Natural Language Inference Models with Domain-Targeted Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Stanton, S.; Izmailov, P.; Kirichenko, P.; Alemi, A. A.; and Wilson, A. G. 2021. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 6906–6919.
- Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Empirical Methods in Natural Language Processing (EMNLP)*, 4323–4332.
- Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2158–2170. Stroudsburg, Penn.: Association for Computational Linguistics.
- Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR-20)*. Addis Ababa, Ethiopia.
- Tjong Kim Sang, E. F.; and De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147.
- Trischler, A.; Wang, T.; Yuan, X.; Howland, J.; Sordoni, A.; Harmar, P.; Offermann, C.; and Suleman, K. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200.
- Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. 2013. OntoNotes Release 5.0. *Linguistic Data Consortium*, 23.
- Williams, A.; Nangia, N.; and Bowman, S. R. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 1112–1122. Association for Computational Linguistics.
- Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; and Zhou, T. 2024. A Survey on Knowledge Distillation of Large Language Models. arXiv:2402.13116.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2.5 Technical Report. arXiv:2412.15115.
- Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *Proceedings of the Fifth International Conference on Learning Representations (ICLR-17)*. Toulon, France.
- Zhang, S.; Luo, Y.; Lyu, Z.; and Chen, X. 2025. ShiftKD: Benchmarking Knowledge Distillation under Distribution Shift. *Neural Networks*, 188: 107504.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.
- Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11962.
- Zhou, A.; Wang, J.; Wang, Y.-X.; and Wang, H. 2023. Distilling Out-of-Distribution Robustness from Vision-Language

Foundation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhu, X.; Li, J.; Liu, Y.; Ma, C.; and Wang, W. 2024. A Survey on Model Compression for Large Language Models. *Transactions of the Association for Computational Linguistics*, 12: 1556–1577.

Appendix

A. Theoretical Background and Proof of Theorem 1

A.1. Background: Invariance across Environments

Standard empirical risk minimization treats all training data as a single pool, allowing models to exploit spurious correlations that change or disappear at test time. To address this, frameworks like Invariant Risk Minimization (IRM) (Arjovsky et al. 2019) conceptually distinguish between *causal* features (which yield a stable, invariant predictive relationship across environments) and *spurious* features (whose importance fluctuates depending on the environment).

Crucially, standard IRM frameworks operate in a traditional supervised learning setting, enforcing invariance globally at the model level (e.g., via regularisation penalties across macroscopic environment datasets). In contrast, our framework adapts this philosophy specifically for *knowledge distillation*, operationalising invariance entirely at the per-sample level. Rather than constraining the student’s representations directly, we posit that a teacher model’s prediction on a single example across counterfactual environments can be decomposed into two parts: an environment-invariant base prediction, and an environment-dependent (spurious) residual. This sample-level view of teacher invariance provides the necessary foundation for the gradient decomposition we derive next.

A.2. Derivation of Definition 2: Exact Gradient Decomposition under the KL Objective

We derive the causal and spurious gradient components g_i^C and g_i^S of Definition 2 directly from the KL distillation objective used in Algorithm 1 (Line 18). Although the softmax is non-linear in the logits, the gradient of the KL loss with respect to the *student’s logits* is exactly linear in the *teacher’s output probabilities*. Since Definition 2 places the IRM decomposition on the teacher’s output distribution (7), the causal/spurious decoupling below is *exact*: no linearisation or Taylor approximation is required.

Distillation loss. For a single training example i observed in environment e_i , the temperature-scaled KL distillation loss is

$$\mathcal{L}_i = \tau_{\text{KD}}^2 \text{KL}(p_T(x_i^{(e_i)}) \parallel p_S(x_i)). \quad (9)$$

The KL gradient is linear in the teacher target. Writing $\text{KL}(p_T \parallel p_S) = \sum_c p_{T,c} \log p_{T,c} - \sum_c p_{T,c} \log p_{S,c}$, the first (entropy) term is constant in θ_S , so minimising $\mathcal{L}_i$ is equivalent to minimising the cross-entropy $H(p_T, p_S)$. Using the standard softmax–cross-entropy derivative, $\partial[-$

$\sum_c p_{T,c} \log p_{S,c}] / \partial z_{S,c} = \frac{1}{\tau_{\text{KD}}} (p_{S,c} - p_{T,c})$, we obtain

$$\nabla_{z_S} \mathcal{L}_i = \tau_{\text{KD}} (p_S(x_i) - p_T(x_i^{(e_i)})), \quad (10)$$

the classical distillation gradient of Hinton, Vinyals, and Dean (2015); the scalar τ_{KD} is absorbed into the learning rate below. The essential observation is that (10) is *affine in the teacher’s probability vector* p_T . The softmax non-linearity acts on the teacher’s logits *before* the loss is formed; once the IRM decomposition is stated at the level of the teacher’s output distribution, the gradient inherits the decomposition additively.

Chain rule. With $J_i = \nabla_{\theta_S} z_S(x_i)$ the Jacobian of the student’s logits,

$$\nabla_{\theta_S} \mathcal{L}_i = (p_S(x_i) - p_T(x_i^{(e_i)}))^T J_i. \quad (11)$$

As in the MSE case, J_i is a Jacobian with respect to the *student’s parameters*, encoding how each weight affects each output logit.

Substituting the teacher decomposition. Substituting $p_T(x_i^{(e_i)}) = \pi_i + s_i^{(e_i)}$ from (7) into (11), the error term splits additively:

$$\nabla_{\theta_S} \mathcal{L}_i = \underbrace{(p_S(x_i) - \pi_i)^T J_i}_{\text{causal gradient (pre-flip)}} - \underbrace{s_i^{(e_i)T} J_i}_{\text{spurious gradient (pre-flip)}}. \quad (12)$$

This step is exact and uses only the linearity of (11) in p_T .

Gradient descent and sign convention. Under gradient descent with step size α , substituting (12) into $\theta_S \leftarrow \theta_S - \alpha \nabla_{\theta_S} \mathcal{L}_i$ and collecting the terms *added* to θ_S :

$$\theta_S \leftarrow \theta_S + \alpha \underbrace{(\pi_i - p_S(x_i))^T J_i}_{g_i^C} + \alpha \underbrace{s_i^{(e_i)T} J_i}_{g_i^S}. \quad (13)$$

The subtraction in gradient descent flips $(p_S - \pi_i)$ to $(\pi_i - p_S)$, so g_i^C points the student toward the teacher’s environment-invariant output distribution π_i ; g_i^S needs no flip because its minus sign in (12) is cancelled by the descent subtraction. Under (A3), $\|g_i^S\|^2 = \Phi \|s_i^{(e_i)}\|^2$ and $\|g_i^C\|^2 = \Phi c_i^2$ with $c_i^2 := \|\pi_i - p_S(x_i)\|^2$, exactly as required in Step 1 of the proof of Theorem 1.

Per-example spurious reliance σ_i^2 . Averaging over environments, $\mathbb{E}_e[\|s_i^{(e)}\|^2] = \sigma_i^2 \geq 0$ defines the per-example teacher spurious reliance. $\sigma_i^2 = 0$ implies $p_T(x_i^{(e)}) = \pi_i$ for all e : the teacher’s output distribution for example i is completely invariant across environments.

Remark 1 (MSE distillation as a special case). *If the distillation loss is instead the squared error $\|f_{\theta_S}(x_i) - f_{\theta_T}(x_i^{(e_i)})\|^2$ on raw outputs, the identical derivation applies with p_T, p_S, z_S replaced by $f_{\theta_T}, f_{\theta_S}, f_{\theta_S}$ and π_i by $\mu_i = \mathbb{E}_e[f_{\theta_T}(x_i^{(e)})]$: the MSE gradient $2(f_{\theta_S} - \mu_i - s_i^{(e_i)})^T \nabla_{\theta_S} f_{\theta_S}$ is likewise affine in the teacher target. Both objectives therefore induce the same causal/spurious gradient structure, and Theorem 1 applies to either.*

Remark 2 (Probability-space vs. logit-space invariance). *Definition 2 states invariance in the space of output distributions: the spurious component is the zero-mean deviation of $p_T(x_i^{(e)})$ around its environment average π_i . Because the softmax is non-linear, zero-mean spurious variation in probability space does not coincide with zero-mean variation in logit space (Jensen’s inequality); the two decompositions define distinct but equally valid notions of invariance. We adopt the probability-space formulation throughout, for two reasons: (i) it is the space in which the KL distillation gradient is linear in the teacher target, making the decomposition in (12) exact; and (ii) it is the natural formalisation of prediction invariance — a teacher is invariant on example i precisely when its predictive distribution does not change across environments.*

A.3. Discussion of Assumptions in Theorem 1

Theorem 1 rests on three assumptions. We discuss their meaning, when they are expected to hold, and what would happen if they were violated.

(A1) Independent spurious components with zero mean.

We assume $s_i^{(e)}$ are independent across training samples, with $\mathbb{E}_e[s_i^{(e)}] = 0$ and $\mathbb{E}_e[\|s_i^{(e)}\|^2] = \sigma_i^2$. The zero-mean condition is the defining property of spurious features under IRM: any signal that has a consistent direction across all environments would, by Definition 1, be part of the causal component μ_i rather than the spurious component $s_i^{(e)}$. Independence across samples holds when counterfactual environments are generated independently per sample, as in our paraphrase-based construction.

(A2) Independence of spurious and causal residuals: $\sigma_i^2 \perp c_i^2$. This is the key structural assumption. σ_i^2 measures how sensitive the teacher is to environmental surface variation for sample i . $c_i^2 = \|\mu_i - f_{\theta_S}(x_i)\|^2$ measures how far the student currently is from the teacher’s causal prediction for the same sample. (A2) requires these two quantities to be independent across training samples.

The assumption holds when spurious correlations arise from *environmental surface confounders* — such as lexical register, syntactic style, or demographic markers in NLI — that affect the teacher’s cross-environment invariance but are unrelated to the semantic difficulty of the inference problem (which drives c_i^2). It would be violated if, for instance, semantically harder samples systematically attracted more spurious surface features, creating a positive correlation between σ_i^2 and c_i^2 . This can be tested empirically by computing the Pearson correlation between D_i (our proxy for σ_i^2) and the per-sample training loss of the student; a correlation near zero supports (A2).

(A3) Homogeneous student gradient norms. We assume $\|\nabla_{\theta_S} f_{\theta_S}(x_i)\|^2 = \Phi$ is constant across all training samples. This simplifies the S2C ratio to a ratio of scalar weighted sums (Step 1 of the proof), making the Chebyshev argument clean. The assumption can be relaxed: if gradient norms Φ_i are allowed to vary but are additionally assumed independent of σ_i^2 (i.e. $\Phi_i \perp \sigma_i^2$), then the Chebyshev step applies to the

products $\Phi_i \sigma_i^2$ and the conclusion of Theorem 1 still holds. We adopt the simpler homogeneous form here for exposition.

A.4. Proposition 1 and Proof.

In the main text, we substitute the theoretical spurious variance σ_i^2 with the empirical distance D_i computed over K counterfactual environments. This substitution is mathematically justified under Assumption (A1), as we formalise below.

Proposition 1 (Convergence of Empirical Invariance Distance). *Let $D_i = \frac{2}{K(K-1)} \sum_{k < k'} \|p_T(x_i^{(k)}) - p_T(x_i^{(k')})\|^2$ be the empirical average pairwise distance between the teacher’s softened output distributions across K independently sampled counterfactual environments. Under Assumption (A1), as $K \rightarrow \infty$, D_i converges in probability to $2\sigma_i^2$.*

Proof of Proposition 1. Substitute the IRM decomposition $p_T(x_i^{(k)}) - p_T(x_i^{(k')}) = (\pi_i + s_i^{(k)}) - (\pi_i + s_i^{(k')}) = s_i^{(k)} - s_i^{(k')}$ into the distance metric. The environment-invariant causal component μ_i cancels out:

Expanding the squared L_2 norm gives:

$$\|s_i^{(k)} - s_i^{(k')}\|^2 = \|s_i^{(k)}\|^2 + \|s_i^{(k')}\|^2 - 2\langle s_i^{(k)}, s_i^{(k')} \rangle. \quad (14)$$

Taking the expectation over the independent environments k and k' , and applying Assumption (A1) (which states $\mathbb{E}[s_i^{(k)}] = 0$ and $\mathbb{E}[\|s_i^{(k)}\|^2] = \sigma_i^2$):

$$\begin{aligned} \mathbb{E}[\|s_i^{(k)} - s_i^{(k')}\|^2] &= \mathbb{E}[\|s_i^{(k)}\|^2] + \mathbb{E}[\|s_i^{(k')}\|^2] \\ &\quad - 2\langle \mathbb{E}[s_i^{(k)}], \mathbb{E}[s_i^{(k')}] \rangle \\ &= \sigma_i^2 + \sigma_i^2 - 2\langle 0, 0 \rangle \\ &= 2\sigma_i^2. \end{aligned} \quad (15)$$

Because D_i is a U-statistic estimating this expected pairwise distance across K independent samples, by the Law of Large Numbers for U-statistics, D_i converges in probability to its expectation as $K \rightarrow \infty$:

$$D_i \xrightarrow{p} 2\sigma_i^2. \quad (16)$$

Since $f(x) = 2x$ is a strictly increasing function for $x \geq 0$, D_i serves as a consistent, strictly increasing proxy for the spurious variance σ_i^2 , justifying its use in the IWD weighting function. $\square$

A.5. Complete Proof of Theorem 1

We prove the four steps in full.

Proof of Theorem 1. Step 1: Reduce $\rho^2(W)$ to a ratio of weighted sums.

Under (A3),

$$\|g_i^S\|^2 = s_i^{(e_i)\top} J_i J_i^\top s_i^{(e_i)} = \Phi_i \|s_i^{(e_i)}\|^2, \quad \|g_i^C\|^2 = \Phi_i c_i^2, \quad (17)$$

so that, taking $\mathbb{E}_e[\|s_i^{(e)}\|^2] = \sigma_i^2$ from (A1),

$$\rho^2(W) = \frac{\sum_i W_i \Phi_i \sigma_i^2}{\sum_i W_i \Phi_i c_i^2}. \quad (18)$$

Step 2: Bound the numerator via the Chebyshev covariance inequality.

Since W_i^{IWD} is a deterministic decreasing function of σ_i^2 and $\Phi_i \perp \sigma_i^2$ by (A3), for large N the Law of Large Numbers gives

$$\sum_i W_i^{\text{IWD}} \Phi_i \sigma_i^2 \xrightarrow{P} \mathbb{E}[\Phi] \cdot \mathbb{E}[a(\sigma^2) \sigma^2],$$

where $a(\sigma^2) := \lim_{N \rightarrow \infty} N W^{\text{IWD}}(\sigma^2)$. (19)

where the factorisation uses $\Phi_i \perp \sigma_i^2$. Because $a(\cdot)$ is strictly decreasing, the Chebyshev (covariance) inequality gives $\mathbb{E}[a(\sigma^2) \sigma^2] \leq \mathbb{E}[a(\sigma^2)] \mathbb{E}[\sigma^2] = \mathbb{E}[\sigma^2]$ (using $\mathbb{E}[a(\sigma^2)] = 1$ from normalisation). Hence the IWD numerator is asymptotically bounded by the uniform numerator $\mathbb{E}[\Phi] \mathbb{E}[\sigma^2]$ (which also uses $\Phi_i \perp \sigma_i^2$).

Step 3: Preserve the denominator.

By (A2), $\sigma_i^2 \perp (c_i^2, \Phi_i)$, so the weights are independent of the products $\Phi_i c_i^2$ and

$$\sum_i W_i^{\text{IWD}} \Phi_i c_i^2 \xrightarrow{P} \mathbb{E}[\Phi c^2] \xleftarrow{P} \sum_i W_i^{\text{unif}} \Phi_i c_i^2. \quad (20)$$

Step 4: Combine to obtain the ratio bound.

From Step 3, for large N , the denominators converge to the same constant $\mathbb{E}[\Phi c^2]$. Substituting this into the S2C ratio from (18) and applying the asymptotic numerator bound from Step 2:

$$\rho^2(W^{\text{IWD}}) \xrightarrow{P} \frac{\mathbb{E}[\Phi] \mathbb{E}[a(\sigma^2) \sigma^2]}{\mathbb{E}[\Phi c^2]} \leq \frac{\mathbb{E}[\Phi] \mathbb{E}[\sigma^2]}{\mathbb{E}[\Phi c^2]} \xleftarrow{P} \rho^2(W^{\text{unif}}). \quad (21)$$

Equality condition. Equality in the Chebyshev covariance inequality holds if and only if σ_i^2 is constant across all training samples (i.e. $\sigma_i^2 = \bar{\sigma}^2$ for all i), in which case $a(\sigma^2) \equiv 1$. In that case, both weighting schemes are asymptotically equivalent and $\rho^2(W^{\text{IWD}}) \xrightarrow{P} \rho^2(W^{\text{unif}})$. $\square$

Remark (Homogeneous case). If $\Phi_i \equiv \Phi$ for all i , everything above reduces to the original finite- N argument (the Chebyshev sum inequality applies, so no Law of Large Numbers is needed for the numerator bound). We present this homogeneous case as a special case in Appendix A.3.

B. Experimental Details

B.1. Dataset Details

Natural-language inference (MNLI). We train on the MultiNLI matched split (Williams, Nangia, and Bowman 2018) (393k premise-hypothesis pairs, three-way labels) and evaluate ID on its matched validation set. For OOD evaluation, we use HANS (McCoy, Pavlick, and Linzen 2019) (30k samples) to test reliance on syntactic shortcuts (lexical overlap, subsequence, and constituent heuristics), and SNLI (Bowman et al. 2015) (570k image caption samples) to evaluate generalization across a linguistic register shift.

Sentiment Classification (SST-2). We train on SST-2 training subsets (Socher et al. 2013) and evaluate ID on the development set. OOD evaluation uses four binary sentiment benchmarks targeting distinct domain and register shifts: SemEval-2017 Task 4A (Rosenthal, Farra, and Nakov 2017) (social media), Customer Reviews (CR) (Hu and Liu 2004) (consumer products), Yelp Polarity (Zhang, Zhao, and LeCun 2015) (business reviews), and Financial PhraseBank (FPB) (Malo et al. 2014) (financial news).

Extractive Question Answering (SQuAD v2). We train on SQuAD v2 subsets (Rajpurkar, Jia, and Liang 2018) (including unanswerable questions) and evaluate ID on the standard development split. OOD performance is evaluated on NewsQA (Trischler et al. 2017) for a news-domain shift, and Natural Questions (NQ) (Kwiatkowski et al. 2019) to test generalization to real Google search queries paired with Wikipedia articles.

Named-entity recognition (CoNLL-2003). Models are trained on English CoNLL-2003 (Tjong Kim Sang and De Meulder 2003) (4 entity types: PER, ORG, LOC, MISC) and evaluated ID on the test set. OOD evaluation uses WNUT-17 (Derczynski et al. 2017) (register and entity-distribution shifts) and OntoNotes 5.0 (Weischedel et al. 2013; Pradhan et al. 2013) (genre and label-schema shifts, projecting its 18-type schema down to PER/ORG/LOC). All setups report entity-level micro-F1.

B.2. Model Details

DeBERTa family. The student is `deberta-v3-xsmall` (He et al. 2021) (22M parameters). Teachers are task-specific `deberta-v3-small` (44M parameters) models fine-tuned on each task’s full training set and selected by ID validation performance. We use the sequence-classification head for MNLI and SST-2, the question-answering head for SQuAD v2, and the token-classification head for NER. Student and teacher share the DeBERTa-v3 tokenizer with a maximum sequence length of 256 tokens for classification and NER, and 384 for SQuAD v2.

Qwen family. The student is `Qwen2.5-0.5B` (Yang et al. 2024) and the teacher is `Qwen2.5-1.5B`. We use the sequence-classification head for MNLI and SST-2, the question-answering head for SQuAD v2, and the token-classification head for NER. Student and teacher share the Qwen2.5 tokenizer with a maximum sequence length of 256 tokens for classification and NER, and 384 for SQuAD v2.

B.3. Baseline Implementation Details

All baselines use the same student architecture, tokenizer, training schedule, and hardware as `INV_KD`. Only the distillation objective differs.

Vanilla KD (Hinton, Vinyals, and Dean 2015). The student minimizes a weighted sum of the ground-truth cross-entropy loss and the forward KL divergence between temperature-scaled teacher and student outputs:

$$\mathcal{L}_{\text{KD}} = (1 - \alpha) \mathcal{L}_{\text{CE}} + \alpha T^2 \text{KL}(\sigma(z_T/T) \parallel \sigma(z_S/T)), \quad (22)$$

Task	Training	ID Eval	OOD Sets	Metric
NLI (MNLI)	MNLI train	MNLI-m dev	HANS, SNLI	Accuracy
Sentiment (SST-2)	SST-2 train	SST-2 dev	Yelp-Polarity, CR	Accuracy
QA (SQuAD v2)	SQuAD v2 train	SQuAD v2 dev	NewsQA, Natural Questions	F1
NER (CoNLL-2003)	CoNLL-2003 train	CoNLL-2003 test	WNUT-17, OntoNotes 5.0	Entity-F1

Table 4: Evaluation datasets. ID sets share the training distribution; OOD sets target specific distribution shifts. Metrics follow standard task practice.

Task	Environment	Text
MNLI	<i>Original</i> (label = entailment)	P: “ <i>The children were playing soccer in the field.</i> ” H: “ <i>Kids were engaged in a sport.</i> ”
	negation	P: “. . . <i>in the field, and no adult was refereeing.</i> ” H: “ <i>Young people participated in athletic activity.</i> ”
	low_overlap	H: “ <i>It was a sport in which the kids were engaged.</i> ”
	syntactic_shortcut_break	
SST-2	<i>Original</i> (label = positive)	“ <i>The film was absolutely wonderful.</i> ”
	negation_invariant	“ <i>The film was not disappointing at all.</i> ”
	social_media_register	“ <i>tbh the film was absolutely wonderful [music] #mustwatch</i> ”
	domain_transposition	“ <i>The product was absolutely excellent.</i> ”
NER	<i>Original</i>	[Barack Obama] _{PER} visited [Paris] _{LOC} .
	lowercase_transform	[barack obama] _{PER} visited [paris] _{LOC} .
	entity_substitution_same_type	[Amina Watanabe] _{PER} visited [Zurich] _{LOC} .
	social_media_register	omg [barack obama] _{PER} pulled up to [paris] _{LOC} [music] #travel

Table 5: Illustrative examples of counterfactual environment generation across three tasks. Each row shows one environment’s rewrite, with the modified span emphasised. Bold text in MNLI/SST-2 shows the injected transformation; NER entities are shown in brackets with their type.

where z_T and z_S are teacher and student logits, T is the temperature, and α is the KD weight. For SQuAD v2, the KL is computed separately for start and end logits. Training uses the original (non-augmented) data.

RevKL (Gu et al. 2024). We replace the forward KL with reverse KL, $\text{KL}(p_S||p_T)$, adapting the objective to classification, extractive QA, and token classification. Training uses the original (non-augmented) data.

DKD (Zhao et al. 2022). Decoupled KD separates the KD loss into target-class (TCKD) and non-target-class (NCKD) terms:

$$\mathcal{L}_{\text{DKD}} = \beta_t \mathcal{L}_{\text{TCKD}} + \beta_{\text{nt}} \mathcal{L}_{\text{NCKD}}. \quad (23)$$

We use the original settings ($\beta_t = 1.0$, $\beta_{\text{nt}} = 1.0$) and adapt the objective to all tasks.

LWD. Layer-Wise Distillation augments output-level KD with MSE matching between corresponding hidden representations. For DeBERTa, we align the first, middle, and last encoder layers; for Qwen, the first, third, and last decoder layers.

AugKD. Vanilla KD trained on the same augmented dataset as IWD, treating all K paraphrases as independent training samples with equal weight. This isolates the effect of invariance weighting from data augmentation alone.

B.4. Training and Hyperparameter Details

Optimiser and schedule. All students use AdamW (Loshchilov and Hutter 2019) ($\beta_1 = 0.9$,

$\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.0) with 200 warmup steps and linear decay. DeBERTa students train for 3 epochs (5 for NER) with learning rate 7×10^{-5} , batch size 32, and fp16 precision. Qwen students train for 3 epochs (5 for NER) with learning rate 2×10^{-5} , batch size 8, gradient accumulation of 2 (effective batch size 16), gradient checkpointing, and bf16 precision.

KD hyperparameters. All methods use temperature $T = 2.0$. The KD weight α is 5 for DeBERTa and 1 for Qwen, selected from $\{0.5, 1, 2, 5\}$ on MNLI ID validation and fixed for all tasks.

IWD hyperparameters. We use $\tau = 1.0$ with batch-mean weight normalization ($\sum_{i \in \mathcal{B}} \tilde{w}_i = |\mathcal{B}|$). The disagreement metric is Jensen–Shannon divergence for MNLI and SST-2, one-minus SQuAD-F1 for SQuAD v2, and one-minus entity-set F1 for NER. The results in the major result table 1 for NLI, NER and Extractive QA tasks were achieved by these hyperparameters while for better results in sentiment analysis task we had to use $\tau = 0.1$ with SymKL as the distance measure of teacher invariance.

Per-task disagreement metric D_i . D_i averages pairwise disagreement over the $\binom{K}{2}$ variant pairs. For MNLI and SST-2, we compute Jensen–Shannon divergence over teacher class probabilities (3-class and 2-class softmax, respectively). For SQuAD v2 and NER, we use one-minus F1 between teacher-decoded answer strings and entity mention sets, respectively.

Task / Dataset	$\tau = 0.1$	$\tau = 0.3$	$\tau = 1.0$	$\tau = 3.0$	$\tau = 10.0$	$\tau = 30.0$
NLI (DeBERTa-v3, 5k)						
MNLI (ID)	81.32 $\pm$ 0.27	82.30 $\pm$ 0.33	82.58 $\pm$ 0.31	82.63 $\pm$ 0.42	82.57 $\pm$ 0.48	82.55 $\pm$ 0.50
HANS (OOD-1)	52.93 $\pm$ 1.74	53.56 $\pm$ 0.60	53.36 $\pm$ 1.35	54.03 $\pm$ 2.28	53.98 $\pm$ 1.78	53.91 $\pm$ 1.61
SNLI (OOD-2)	76.39 $\pm$ 0.87	78.21 $\pm$ 0.51	78.61 $\pm$ 0.29	78.52 $\pm$ 0.55	78.76 $\pm$ 0.08	78.84 $\pm$ 0.20
Sentiment Analysis (DeBERTa-v3, 5k)						
SST-2 (ID)	92.20 $\pm$ 0.23	92.13 $\pm$ 0.35	92.28 $\pm$ 0.24	92.05 $\pm$ 0.33	92.09 $\pm$ 0.23	92.13 $\pm$ 0.07
SemEval (OOD-1)	88.16 $\pm$ 0.36	88.48 $\pm$ 0.21	88.45 $\pm$ 0.46	88.63 $\pm$ 0.32	88.62 $\pm$ 0.45	88.61 $\pm$ 0.51
CR (OOD-2)	88.21 $\pm$ 0.93	88.30 $\pm$ 0.96	88.21 $\pm$ 0.77	87.77 $\pm$ 0.96	87.85 $\pm$ 0.93	87.94 $\pm$ 1.34
Yelp-Polarity (OOD-3)	91.60 $\pm$ 0.56	91.67 $\pm$ 0.43	91.70 $\pm$ 0.49	91.43 $\pm$ 0.53	91.43 $\pm$ 0.52	91.51 $\pm$ 0.55
FPB (OOD-4)	92.65 $\pm$ 0.16	93.05 $\pm$ 0.66	92.78 $\pm$ 0.67	92.68 $\pm$ 0.75	93.24 $\pm$ 0.43	92.93 $\pm$ 0.36
Extractive QA (DeBERTa-v3, 5k)						
SQuAD-v2 (ID)	47.64 $\pm$ 2.72	47.14 $\pm$ 2.60	47.51 $\pm$ 1.27	46.75 $\pm$ 0.75	45.80 $\pm$ 0.87	45.36 $\pm$ 0.74
NewsQA (OOD-1)	29.20 $\pm$ 2.00	27.56 $\pm$ 1.97	24.82 $\pm$ 0.41	23.00 $\pm$ 1.61	22.95 $\pm$ 0.91	21.57 $\pm$ 1.03
Natural Questions (OOD-2)	45.32 $\pm$ 1.47	42.99 $\pm$ 2.68	35.87 $\pm$ 0.29	33.37 $\pm$ 2.19	38.32 $\pm$ 1.87	35.49 $\pm$ 2.47
NER (DeBERTa-v3, 5k)						
CoNLL-2003 (ID)	91.42 $\pm$ 0.08	91.72 $\pm$ 0.23	91.62 $\pm$ 0.22	91.50 $\pm$ 0.49	91.06 $\pm$ 0.24	91.05 $\pm$ 0.25
WNUT-17 (OOD-1)	45.54 $\pm$ 0.65	45.89 $\pm$ 0.18	46.04 $\pm$ 1.11	45.26 $\pm$ 2.19	43.21 $\pm$ 0.84	43.25 $\pm$ 0.76
OntoNotes 5.0 (OOD-2)	78.51 $\pm$ 0.25	78.60 $\pm$ 0.26	78.56 $\pm$ 0.30	78.13 $\pm$ 0.93	77.01 $\pm$ 0.64	77.00 $\pm$ 0.58

Table 6: Sensitivity analysis evaluating the impact of the distillation temperature parameter τ on IWD across four task families using DeBERTa-v3 (trained on 5k subsamples).

Counterfactual data generation. Variants are generated offline using GPT-4.1-mini with prompts preserving task labels (sentiment, entailment, answer spans, or entity annotations) while varying superficial features such as register, syntax, and vocabulary. For NER, entity surface forms are preserved while context is paraphrased. Each sample has $K \in [2, 6]$ variants (mean ≈ 5); samples with $K < 2$ or label-inconsistent variants are removed.

Hardware and seeds. Experiments use 2 NVIDIA RTX A4500 GPUs (20 GB each). The Qwen SQuAD v2 teacher uses batch size 4 with gradient accumulation of 8. Each configuration is evaluated over five seeds {42, 43, 44, 45, 46}, reporting mean and standard deviation.

B.5. Counterfactual Environment Types

This subsection lists the paraphrase environments used to generate counterfactual variants. Each environment targets a distinct spurious surface feature while preserving labels. All environments are applied to training samples, and label-inconsistent variants are discarded. The number of variants varies across experiments. Table 5 shows several examples of environments data generation.

MNLI (7 environments). We generate variants by perturbing lexical overlap (`high/low_overlap`), negation patterns (`negation`), irrelevant context (`distractor`), syntactic heuristics (`syntactic_shortcut_break`), length bias (`length_decorrelation`), and voice structure (`voice_alternation`) while preserving labels.

SST-2 (8 environments). Variants perturb sentiment-related shortcuts through negation invariance, lexical overlap (`lexical_paraphrase_low_overlap`), domain and register shifts (`domain_transposition`,

`social_media_register`), sentiment cue removal, intensifier/hedge changes, and comparative reframing while preserving polarity.

SQuAD v2 (6 environments). Variants target QA shortcuts through distractor insertion, question paraphrasing, answer-position shifts, decoy answer spans, entity substitution, and voice alternation while preserving the gold answer.

NER (7 environments). Variants include 3 programmatic transforms (lowercasing, typographical noise, and punctuation perturbation) and four LLM-based transforms targeting entity substitution, emerging entities, social-media register, and syntax variation. Entity annotations are preserved for all.

C. Ablation Experiments and Results

Sensitivity to Weight Temperature (τ). Table 6 evaluates the weighting temperature $\tau \in \{0.1, 0.3, 1.0, 3.0, 10.0, 30.0\}$, which dictates the sharpness of sample re-weighting relative to teacher divergence (D_i). Classification tasks (MNLI, SST-2) and NER are robust across $\tau \in [0.3, 3.0]$, with ID and mean OOD metrics varying by less than 0.7 and 1.5 points, respectively. In contrast, extractive QA (SQuAD v2) is highly sensitive due to high variance in teacher answer-span fluctuations, benefiting dramatically from aggressive downweighting ($\tau = 0.1$). Lowering τ from 30.0 to 0.1 yields substantial OOD gains of +7.6 F1 on NewsQA (29.20) and +9.8 F1 on Natural Questions (45.32). Thus, while our default setting of $\tau = 1.0$ is optimal for classification and sequence tagging, tasks with high divergence variance like QA benefit from smaller temperatures ($\tau \in \{0.1, 0.3\}$).

Sensitivity to Distance Metric Selection. The optimal teacher invariance distance metric depends directly on the structure of the output space. For classification tasks (MNLI,

Distance Metric	ID	OOD 1	OOD 2
MNLI Benchmark			
	<i>MNLI-m</i>	<i>HANS</i>	<i>SNLI</i>
Jensen–Shannon (D_{JS})	82.60 ± 0.28	53.35 ± 1.38	78.68 ± 0.34
Symmetric KL (D_{SymKL})	81.38 ± 0.40	53.01 ± 2.12	77.43 ± 0.74
L_2 Distance	82.38 ± 0.27	52.76 ± 1.03	78.58 ± 0.33
SST-2 Sentiment Benchmark			
	<i>SST-2</i>	<i>CR</i>	<i>Yelp</i>
Jensen–Shannon (D_{JS})	92.28 ± 0.24	88.21 ± 0.77	91.70 ± 0.49
Symmetric KL (D_{SymKL})	92.01 ± 0.54	87.85 ± 1.63	91.73 ± 0.44
L_2 Distance	92.01 ± 0.37	88.30 ± 1.48	91.55 ± 0.55
Extractive QA Benchmark			
	<i>SQuAD-v2</i>	<i>NewsQA</i>	<i>NQ</i>
Span F1	26.60 ± 0.70	25.21 ± 0.22	37.35 ± 0.96
Jaccard Distance	27.03 ± 1.13	25.36 ± 0.23	37.78 ± 0.29
Levenshtein Distance	27.10 ± 1.51	24.65 ± 0.23	38.19 ± 1.38
NER Benchmark			
	<i>CoNLL-03</i>	<i>WNUT17</i>	<i>OntoNotes</i>
Entity Set F1	91.62 ± 0.22	46.04 ± 1.11	78.56 ± 0.30
Entity Set Jaccard	91.62 ± 0.33	45.80 ± 0.57	78.46 ± 0.37
Entity Flip Rate	91.62 ± 0.33	45.80 ± 0.57	78.46 ± 0.37

Table 7: Sensitivity of IWD to distance metric formulations across tasks on DeBERTa-v3 ($N = 5k$).

Variant Strategy & Metric	MNLI-M (ID)	MNLI-MM (ID)	HANS (OOD1)	SNLI (OOD2)
JS Distance (Primary Metric)				
All Variants	84.238	84.947	57.323	80.542
Original Variant	82.812	83.411	56.206	77.433
Random Variant	82.873	83.441	53.103	78.185
Sym KL Divergence				
All Variants	83.311	84.121	54.420	78.784
Original Variant	82.170	83.197	52.076	75.706
Random Variant	82.292	82.994	52.643	77.027
L_2 Distance				
All Variants	<u>84.014</u>	<u>84.875</u>	55.036	<u>79.617</u>
Original Variant	82.934	83.868	<u>55.626</u>	77.037
Random Variant	82.638	83.390	52.923	78.195

Table 8: Ablation study evaluating environment variant selection strategies during student distillation across different distance metrics using DeBERTa-v3 on ID and OOD benchmarks.

SST-2), bounded and symmetric Jensen–Shannon divergence (D_{JS}) achieves peak performance by preventing extreme target probability tails from overly skewing sample weights compared to Symmetric KL or L_2 distance. For structured prediction, metrics directly targeting output boundary shifts yield superior signals: Levenshtein distance performs best for Extractive QA (38.19 F1 on NQ), while Entity Set F1 achieves top accuracy for NER (46.04 F1 on WNUT17). Aligning the metric to the output modality—probability divergences for classification and structural metrics for span

extraction—maximizes IWD’s ability to isolate non-causal prediction variance.

Environment Variant Utilization. Table 8 compares three student training presentation strategies using teacher divergence weights (D_i) computed over the full set of variants: (1) *All Variants* (distilling on every generated environment state), (2) *Original Variant* (distilling on clean inputs only), and (3) *Random Variant* (selecting a single random variant per sample). Distilling across **All Variants** consistently achieves top performance across all benchmarks. Restricting student inputs to the *Original* or *Random* variant causes substantial drops (e.g., losing up to 4.2 percentage points on HANS). Evaluating across multiple distance measures (JS , Symmetric KL , L_2) confirms this trend is metric-agnostic, with **JS Distance** providing the most effective weighting signal overall.

Weight Normalisation. We compare default batch-mean normalisation ($\tilde{w}_i = |\mathcal{B}| \cdot w_i / \sum_{j \in \mathcal{B}} w_j$) against unnormalised weighting across MNLI and SST-2 (Table 9). Performance is effectively equivalent: on MNLI, ID accuracy shifts by < 0.17 points and mean OOD performance differs by ≤ 0.07 points across DeBERTa-v3 and Qwen-2.5. On SST-2, unnormalised weighting shows minor average OOD gains (+0.63 on DeBERTa-v3, +0.41 on Qwen-2.5) but fluctuates across individual datasets (e.g., +1.32 on FPB vs. -0.22 on SemEval-2017 for DeBERTa-v3), reflecting single-seed variance. We retain batch-mean normalisation because it: (i) decouples the effective learning rate from the absolute scale of D_i , stabilizing τ tuning across tasks; and (ii) strictly preserves the per-sample relative importance weights required by our theoretical bound.

Backbone	Norm	ID	HANS (OOD1)	SNLI (OOD2)
Task: MNLI				
DeBERTa-v3	Batch-Mean	83.25	51.98	79.66
	None	83.22	51.80	79.91
	Δ	-0.03	-0.18	+0.25
Qwen-2.5	Batch-Mean	81.87	53.19	78.39
	None	81.70	53.20	78.23
	Δ	-0.17	+0.02	-0.16

Backbone	Norm	ID	SemEval	CR	Yelp	FPB
Task: SST-2						
DeBERTa-v3	Batch-Mean	91.40	89.27	85.64	91.38	90.19
	None	91.74	89.05	86.70	91.74	91.51
	Δ	+0.34	-0.22	+1.06	+0.36	+1.32
Qwen-2.5	Batch-Mean	92.89	74.19	73.40	78.45	83.58
	None	93.00	73.91	72.34	79.13	85.87
	Δ	+0.11	-0.28	-1.06	+0.68	+2.29

Table 9: Ablation study comparing batch-mean normalization against no weight normalization on MNLI and SST-2 across DeBERTa-v3 and Qwen-2.5 backbones.

Task	w_i (Mean $\pm$ Std)	% $w_i \geq 0.9$	% $w_i < 0.5$	Effective N
MNLI	0.87 ± 0.10	40.8%	0.0%	4,931 (98.6%)
SST-2	0.95 ± 0.08	78.0%	0.0%	4,966 (99.3%)
SQuAD v2	0.86 ± 0.20	63.5%	4.4%	4,748 (95.0%)
NER	0.48 ± 0.06	0.0%	58.7%	4,931 (98.6%)

Table 10: Distribution of per-sample IWD weights across training data at $\tau = 1.0$ and batch-mean normalization ($N = 5,000$ per task). Effective sample size $(\sum_i w_i)^2 / \sum_i w_i^2$ measures information retention relative to uniform weighting, where values near N indicate gentle re-weighting rather than aggressive filtering.

Weight Distribution Across Tasks. Table 10 reports IWD weight statistics across $N = 5,000$ samples at $\tau = 1.0$. High effective sample sizes across all tasks (95.0%–99.3%) confirm that IWD performs smooth continuous re-weighting rather than aggressive sample filtering. Weight profiles naturally adapt to task complexity: classification tasks (SST-2, MNLI) retain high mean weights (0.95 and 0.87), whereas SQuAD v2 exhibits higher variance (± 0.20) and 4.4% low-weight samples ($w_i < 0.5$) due to teacher sensitivity on complex extractive spans. On NER, higher overall perturbation sensitivity yields a lower mean weight (0.48 ± 0.06), yet a 98.6% effective sample size confirms that relative importance ordering is preserved while maintaining overall data capacity.