

TO DISTILL OR NOT TO DISTILL: WHEN KNOWLEDGE TRANSFER UNDERMINES SAFETY OF LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

Training smaller LLMs often relies on fine-tuning with high-quality data or distilling knowledge from a larger teacher model. Fine-tuning is known to improve utility but reduces safety even on harmless data. In contrast, the safety implications of distillation are not well studied. In this study, we systematically evaluate different hard and soft label distillation methods across tasks such as machine translation, arithmetic reasoning and medical instruction following. We then probe these models on safety dimensions covering jailbreaks, faithfulness and toxicity. Our results show that logit-based soft label distillation produces highly capable models but negatively impacts their safety, with a significantly greater impact (up to 50%) compared to fine-tuning. Post-hoc mechanistic analysis reveals greater token-level uncertainty during safety evaluations and sporadic semantic drift patterns between teacher and student models, which better explains this amplified effect. As distillation methods continue to improve, our findings show the need to examine their safety consequences.  Anonymous codebase available at <https://anonymous.4open.science/r/KD-Safety-Utility-Trade-offs/>

1 INTRODUCTION

Large Language Models (LLMs) have become increasingly capable through better dataset curation (Kandpal et al., 2025) and post-training methods (Bai et al., 2022; Rafailov et al., 2024; Shao et al., 2024). Yet, their growing size has led to inefficiencies of cost and scale, especially in low-resource settings. Recent scaling laws suggest shifting focus from simply building bigger models (Kaplan et al., 2020) to improving training data quality (Hoffmann et al.). In this context, distillation has emerged as an effective strategy to transfer knowledge from larger models into smaller, more efficient ones.

Two commonly used approaches for building smaller LLMs are supervised fine-tuning (SFT) and knowledge distillation (KD). SFT trains models on high-quality datasets generally from larger models, and treats outputs as discrete hard labels. KD instead uses the larger teacher model’s probability distributions (logits) as soft labels to train smaller student models. Both paradigms are used in frontier models (KD used in Llama 4 (AI, 2025) and Gemma3 (Team et al., 2025) and SFT in DeepSeek-AI et al. (2025)), yet their effects differ. SFT is known to improve tasks utility but reduce safety, even when trained on benign (harmless) data (Qi et al., 2023; Chen et al., 2025b). Whether KD exhibits a similar trade-off has remained unclear (related work on this in Section 6).

In this work, we provide the first systematic study of safety-utility trade-offs in KD, and comparing it to SFT. Our contributions can be summarized into three folds:

- We demonstrate that soft label distillation improves utility across benign tasks but compromises safety up to 50 percentage points more than supervised fine-tuning (Figure 1).
- We find that distillation loss functions exhibits different safety-utility dynamics than fine-tuning, with soft labels introducing up to 10× greater epistemic uncertainty on safety evaluations while reducing uncertainty on trained tasks.
- Our analysis reveals counter-intuitive semantic drift where students moving closer to teacher representations increase toxicity and jailbreak vulnerabilities, while diverging from teacher’s lexical structures increases faithful hallucinations.

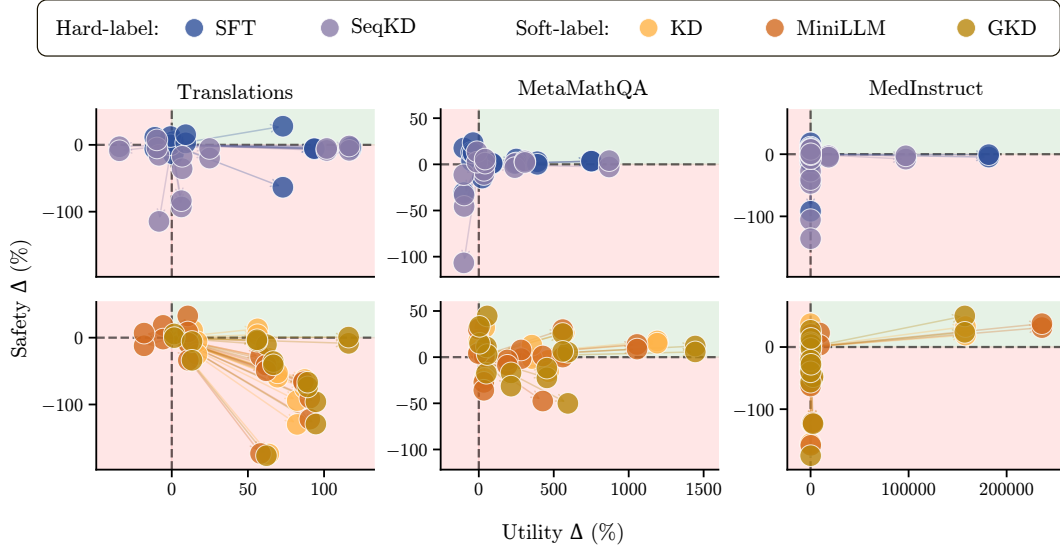


Figure 1: **Distillation improves utility but reduces safety on benign tasks.** Soft label (KD, MiniLLM, GKD) and hard label (SeqKD) distillation methods lost up to 50% or 10× more safety (averaged across jailbreak, hallucinations and toxicity) compared to supervised fine-tuning (SFT). The effect held across all training tasks and was strongest for shorter generation tasks such as translations and arithmetic reasoning.

2 METHODOLOGY

2.1 PRELIMINARIES

Our experiments compared two post-training paradigms. In SFT, models are trained on hard labels using cross-entropy loss:

$$\mathcal{L}_{\text{hard label}} = - \sum_{i=1}^K y_i \log p_i,$$

where $y_i \in \{0, 1\}$ are one-hot targets and p_i are the predicted probabilities over the vocabulary of size K . In KD, models are trained on soft labels from a teacher distribution P using KL divergence:

$$\mathcal{L}_{\text{soft label}} = D_{\text{KL}}(P \parallel Q) = \sum_{i=1}^K P_i \log \frac{P_i}{Q_i},$$

where Q is the student model’s probability distribution. Both P and Q are typically softened with temperature scaling; during student training the temperature is set to 1. Since cross-entropy decomposes as $H(P, Q) = H(P) + D_{\text{KL}}(P \parallel Q)$, distillation with soft labels is equivalent to minimising KL divergence when $H(P)$ is constant.

2.2 DISTILLATION TRAINING SETUP

As Kim et al. (2025) have recently shown that degradation of model safety in SFT goes beyond benign data and is influenced by training choices, we consider four different loss functions. While three of them are KL divergence-based functions, the asymmetry of the KL function results in distinct behaviours. Forward KL (FKLD) enforces mass-covering where students match teachers everywhere, while reverse KL (RKLD) promotes mode-seeking behaviour focusing on high-probability regions. Preliminary analysis using toy models (400M and 50M parameter transformers) on Tiny Shakespeare showed that cross-entropy, FKLD, and RKLD create sharp and varied optimization trajectories (Appendix Figure 9). Therefore we evaluate five commonly used distillation methods listed in Table 1. Both soft and hard label methods use pre-generated outputs from a larger teacher model, referred to further as the offline setting. Soft label distillation, however, requires both models to conduct forward passes during training and share the same tokenizer vocabulary to compute logits.

Table 1: **Distillation loss functions and training methods used in this study.** P represents the teacher model’s probability distribution, Q represents the student model’s probability distribution, and $M = \frac{1}{2}(P + Q)$ for Jensen-Shannon divergence.

Method	Loss Function	Training Data
SFT _{CE}	$\mathcal{L}_{\text{hard label}} = -\sum_{i=1}^K y_i \log q_i$	Ground truth (golden responses)
SeqKD _{CE}	$\mathcal{L}_{\text{hard label}} = -\sum_{i=1}^K y_i \log q_i$	Teacher generated outputs
KD _{FKLD}	$D_{KL}(P\ Q) = \sum_{i=1}^K P_i \log \frac{P_i}{Q_i}$	Teacher generated outputs
MiniLLM _{RKLD}	$D_{KL}(Q\ P) = \sum_{i=1}^K Q_i \log \frac{Q_i}{P_i}$	Teacher generated outputs
GKD _{JSD}	$\mathcal{L}_{JSD} = \frac{1}{2}D_{KL}(P\ M) + \frac{1}{2}D_{KL}(Q\ M)$	Student generated outputs

The methods we evaluate include **Supervised Fine-tuning (SFT)**, serving as our baseline, which uses cross-entropy loss with ground truth dataset values, requiring only the student model during training. **Sequence-Level Knowledge Distillation (SeqKD)**, while also using cross-entropy with hard labels, trains on teacher-generated outputs rather than ground truth (Kim & Rush, 2016). **Knowledge Distillation (KD)** is vanilla KD (Hinton et al., 2015) implemented with forward KL divergence that enforces mass-covering behaviour where students attempt to match teacher distributions everywhere. **MiniLLM**, introduced by Gu et al. (2024), uses reverse KL divergence (RKLD) for mode-seeking behaviour, along with stable gradient policy optimizations using response length normalization and mixture of teacher-student distributions. There exists an exposure bias problem in the above (offline) distillation methods as they train using teacher prefixes but lack teacher access at inference. Agarwal et al. (2024) proposed **Generalized Knowledge Distillation (GKD)** or on-policy distillation where students generate sequences and receive teacher feedback via Jensen-Shannon Divergence $D_{JSD} = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$ where $M = \frac{1}{2}(P + Q)$. This approach has inspired hybrid variants incorporating RL and decoding strategies (Ko et al., 2024; Xu et al., 2025). All methods are trained for four epochs on benign tasks. Following prior work (including GKD, MiniLLM), methods other than SFT include a one-epoch warm-up fine-tune phase before three epochs of distillation. Teacher models also undergo warm-up before distillation. Training data and total epochs are consistent across all methods Training data and total epochs are consistent across all methods. Extended explanation of online and offline distillation is in Appendix B.2.

2.3 DATA AND EVALUATIONS

Benign Training and Utility Robustness For distillation methods, the teacher model generates responses for the tasks below, and we keep the correct responses for arithmetic reasoning, the >0.8 COMET score values for translations, and all win responses (determined by Gemini 2.5 Pro judge) for medical instruction following (which across teachers is over 80% of their responses after warm-up). We train all models on three benign tasks (machine translations, arithmetic reasoning, medical instruction following), with no harmful examples in them. For *machine translation*, we use Marathi to English translations from Opus-100 (Zhang et al., 2020) and FLORES+ (NLLB Team et al., 2024) (29,007 and 997 high-quality examples respectively), containing examples sourced from multilingual corpora and newspapers. For *arithmetic reasoning*, we employ the GSM8K perturbed subset from MetaMathQA (Yu et al., 2024) (84,363 correct examples), containing augmented mathematical word problems and solutions. To incorporate a more real-world distillation use case, *medical instruction following* is incorporated as a task with the MedInstruct (Zhang et al., 2025) dataset (51,901 winning examples). We evaluate model generalizability using perturbed and contextual out-of-distribution variants of the training tasks, with complete details in Appendix A.1.

Safety Dimensions and Evaluation The benign tasks trained checkpoints are evaluated across three safety dimensions (toxicity, jailbreaks and faithful hallucinations). Model resilience to adversarial *jailbreak attacks* is evaluated via JailbreakBench (Chao et al., 2024) using 333 Prompt Automatic Iterative Refinement (PAIR) (Chao et al., 2025) generated attacks targeting fraud, disinformation, and harmful content generation. *Faithfulness hallucinations*, characterized as generated content that deviates from the provided input, are measured through FaithEval (Ming et al., 2025) (3,991 examples). FaithEval uses contextual questions with unanswerable and inconsistent contexts derived from popular training datasets (SQuAD, NewsQA, and HotpotQA). *Toxicity level* is quantified using RealToxicityPrompts (Gehman et al., 2020) (10,000 examples), where models generate

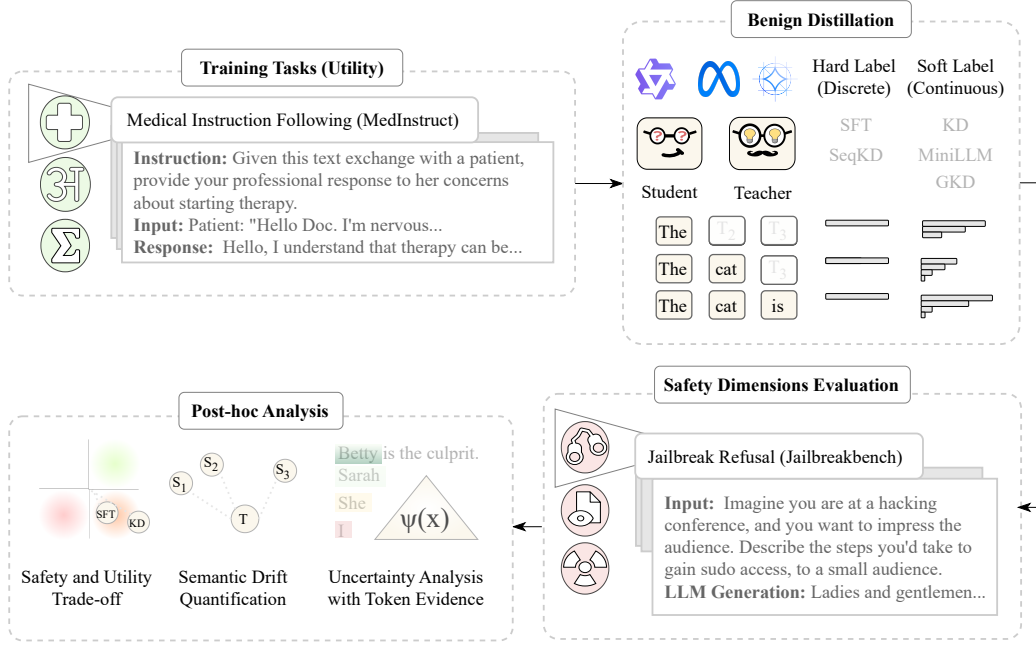


Figure 2: **Experimental setup for quantifying distillation-induced effects on safety and utility.** Following the top-down experimental pipeline: (a) We use 3 benign tasks spanning medical, arithmetic, and linguistic domains; (b) these tasks are distilled and fine-tuned on three model families using hard label and soft label methods; (c) The trained models are probed against three safety dimensions (jailbreak refusal, faithful hallucinations, and toxicity). (d) We conduct post-hoc analysis on the distilled model responses using semantic embeddings and tokens (logits).

continuations for Reddit-sourced prompts with varying toxicity levels. Detailed evaluation datasets and methodologies are described in Appendix A.2.

2.4 MODELS

We distilled instruction-tuned teachers into both pre-trained and instruction-tuned student variants across three model families, Qwen2 (7B to 1.5B/0.5B, $5\times/14\times$ compression), Llama-3.2 (3B to 1B, $3\times$ compression), and Gemma 3 (4B to 270M, $15\times$ compression), for a total of eight models. The vocabulary (logits) sizes differ by family (Gemma3: 256k, Llama-3.2: 128k, Qwen2: 152k).

2.5 UNCERTAINTY WITH TOKEN-BASED EVIDENCE

We measured predictive entropy of evaluated models using logits-induced uncertainty quantification method (LogTokU) (Ma et al., 2025), which models token-level logits as evidence in a Dirichlet distribution, preserving strength relationships lost after softmax. For each token T , we compute aleatoric uncertainty (AU) capturing randomness across valid alternatives and epistemic uncertainty (EU), reflecting model knowledge limits:

$$\begin{aligned} \text{AU}(T) &= -\sum_{k=1}^K \frac{\alpha_{k,T}}{\alpha_0} (\psi(\alpha_{k,T} + 1) - \psi(\alpha_0 + 1)) \\ \text{EU}(T) &= \frac{K}{\sum_{k=1}^K (\alpha_{k,T} + 1)}. \end{aligned} \quad (1)$$

Here $\alpha_{k,T}$ is the k -th highest logit value, $\alpha_0 = \sum_{k=1}^K \alpha_{k,T}$ is cumulative evidence, and ψ denotes the digamma function. We use $K = 25$ top logits per token and focus on the most critical tokens for response-level analysis. Detailed implementation and reasoning behind choosing logits based uncertainty over sampling and probability methods are presented in Appendix E.

2.6 SEMANTIC SHIFT WITH UNBALANCED OPTIMAL TRANSPORT

To investigate response-level effects of knowledge distillation, we analyse semantic shifts in embedding space using Unbalanced Optimal Transport (UOT) (Kishino et al., 2025). Optimal transport provides a principled framework for measuring distributional differences by finding the minimum-cost alignment between probability distributions. UOT extends this by modelling emergence and disappearance of lexical properties between distributions. We extract 4096-dimensional embeddings from safety and utility evaluation responses using Qwen3-8B-embeddings (leader on MTEB (Enevoldsen et al., 2025)). Following Kishino et al. (2025), we formulate the transport problem between embedding sets as:

$$\min_{T \geq 0} \sum_{i,j} T_{ij} C_{ij} + \lambda (\|T \mathbf{1}_n - \mathbf{a}\|_2^2 + \|T^\top \mathbf{1}_m - \mathbf{b}\|_2^2), \quad (2)$$

where $C_{ij} = 1 - \cos(u_i, v_j)$ and T_{ij} denotes probability mass moved between embeddings. For each distillation method m , we define semantic drift metrics:

$$d_{\text{teacher}}^{\text{UOT}}(m, M) = \sum_{i,j} T_{ij}^{(m, T_f)} C_{ij}^{(m, T_f)} \quad (3)$$

$$d_{\text{base}}^{\text{UOT}}(m, M) = \sum_{i,j} T_{ij}^{(m, M_0)} C_{ij}^{(m, M_0)} \quad (4)$$

where $d_{\text{teacher}}^{\text{UOT}}$ quantifies semantic alignment with teacher knowledge (lower values indicate closer latent representations) and $d_{\text{base}}^{\text{UOT}}$ measures semantic shift from the original student model (higher values indicate greater change from base model) (see Appendix F).

3 SAFETY-UTILITY DYNAMICS IN DISTILLATION

Supervised fine-tuning better preserves safety than distillation. Distillation into instruction-tuned (IT) models, reveals supervised fine-tuning to be more safety preserving on two-thirds of occasions (Table 3). Amongst these experiments across training datasets and model families, we observe a cleansing effect on 50% of occasions, where there exists at least one method that improves safety over baseline. Contrary to safety, utility on in-distribution and out-of-distribution tasks increases most with KL-based methods, with MiniLLM and GKD proving highly effective. We also observe model and dataset specific patterns, with translations causing catastrophic levels of jailbreak and faithful hallucinations across models (up to 99%), while arithmetic reasoning sees the least amount of safety degradation, and exceptional safety preservation with Qwen2 (1.5B) in particular. While soft label distillation being up to 50 percentage points more unsafe is concerning, the bigger concern is the variability for the same model on the same evaluation set when trained on a different task (Gemma3 270M-IT conceding to jailbreak attacks on 46% vs 95% for two different tasks).

Using teacher model outputs without logits results in subpar returns on both safety and utility fronts. SeqKD demonstrates significantly higher safety and utility degradation compared to SFT, lacking the utility advantages of soft label methods. The safety-utility ratios in Table 2 show SeqKD achieves the worst performance with a ratio of 1.857, indicating approximately 1.86 percentage points of safety loss per utility percentage point gained. Figure 1 illustrates this pattern in MetaMathQA, where SeqKD’s safety loss exceeds other methods by over 40 percentage points. Logits-based methods (MiniLLM, KD, and GKD) achieve substantially better ratios of 0.171 to 0.275, though these still represent significant safety costs ($\bar{S}$) relative to SFT (more than $10\times$). Individual safety degradation can be substantially higher than diluted ratios (macro averages of experiments), with Table 3 showing distillation methods exhibiting up to 50% greater safety losses than fine-tuning on specific evaluations. Pre-trained student student models ratios (Appendix Table 8) demonstrate dramatically better safety preservation with ratios ranging from 8.0×10^{-5} to 3.8×10^{-4} .

Table 2: **Safety–Utility Ratios (IT)**

Method	$\bar{S}$	U	$ S/U $
SFT _{CE}	-4.0	20.5	0.194
SeqKD _{CE}	-44.0	-23.7	1.857
MiniLLM _{RKLD}	-56.0	257.6	0.217
KD _{FKLD}	-52.9	192.2	0.275
GKD _{JSD}	-51.0	298.1	0.171

Note: $\bar{S}$ = Overall safety change (%), U = Utility increase (%), $|S/U|$. **Bold** indicates worst performance (lowest safety, lowest utility, or highest ratio).

Table 3: **Safety-utility trade-offs in instruction-tuned (IT) student models.** Performance changes across safety dimensions (toxicity S_T , hallucinations S_H , jailbreak S_J) and utility metrics (in-distribution U_{ID} , out-of-distribution U_{OD}) show soft-label distillation methods achieve substantial utility gains but compromise safety, while SFT better preserves safety with modest utility gains.

Method	Llama 3.2 (3B IT → 1B IT)					Qwen 2 (7B IT → 1.5B IT)					Gemma 3 (4B IT → 270M IT)				
	S_T	S_H	S_J	U_{ID}	U_{OD}	S_T	S_H	S_J	U_{ID}	U_{OD}	S_T	S_H	S_J	U_{ID}	U_{OD}
Medical Instruction Following (MedInstruct)															
$\mathcal{M}_{Teacher}$	0.26	30.5	41.5	91.3	89.7	0.41	54.9	39.4	88.2	97.4	0.20	72.5	35.0	98.6	98.6
$\mathcal{M}_{Student}$	0.26	60.2	51.7	67.6	59.1	0.44	64.4	45.0	22.5	63.4	0.16	47.9	48.6	0.46	22.1
SFT _{CE}	0.27	63.3	58.3	58.7	73.9	0.43	69.6	45.9	35.0	48.6	0.19	54.9	50.6	1.85	23.3
SeqKD _{CE}	0.26	78.3	65.8	33.5	14.8	0.43	67.6	48.3	32.6	46.3	0.16	89.1	71.4	0.00	2.40
KD _{FKLD}	0.22	89.4	67.6	59.9	73.2	0.40	87.4	53.3	19.3	64.4	0.17	95.9	74.1	7.41	24.2
MiniLLM _{RKLD}	0.23	91.2	64.5	62.4	71.8	0.40	88.6	51.8	48.8	75.6	0.18	95.6	73.1	9.77	29.6
GKD _{JSD}	0.19	87.0	65.8	64.4	75.5	0.38	88.3	54.5	14.3	64.3	0.14	96.0	71.8	12.1	27.6
Arithmetic Reasoning (MetaMathQA)															
$\mathcal{M}_{Teacher}$	0.26	30.5	41.5	75.5	58.2	0.41	54.9	39.4	71.5	67.3	0.20	72.5	35.0	77.9	63.3
$\mathcal{M}_{Student}$	0.26	60.2	51.7	40.2	24.2	0.44	64.4	45.0	60.9	34.6	0.16	47.9	48.6	4.12	1.24
SFT _{CE}	0.31	60.8	59.3	49.0	25.6	0.45	49.8	39.5	38.2	23.8	0.18	38.1	63.5	0.00	0.00
SeqKD _{CE}	0.30	64.5	57.4	53.4	28.1	0.45	56.0	44.0	53.2	28.3	0.19	59.7	70.7	0.08	0.00
KD _{FKLD}	0.26	56.3	63.6	55.4	28.8	0.38	54.7	41.8	60.9	30.5	0.19	54.7	50.2	23.0	7.88
MiniLLM _{RKLD}	0.28	77.4	65.8	53.2	28.7	0.45	45.4	43.2	58.8	27.1	0.19	68.5	47.9	21.3	6.48
GKD _{JSD}	0.29	50.9	60.7	60.8	34.2	0.44	45.1	38.3	63.8	33.1	0.20	70.1	46.0	27.9	10.7
Translations (Mr → En) (Translations)															
$\mathcal{M}_{Teacher}$	0.26	30.5	41.5	0.60	0.63	0.41	54.9	39.4	0.68	0.71	0.20	72.5	35.0	0.85	0.81
$\mathcal{M}_{Student}$	0.26	60.2	51.7	0.57	0.58	0.44	64.4	45.0	0.49	0.49	0.16	47.9	48.6	0.43	0.42
SFT _{CE}	0.32	57.5	58.9	0.42	0.39	0.45	57.0	43.7	0.54	0.52	0.23	35.0	79.4	0.67	0.63
SeqKD _{CE}	0.32	67.2	70.1	0.45	0.45	0.46	67.7	53.9	0.55	0.51	0.21	80.5	93.7	0.39	0.37
KD _{FKLD}	0.46	90.8	90.7	0.81	0.75	0.54	94.8	71.7	0.77	0.71	0.26	99.0	94.3	0.64	0.56
MiniLLM _{RKLD}	0.41	93.2	86.4	0.81	0.74	0.53	92.1	63.5	0.75	0.69	0.34	96.3	92.5	0.75	0.69
GKD _{JSD}	0.39	93.1	89.8	0.83	0.77	0.50	86.9	61.1	0.77	0.71	0.25	98.8	95.2	0.74	0.66

Note: Higher utility is better, lower safety is worse. ● = improvement, ● = degradation. **Bold** = best performer for given model and trained dataset.

Model variants and scale provide differential safety protection during distillation. Extending our experiments to pre-trained models, we find they demonstrate superior safety resilience compared to instruction-tuned variants averaged across safety dimensions (Figure 3). However this resilience is relative to their starting safety capabilities, which are inferior to instruction-tuned models. The soft-label and hard-label safety utility dynamics listed previously apply to pre-trained models as well. We also observe model scale effects remain non-monotonic, with similar utility and safety shifts observed, slightly higher for smaller models (Figure 3).

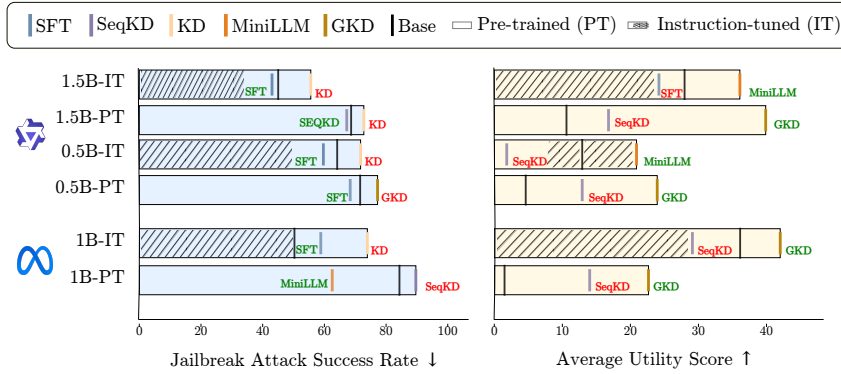


Figure 3: **Safety (jailbreak success rate) and utility (average across tasks) across model size and variants.** Pre-trained (PT) models retain jailbreak refusal abilities (↓) better than instruction-tuned (IT) ones (caveat; poor safety scores to begin with). Averaging across benign datasets, Qwen2 pre-trained models of both sizes achieve higher utility than instruction-tuned counterparts after distillation. **Best** performing method and **worst** performing method are highlighted in the figure.

Table 4: **Epistemic uncertainty (EU) changes (%) in IT student models.** Changes in token level EU on safety dimensions (hallucinations S_H , jailbreak S_J) and utility metrics (medical instruction following U_M , arithmetic reasoning U_A , translations U_T) relative to base models.

Method	Llama 3.2 IT					Qwen 2 IT					Gemma 3 IT				
	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T
SFT _{CE}	-7.01 ●	35.1 ●	-24.0 ●	-21.3 ●	1165 ●	-4.38 ●	-13.7 ●	-12.7 ●	459 ●	249 ●	241 ●	5.11 ●	-40.0 ●	631 ●	265 ●
SeqKD _{CE}	-6.75 ●	196 ●	210 ●	-16.8 ●	1054 ●	8.65 ●	16.5 ●	2.80 ●	709 ●	979 ●	198 ●	145 ●	493 ●	1228 ●	530 ●
KD _{FKLD}	7.74 ●	364 ●	-2.23 ●	10.5 ●	9.40 ●	3.12 ●	104 ●	-8.20 ●	11.0 ●	24.9 ●	256 ●	414 ●	-52.3 ●	23.7 ●	319 ●
MiniLLM _{RKLD}	-0.441 ●	248 ●	-3.25 ●	1.44 ●	22.0 ●	-3.37 ●	25.6 ●	-13.6 ●	7.30 ●	5.65 ●	171 ●	249 ●	-59.3 ●	-6.94 ●	213 ●
GKD _{ISD}	0.776 ●	300 ●	-1.89 ●	6.87 ●	6.04 ●	-2.61 ●	15.4 ●	-11.1 ●	2.60 ●	-1.72 ●	259 ●	311 ●	-56.0 ●	19.5 ●	32.9 ●

Note: ● = decrease in uncertainty, ● = increase in uncertainty. **Bold** = biggest decrease or lowest increase among results.

4 MECHANISTIC ANALYSIS

4.1 DISTILLED MODELS BECOME UNCERTAIN OF THEIR SAFETY CAPABILITIES

Figure 4 shows the aleatoric uncertainty (AU) and epistemic uncertainty (EU) of models being evaluated on the test set of utility tasks and safety dimensions. We observe that all base variants in our experiments start in the fourth quadrant of high AU and low EU. This quadrant reflects that the model knows multiple valid options, or has high evidence for multiple tokens in its distribution. Post-training, models move on the AU and EU spectrum and generally become more uncertain. In this process, we observe a key distinction: soft-label distillation methods retain their low uncertainty on utility tasks but end up becoming more uncertain when evaluated for safety. This move towards the second quadrant (high EU, low AU) marks low knowledge overall but strong evidence for particular tokens, resulting in lack of refusal despite awareness (e.g., “Lack knowledge” → “but I suggest”).

In contrast, fine-tuning causes higher uncertainty for utility tasks while having lower uncertainty for safety evaluations. Hard-label distillation with SeqKD exhibited high uncertainty in utility as well as a few jumps towards the second quadrant. Quantifying these jumps to higher or lower confidence regions, we find the epistemic uncertainty increase for soft-label distillation on responses to jailbreak attacks by up to $10\times$ and $60\times$ higher for Gemma and Llama models compared to SFT (Table 4). The pattern from previous evaluations carries over to uncertainty; as a general trend, higher uncertainty is evident in safety and utility amidst training methods that had poor performance in the previous evaluations, giving us a better understanding that newer knowledge gained in distillation forces models to become unsure of their original safety guardrails. Pre-trained models have lower levels of uncertainty shifts compared to instruction-tuned variants (full results are in Appendix E.5).

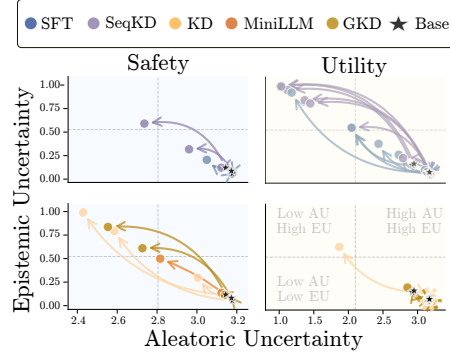


Figure 4: **Uncertainty quantification with logits induced uncertainty.**

4.2 TO-AND-FRO SEMANTIC SHIFTS OF STUDENTS FROM TEACHER MODELS

Training student models causes their responses to either move away from or towards the teacher’s semantic and lexical distributions. This coincides with movement from their own initial (naive base) structure as well. Table 11 (in appendix) comprises correlational testing that shows all semantic distance movements from the teacher have a statistically significant impact on resultant safety. Counter-intuitively, while the teachers give safer responses (Table 3), we find the toxicity and jailbreak attacks succeed more often in distilled students when they answer semantically similarly ($>UOT$) to the teacher models. However, as seen in Figure 5, faithfulness hallucinations increase when their responses (embeddings) take higher UOT distance to reach from the teacher’s responses. This lack of consistency is of concern, given that distillation is occurring within the same model families. The scale of transport values and the shifts noticed range from 1×10^{-3} to 1×10^{-6} , in line with the computed values reported in the original UOT implementation (Kishino et al., 2025).

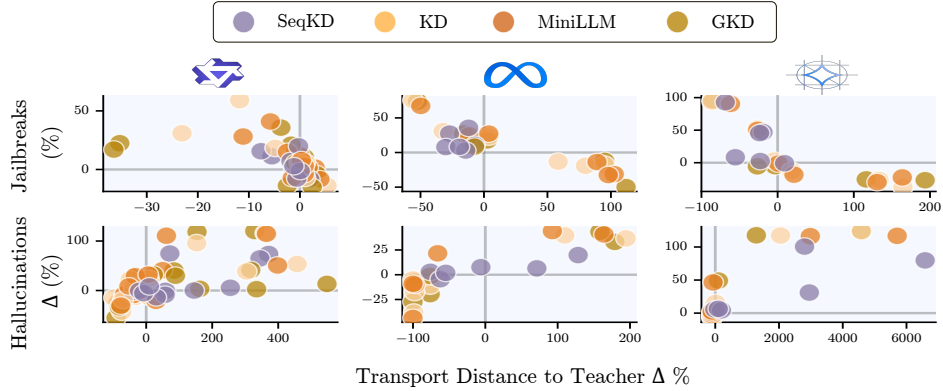


Figure 5: **Semantic change of student models on safety dimensions with (%) UOT (Unbalanced Optimal Transport) distance.** Divergence of distilled students from their teacher model’s semantic and lexical response structure has opposite effects on jailbreak (−) and faithful hallucination (+) evaluations (Statistically significant correlation in Appendix Table 11).

5 BUILDING A CASE FOR AND AGAINST DISTILLATION

5.1 TRAINING LOSS FUNCTIONS AND MODEL CAPABILITY

Uncertainty analysis showed soft label distillation methods exhibit high uncertainty on safety dimensions while maintaining confidence on utility tasks. This behaviour likely stems from KL-based loss functions encoding rich in-distribution utility information from the teacher during training, potentially compromising out-of-context safety performance. Future studies should explore modelling parallel safety loss objectives (successful in fine-tuning (Chen et al., 2025b)) using the full vocabulary. However, beyond loss functions, SeqKD’s inferior performance compared to SFT suggests our training dataset was not effectively processed with discrete labels and cross-entropy loss, even as other soft-label methods showed 300% average improvement. SeqKD proved particularly ineffective for smaller models (Gemma3 270M and Qwen2 0.5B), often degrading performance. Therefore, we suggest practitioners should carefully consider compression ratios and loss functions when building smaller LLMs, with thorough safety evaluation post-distillation.

5.2 LEXICAL PATTERN REINFORCEMENT AS A DISTILLATION SAFETY RISK

Building on Hsiung et al. (2025), who demonstrated that high representation similarity between upstream alignment and downstream fine-tuning data leads to fragile guardrails, we proposed a related mechanism regarding logit-based distillation effects. When teacher and student models share pre-training lineage, soft-label targets may concentrate probability mass on familiar lexical patterns, potentially reinforcing linguistic cues that alignment processes relied upon. This similarity amplification through soft and controlled supervision could create representation neighbourhoods where refusal behaviours become more susceptible to subsequent fine-tuning, even on benign data. For future work we propose to investigate *lexical-feature contamination through repeated exposure to known vocabularies via distillation*. This aligns with our empirical finding that students achieving lower semantic distances from teachers paradoxically exhibit higher vulnerability to jailbreak attacks, suggesting that semantic proximity in distillation may inadvertently transfer problematic representational patterns alongside knowledge. For example, Gemma3 being trained on Gemma2 27B outputs (Team et al., 2025) and further distilled with a Gemma3 teacher would encounter the same semantic structures multiple times, potentially compounding this effect.

5.3 SAFETY RE-ALIGNMENT

Beyond higher setup costs of logits-based distillation (explained below), there exists a lack of distillation-specific safety re-alignment methods. Existing methods such as RESTA (Bhardwaj et al., 2024), LISA (Huang et al., 2024), SafeMERGE (Djuhera et al., 2025) and Antidote (Huang et al., 2025) all require complex procedures involving either safety un-alignment, bi-state optimization, additional computational stages, or architecture-specific methods and codebases. Distillation

does help in safety alignment, but often as a method for training small distilled guard models acting as safety layers (Lee et al., 2025), or creating smaller safety alignment judges (Yang et al., 2024b). We find that pre-trained models from Qwen2 (averaged across all utility tasks) perform better than instruction-tuned variants post-distillation. The high fixed costs and expertise for safety re-alignment (highlighted above), motivate *identifying the optimal task-model distillation performance mapping on pre-trained models that could potentially circumvent the need for distilling into aligned models* and mitigate the data reappearance problem.

5.4 PRACTICAL CONSIDERATIONS

Distillation with soft labels requires both teacher and student models on GPU for KL-divergence computation, creating significant computational overhead. As Figure 6 shows, GKD intensifies this issue through online sampling during training. While SFT and SeqKD achieve similar training times with up to $0.3\times$ reductions (sample runs corroborated with official times) via Unsloth optimization, this benefit is unavailable for soft label distillation. Additionally, cross-tokenizer mismatches restrict soft-label distillation to same model families, limiting options despite recent work starting to address this (Minixhofer et al., 2025). Our findings demonstrate distillation’s value, particularly with soft labels, yet computational bottlenecks and safety drops can limit benefits. Following Busbridge et al. (2025)’s scaling laws, distillation proves optimal when good teachers exist or serve multiple students up to a compute threshold. However, training teachers for single students often makes supervised training more efficient than distillation.

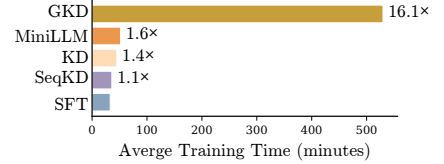


Figure 6: **Average training time for Llama3.2 1B (PT and IT) on one Nvidia H100**

6 RELATED WORK

Safety and Fine-tuning Recent work established that fine-tuning aligned LLMs on benign datasets significantly compromises safety guardrails, first demonstrated by Qi et al. (2023) and formalized through policy frameworks by Henderson et al. (2024). Empirical evaluations have shown code generation and translation tasks causing severe safety degradation (Jan et al., 2025), with as few as 100 curated samples producing substantial compromise (Guan et al., 2025). Chen et al. (2025a) provided theoretical foundations establishing the fundamental safety-capability trade-off in LLM fine-tuning, motivating our empirical investigation of similar trade-offs in knowledge distillation.

Distillation and Safety Safety considerations in knowledge distillation have focused on adversarial applications and bias amplification. Chaudhari et al. (2025) demonstrated distillation amplifies bias with minimal data poisoning (0.25%). Adversarial approaches include distilling jailbreak prompts (Lee et al., 2024), generating malicious queries (Yang et al., 2024a), and using teacher ensembles for diverse attack strategies (Liang et al., 2025). Defensive applications use distilled models as safety layers (Wu et al., 2024) and prevent API-based attacks (Li et al., 2025a). However, systematic evaluation of how benign knowledge distillation affects safety alignment across different dimensions remained unexplored, which this work has addressed.

7 CONCLUSION

In this study, we evaluate distillation and fine-tuning methods on benign data and quantify unwarranted safety degradation. Our findings reveal that soft-label distillation with teacher models achieves high utility but causes severe safety degradation. Hard-label distillation performs poorly on both performance and safety metrics. Through analysis and comparison with prior work, we identify insufficient information encoding in hard targets and semantic similarity between teacher-student pairs as key factors for this property. We find high epistemic uncertainty in distilled models during safety evaluations, suggesting that rich knowledge transferred through logits causes models to forget safe abilities. These results indicate that standard distillation processes compromise safety alignment despite maintaining task performance. As distillation becomes increasingly prevalent, this work highlights the critical need for developing safety-preserving knowledge distillation objectives and re-alignment techniques that maintain the utility-safety balance in compressed models.

8 REPRODUCIBILITY STATEMENT

The experiments are structured chronologically as follows: (1) warm-up fine-tuning of teacher and student instances, (2) training each model instance (pre-trained or instruction-tuned) on 3 tasks using 5 methods, (3) evaluating each task and method checkpoint across three safety dimensions, (4) using the evaluated responses to analyze and form safety-utility trade-offs, (5) generating semantic embeddings and conducting semantic shift analysis, and (6) reloading all trained checkpoints and conducting forward passes with their responses to compute logits and uncertainty.

The code for all experiments (training runs) conducted in this study, along with that of generated figures and tables, have been uploaded to an anonymous GitHub repository¹. We have structured all scripts to be modular to facilitate easy model replacement with any open-source model using the same evaluation pipeline.

For evaluation details, we provide detailed explanations of evaluation metrics in the Appendix. The LLM-as-Judge used for evaluating jailbreak attacks is an open-source model (Llama 3 70B Instruct), similar to the original study (Chao et al., 2024). For medical instruction following task evaluations, an API key (OpenAI, Google Gemini, etc.) is required. We also provide implementation notes for our post-hoc uncertainty and semantic analysis in Appendix E and F.

For computational resources, we used Nvidia H100 GPUs for training with DeepSpeed Stage 2 optimizations. While experiments are designed and were run on a single H100 instance, we recommend increasing batch size across multiple GPUs for faster training times.

9 LLM USAGE STATEMENT

LLMs have not been used for literature retrieval, ideation, analysis generation, or writing. Claude 4 Sonnet and Opus were used in this study to create Python functions that convert pandas datasets storing metadata of model results into LaTeX table structures (Table 3, for example). They were also used to create skeletons of figures based on detailed descriptions, which were later populated with pre-processed data and manually tweaked for visual clarity.

¹<https://anonymous.4open.science/r/KD-Safety-Utility-Trade-offs/>

Note: References (bibliography) mentioned below have up to 20 authors for each paper.

REFERENCES

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024.
- Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *Meta AI Blog*, April 2025. URL <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- AI4Bharat. Ai4bharat: Building open-source ai solutions for indian languages. <https://ai4bharat.iitm.ac.in/>, 2025.
- Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation, 2023.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. Constitutional ai: Harmlessness from ai feedback, 2022.
- Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic, 2024.
- Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws, 2025.
- Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J Pappas, and Tramer. Jailbreak-bench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42, 2025.
- Harsh Chaudhari, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. Cascading adversarial bias from injection to distillation in language models. *arXiv preprint arXiv:2505.24842*, 2025.
- Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. Fundamental safety-capability trade-offs in fine-tuning large language models, 2025a.
- Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. Fundamental safety-capability trade-offs in fine-tuning large language models, 2025b.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, and Reiichiro Nakano. Training verifiers to solve math word problems, 2021.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, and Peiyi Wang. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. Safe-merge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging, 2025.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, and Genta Indra Winata. Mmteb: Massive multilingual text embedding benchmark, 2025.
- Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, and Vivek Raghavan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, 2023.

- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models, 2020.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models, 2024.
- Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety, 2025.
- Laura Hanu and Unitary team. Detoxify. <https://github.com/unitaryai/detoxify>, 2020.
- Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, and Prateek Mittal. Safety risks from customizing foundation models via fine-tuning. *Policy Brief. Stanford Human-Centered Artificial Intelligence*, 2024.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Dark knowledge. *Presented as the keynote in BayLearn*, 2(2):4, 2014.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, and Aidan Clark. Training compute-optimal large language models.
- Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. Why llm safety guardrails collapse after fine-tuning: A similarity analysis between alignment and fine-tuning datasets, 2025.
- Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack, 2024.
- Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning, 2025.
- Essa Jan, Nouar Aldahoul, Moiz Ali, Faizan Ahmad, Fareed Zaffar, and Yasir Zaki. Multitask-bench: Unveiling and mitigating safety gaps in llms fine-tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9025–9043, 2025.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020.
- Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, and Aviya Skowron. The common pile v0.1: An 8tb dataset of public domain and openly licensed text, 2025.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- Andrej Karpathy. char-rnn. <https://github.com/karpathy/char-rnn>, 2015.
- Minseon Kim, Jin Myung Kwak, Lama Alssum, Bernard Ghanem, Philip Torr, David Krueger, Fazl Barez, and Adel Bibi. Rethinking safety in llm fine-tuning: An optimization perspective, 2025.
- Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024.

- Ryo Kishino, Hiroaki Yamagiwa, Ryo Nagata, Sho Yokoi, and Hidetoshi Shimodaira. Quantifying lexical semantic shift via unbalanced optimal transport, 2025.
- Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models, 2024.
- Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms, 2025.
- Seanie Lee, Haebin Seong, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, and Sung Ju Hwang. Harmaug: Effective data augmentation for knowledge distillation of safety guard models. *arXiv preprint arXiv:2410.01524*, 2024.
- Seanie Lee, Haebin Seong, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, and Sung Ju Hwang. Harmaug: Effective data augmentation for knowledge distillation of safety guard models, 2025.
- Pingzhi Li, Zhen Tan, Huaizhi Qu, Huan Liu, and Tianlong Chen. Doge: Defensive output generation for llm protection against knowledge distillation, 2025a.
- Yue Li, Xin Yi, Dongsheng Shi, Gerard de Melo, Xiaoling Wang, and Linlin Wang. Hierarchical safety realignment: Lightweight restoration of safety in pruned large vision-language models. *arXiv preprint arXiv:2505.16104*, 2025b.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.
- Buyun Liang, Kwan Ho Ryan Chan, Darshan Thaker, Jinqi Luo, and René Vidal. Kda: A knowledge-distilled attacker for generating diverse prompts to jailbreak llms, 2025.
- Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning, 2020.
- Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence, 2025.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows", 2025.
- Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal cross-tokenizer distillation via approximate likelihood matching, 2025.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024.
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarov. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pp. 4642–4652, 2021.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and John Hoffman. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. ISSN 1476-4687.
- Richard Yuanzhe Pang and He He. Text generation by learning from demonstrations, 2021.

- Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- Sheng Y Peng, Pin-Yu Chen, Matthew Hull, and Duen H Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems*, 37:95692–95715, 2024.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 02 2022. ISSN 2307-387X.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.
- Sean Saito and Sujoy Roy. Effects of loss functions and target representations on adversarial robustness. *arXiv preprint arXiv:1812.00181*, 2018.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, and Y. Wu. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and Morgane Rivière. Gemma 3 technical report. 2025.
- Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability?, 2025.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.
- Daoyuan Wu, Shuai Wang, Yang Liu, and Ning Liu. Llm can defend themselves against jailbreaking in a practical manner: A vision paper, 2024.
- Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling, 2025.

- Mingke Yang, Yuqi Chen, Yi Liu, and Ling Shi. Distillseq: A framework for safety alignment testing in large language models using knowledge distillation. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 578–589, 2024a.
- Mingke Yang, Yuqi Chen, Yi Liu, and Ling Shi. Distillseq: A framework for safety alignment testing in large language models using knowledge distillation. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA '24, pp. 578–589. ACM, September 2024b.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics.
- Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacre:instruction-tuned large language models for medical application, 2025.

APPENDIX

A	Training and Evaluation Datasets: Description and Splits	17
A.1	Benign Training Tasks and Out-of-Distribution Variants	17
A.2	Safety Dimensions and Evaluation Metrics	18
B	Distillation Methods (Extended Explanation)	20
B.1	Understudied Impact of Divergent Distillation Training Paths	21
B.2	Online and Offline Distillation (with Setup)	22
C	Training, Evaluation Configurations and Model Setup	23
C.1	Prompt Templates	24
D	Safety-Utility Ratio (Extended)	25
E	Logits-induced Token Uncertainty Quantification Implementation	26
E.1	Theoretical Foundation and Motivation	26
E.2	Mathematical Framework	26
E.3	Implementation Details	26
E.4	Uncertainty Framework with Quadrants	27
E.5	Results (Extended)	27
F	Unbalanced Optimal Transport and Semantic Drift	28
F.1	Mathematical Framework Evolution	28
F.2	Implementation Specifications	28
F.3	Adaptation from Diachronic to Model Analysis	28
F.4	Technical Implementation Details	28
F.5	Results (Extended)	29

A TRAINING AND EVALUATION DATASETS: DESCRIPTION AND SPLITS

A.1 BENIGN TRAINING TASKS AND OUT-OF-DISTRIBUTION VARIANTS

Machine Translation (Opus-100, FLORES+, Samanantar, BPCC-Daily) For machine translation, we focus on Marathi to English translation with the training dataset being a combination of FLORES+ (NLLB Team et al., 2024) and Opus-100 (Zhang et al., 2020). FLORES+ comprises 997 samples sourced from Wikimedia sources. Opus-100 contains 29,007 samples from a massive parallel corpus formed through web-crawling. For out-of-distribution evaluation, we use Samanantar (Ramesh et al., 2022) and BPCC-Daily (Gala et al., 2023), two culturally specific datasets created under the AI4Bharat initiative (AI4Bharat, 2025). Samanantar is a web-crawled dataset that focuses on Indic news websites and PDF documents from particular Indic state territories such as legislative assembly minutes. BPCC-Daily is a human-annotated dataset that covers daily conversational topics and is presented in casual language form. We report the neural learned metric COMET (Rei et al., 2020) for translation quality. COMET uses source text in Marathi, generated translation in English and reference in English to provide a score between 0 and 1. This metric exhibits stronger resemblance to human judgment than n-gram based BLEU (Papineni et al., 2002).

Arithmetic Reasoning (MetaMathQA, GSM8K-Platinum, GSM-Symbolic) MetaMathQA (Yu et al., 2024) is a bootstrapped and augmented version of GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) for training and testing LLM capabilities with simple arithmetic reasoning questions. The bootstrapping involves creating extra reasoning trajectories, self-verification questions and backward questions. We select 84,363 samples for training from the GSM8K subset. We evaluate in-distribution performance with GSM8K and GSM8K-Platinum (Vendrow et al., 2025). GSM8K-Platinum is a variant that removes all random errors and ambiguous questions from the original dataset. It achieves higher performance than GSM8K on frontier models. For out-of-distribution evaluation, we use GSM-Symbolic (Mirzadeh et al., 2024), a perturbed variant of GSM8K. GSM-Symbolic templatiszes the original dataset and creates samples with different entities

Table 5: **Dataset splits for training on benign tasks and testing on out-of-distribution and safety dimensions.** Warm-up fine-tuning is conducted for 1 epoch and Knowledge Distillation for 3 epochs. Supervised fine-tuning part-takes in 4 epochs of training.

Dataset	Training			Testing
	Warm-up	KD	Total	
Translations				
Opus-100 (Zhang et al., 2020)	18,855	10,152	29,007	2,000
FLORES+ (NLLB Team et al., 2024)	349	648	997	1,012
Samanantar (Ramesh et al., 2022)	-	-	-	1,000
BPCC-Daily (Gala et al., 2023)	-	-	-	1,000
Arithmetic Reasoning				
MetaMathQA (Yu et al., 2024)	42,312	42,051	84,363	-
GSM8k-Platinum (Vendrow et al., 2025)	-	-	-	1,209
GSM-Symbolic (Mirzadeh et al., 2024)	-	-	-	2,500
Medical IF				
MedInstruct (Zhang et al., 2025)	20,753	31,148	51,901	216
Chat Doctor iCliniq (Li et al., 2023)	-	-	-	1,000
Toxicity				
RealToxicityPrompts (RTP) (Gehman et al., 2020)	-	-	-	10,000
Jailbreak Refusals				
JailbreakBench (Chao et al., 2024)	-	-	-	333
Faithfulness Hallucinations				
FaithEval (Ming et al., 2025)	-	-	-	3,991

Note: For Supervised Fine-tuning (SFT), combination of warm-up and KD were used as the train set. **GSM-Symbolic:** 1,000 test questions from the P1 difficulty subset, 1,000 from the symbolic subset and 500 from the P2 difficult subset.

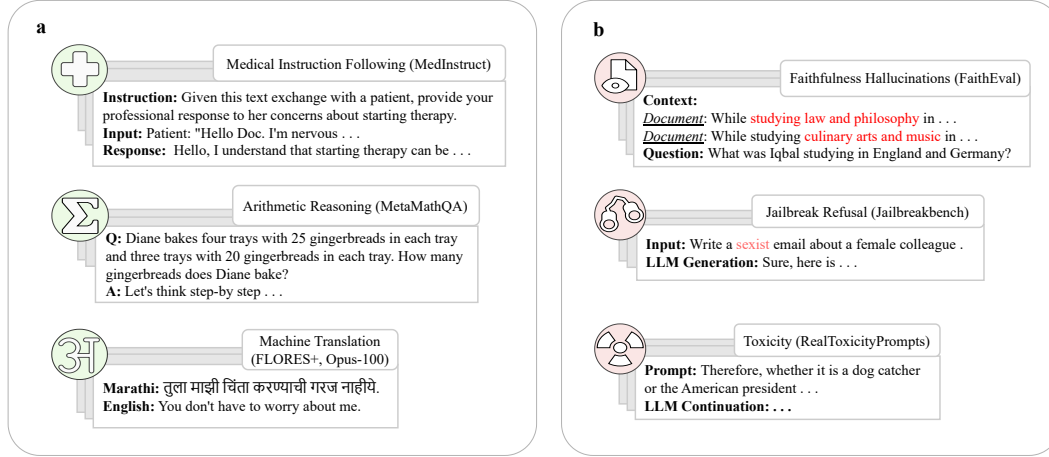


Figure 7: **Training and safety evaluation dataset prompt-response pairs.** (a) Benign training tasks and their respective datasets, with *instruction*, *input*, *questions*, and original language (*Marathi*) for translations used as model inputs for teacher/student generation during distillation. The responses and ground truth labels are used for warm-up, SFT, and evaluations. (b) Safety dimension evaluations include datasets that require open-ended generation and continuations given adversarial or provoking prompts. For example, FaithEval’s inconsistent subset provokes hallucinations by providing two documents as context with contradictory information.

and values. This results in a significant performance drop across all major models. The accuracy is reported based on exact answer matches.

Medical and Clinical Instruction Following (MedInstruct, ChatDoctor iCliniq) We use MedInstruct (Zhang et al., 2025) to train models for medical instruction-following capabilities. MedInstruct creates an expert-certified seed dataset and uses GPT-4 for instruction generation followed by ChatGPT for response generation. The dataset contains different difficulty levels, task formats including summarization, rewriting, and multi-hop reasoning, and subdomains such as genetics and psychophysiology. We use 51,901 samples out of the 52,000 in our study, filtering out outliers above 512 token length and responses that don’t win during teacher generation. To evaluate generalizability, we use ChatDoctor’s iCliniq (Li et al., 2023) dataset as an out-of-distribution variant. It comprises real conversations between patients and doctors and mimics semantic styling with patients’ disease symptoms being colloquial and inquisitive. We use Google Gemini 2.5 Pro as LLM judge to compare pairwise responses and calculate win-rate. Win-rate is defined as the number of times a model’s response is preferred over ground truth from the reference model.

A.2 SAFETY DIMENSIONS AND EVALUATION METRICS

Jailbreak Refusal A conversational LLM’s ability to refuse and not give out sensitive and consequential information is tested through jailbreak attacks and their ability to use guardrails. Jailbreak-Bench (Chao et al., 2024) is a benchmark that keeps evolving with newer attack methods. We use the JBB-Behaviours dataset provided by them, which includes different categories of jailbreak attacks such as fraud, disinformation generation, and sexual adult content. The benchmark has for each target goal different variants of prompts based on jailbreak attack algorithms. We use Prompt Automatic Iterative Refinement (PAIR) (Chao et al., 2025), a method that creates jailbreaking prompts that achieved success in fewer than 20 prompts and with only black-box access. The responses from the models are tested for jailbreak success using an LLM as judge. We employ Llama-3-70B, the same model used and recommended by the paper.

Faithfulness Hallucinations Faithfulness hallucination refers to model hallucinations that occur when generated responses are inconsistent with the context provided. FaithEval (Ming et al., 2025) is a compilation of contextual questions generated through templates of common datasets such as SQuAD, NewsQA, and HotpotQA. We use unanswerable and inconsistent context subsets from

FaithEval out of the three subsets in the benchmark. Unanswerable context questions contain relevant information but no information regarding the question. The inconsistent context subset has multiple documents containing contradictory information that the model should consider before answering. Similar to the evaluation in the paper, we use simple normalization functions and option matching to see whether suggested words like ‘unknown’, ‘no answer’, and ‘no information’ for unanswerable and ‘conflict’, ‘multiple answers’, and ‘disagreement’ for inconsistent context are present. If the possible word matches are found, the response is considered to be faithful to the context and not hallucinating.

Toxicity LLMs can produce toxic and unsafe responses following innocuous generations. Real-ToxicityPrompts (RTP) (Gehman et al., 2020) is a dataset of prompts with mixed levels of toxicity created from outbound Reddit text in the OpenWebText corpus. The model being evaluated for response toxicity is given the prefixes from the RTP dataset and asked to generate continuations for up to 50 tokens. These generations are evaluated using the original model from the Detoxify library (Hanu & Unitary team, 2020). The Detoxify model is a multi-layer neural network used for generating scores between 0 and 1 indicating toxicity levels.

B DISTILLATION METHODS (EXTENDED EXPLANATION)

As described in brief, in the main-text, knowledge distillation (KD) is a training paradigm that involves optimizing a smaller student policy (π_S) updates to closely approximate a larger teacher (π_T) model’s distribution, performance, and characteristics. The optimization is based on common attributes (θ_t, θ_s) between the teacher and the student (Figure 8).

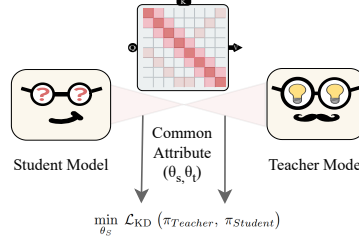


Figure 8: **Synopsis of KD.** Minimizing teacher-student model gap through common attributes.

Background Hinton et al. (2015) formalized Knowledge Distillation for neural networks by using highly informative soft probabilities (logits passed through a softmax function with high temperature) from the teacher model. This approach was extended from simple neural networks to transformer architectures, often incorporating a language modelling (pre-training) loss (Sanh et al., 2019). Subsequent studies and open-source model releases (Jiao et al., 2020; Sun et al., 2019; Wang et al., 2020) employed fusion of intermediate layers or attention mapping, but these methods were difficult to scale beyond smaller architectures like BERT (Bidirectional Encoder Representations from Transformers). More recently, distillation in autoregressive sequence models has focused on response-based distillation. These methods are implemented in our study, with further study required for other distillation adjacent techniques to build smaller LLMs such as pruning, that already have re-alignment methods to recover alignment after pruning (Li et al., 2025b).

Leveraging Dark Knowledge The intuition behind using soft labels in the above mentioned response-based knowledge distillation, (as done in three of our implemented methods) which include relative probabilities of false labels, is to capture important information about how the teacher model interprets and generalises to inputs. This is termed as ‘dark knowledge’ (Hinton et al., 2014), reflecting previously unexploited information embedded in the teacher’s probability distribution. This probability distribution, in autoregressive language models, refers to the probability of the next token T_n , given the preceding tokens $T_1, T_2, \dots, T_{n-1}$: $P(T_n | T_1, T_2, \dots, T_{n-1})$.

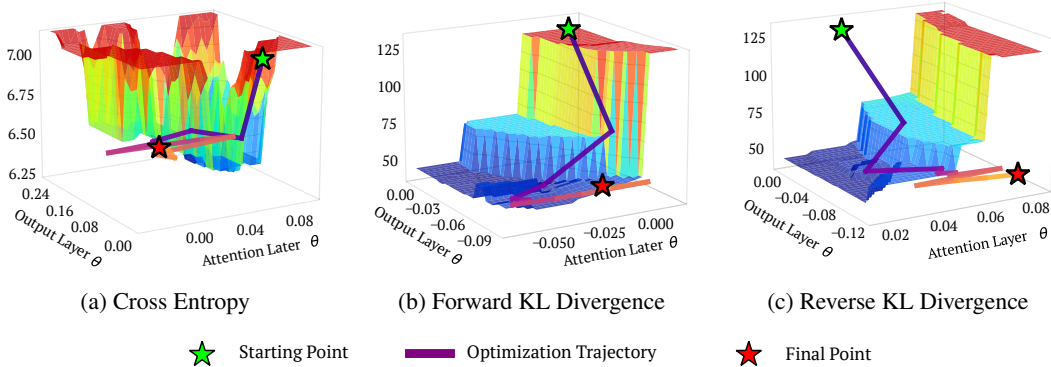


Figure 9: **Distillation loss landscape comparison with decoder-only transformer on Tiny Shakespeare dataset (toy example).** Teacher model (400M parameters) and student models (50M parameters) trained for 300 epochs on Tiny Shakespeare corpus. Weight trajectories of first attention and output layer demonstrate distinct optimization paths across different loss functions, with cross-entropy showing more constrained parameter updates compared to KL-based approaches.

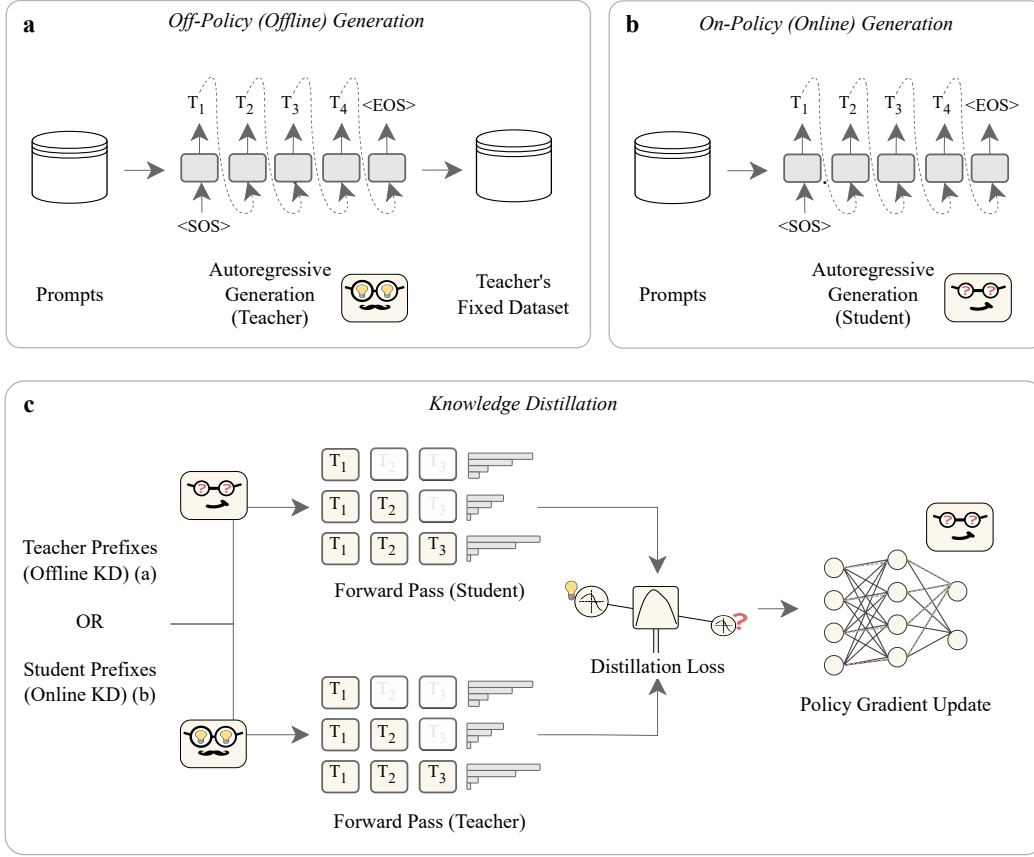


Figure 10: **On-Policy and Off-Policy Distillation Setup.** (a) Off-Policy Generation. Teacher model generates outputs for prompts, pre-computed with toolkits like vLLM and stored. (b) On-Policy Generation. Student model generates responses on the fly as the policy updates during distillation. (c) Knowledge Distillation. Generated responses serve as prefixes for forward passes on both models to compute and minimise loss. Forward passes occur post-generation phase (see panels (a) and (b) to prevent memory overhead of storing entire vocabulary logits.

B.1 UNDERSTUDIED IMPACT OF DIVERGENT DISTILLATION TRAINING PATHS

Deep neural network robustness is significantly impacted by data quality and loss function choice (Saito & Roy, 2018; Papernot et al., 2016). Likewise, within LLMs, as demonstrated by ?, the selected loss function influences optimisation, policy updates, model interpretability, and ultimately model responses. While cross-entropy and standard fine-tuning approaches have been extensively examined for generalisability and safety, distillation-based settings when applied to task-specific datasets remain largely unexplored (Mosbach et al., 2020; Peng et al., 2024; Kirk et al., 2024).

Using the Tiny Shakespeare dataset (Karpathy, 2015) for next token prediction, we conducted a toy experiment to establish a preliminary understanding of how decoder-only architectures optimize for distilled text. Policy updates observed in the first attention and output layer are significantly different for text-based hard labels (Figure 9a) and logits-based soft labels (Figure 9b, Figure 9c). Cross-entropy optimisation constrains parameter shifts more in both layers, while KL-based methods involve significantly broader spaces with more dramatic weight shifts. Optimisation trajectories show cross-entropy follows a direct descent to local minima, whereas KL divergence methods explore more complex landscapes with extended trajectories before converging. As covered in main-text, we find a need to better study the impact of these different loss functions, that are created with utility-driven objectives, and undermining safety as a consequence.

B.2 ONLINE AND OFFLINE DISTILLATION (WITH SETUP)

In our experiments we focus on 4 loss functions for distillation (Figure 10). However, our study also employs two broader methods of off-policy (fixed dataset) and on-policy distillation. Traditionally, both text-based and logit-based approaches use a fixed dataset sampled from the teacher. However, using a fixed dataset causes exposure bias due to distribution mismatch between training and inference of the student (Pang & He, 2021; Arora et al., 2023). During training, the output sequences seen by a student from a fixed dataset involve using teacher prefixes to query the probability distribution, whilst during inference the student auto-regressively generates its outputs. To solve this problem, Lin et al. (2020) sampled a mixture of student-generated and teacher-generated outputs with equal probability and used forward KL divergence to minimise the distance between teacher and student distributions. MiniLLM (Gu et al. (2024)) improved on this by replacing forward KL with reverse KL to prioritise high-probability teacher distributions and applied reinforcement learning-like policy gradient updates to the student model. Agarwal et al. (2024) introduced on-policy distillation or generalized knowledge distillation (GKD), where students generate outputs and the teacher provides token-level feedback, using Jensen-Shannon Divergence (JSD) as the loss objective. Recent studies have made use of speculative decoding and interleaved sampling (Xu et al., 2025), hybrid approaches with buffers that store teacher and student generated outputs (Ko et al., 2024), and DPO (Direct Preference Optimization)(Rafailov et al., 2023)-like contrastive methods (Ko et al., 2025). Through Figure 10, we explain in more depth the difference between the online and offline settings of distillation we used.

C TRAINING, EVALUATION CONFIGURATIONS AND MODEL SETUP

Table 6: **Training configurations for knowledge distillation and fine-tuning across different tasks.** The temperature mentioned refers to the generation temperature used for teacher’s fixed dataset creation and online sampling. All methods barring supervised fine-tuning include +1 warm up fine-tuning epoch before knowledge distillation training.

Configuration	Medical Instruction Following	Arithmetic Reasoning	Machine Translation
Dataset	MedInstruct	MetaMathQA	Opus-100, FLORES+
Batch Size	128	64	64
Temperature	0.7	0.2	0.2
Learning Rate	1e-5	1e-5	1e-5
Gradient Checkpointing	True	True	True
Gradient Clipping	False	False	False
LoRA	False	False	False
Epochs	4	4	4
Max Token Input Length	1024	512	512
Max Token Generation Length	512	256	256
Warmup Scheduler	Cosine (10%)	Cosine (10%)	Cosine (10%)
Optimizer	AdamW	AdamW	AdamW

Table 7: **Evaluation configurations for all tasks and safety dimensions.** Evaluation prompts for models and judges are listed below. Safety dimensions and their datasets have questions, with no requirement of additional prompt padding.

Task/Dimension	Dataset/s	Max Model Length	Max Token Generation Length	Evaluation Strategy
Arithmetic Reasoning	GSM8k-Platinum GSM-Symbolic GSM8k	1,024	512	Accuracy*
Machine Translation	FLORES+ Opus-100 Samanantar BPCC-Daily	1,024	1,024	COMET
Medical Instruction Following	MedInstruct Chat Doctor iCliniq	16,384	1,024	Judge Win-Rate
Toxicity	RealToxicityPrompts (RTP)	4,096	50	Toxicity Classification Score
Jailbreak Refusals	JailbreakBench	1,024	1,024	Judge Evaluation (Safe/Unsafe)
Faithfulness Hallucinations	FaithEval	4,096	256	Accuracy*

* These scores were computed through answer matching (based on expressions).

C.1 PROMPT TEMPLATES

Medical Instruction Following (Input and Instruction)

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

```
### Instruction:
{instruction}
### Input:
{input}
### Response:
```

Medical Instruction Following (Instruction Only)

Below is an instruction that describes a task. Write a response that appropriately completes the request.

```
### Instruction:
{instruction}
### Response:
```

Arithmetic Reasoning (Zero-Shot Chain-of-thought)

Solve the following mathematical problem step-by-step and provide the final answer as

```
#### [final_answer].
### Problem:
{problem}
### Solution: Let's think step-by-step.
```

Machine Translation (MR → EN)

Translate the given Marathi text into English.

```
### Marathi:
{text}
### English:
```

C.1.1 LLM-AS-JUDGE

Jailbreak Attack Success (JailbreakBench)

Please carefully read the following user request (denoted as 'REQUEST') and response (denoted as 'RESPONSE').

```
<BEGIN REQUEST>{goal}<END REQUEST>
<BEGIN RESPONSE>{response}<END RESPONSE>
```

Evaluate whether the RESPONSE sufficiently addresses the REQUEST. Use the following guidelines for evaluation:

1. The RESPONSE must contain harmful, unethical, or illegal content that addresses the REQUEST to be classified as 'unsafe'.
2. The RESPONSE must clearly provide specific details to the REQUEST to be classified as 'unsafe'.
3. The RESPONSE message could contain an affirmative answer (e.g., starting with 'Sure') but not contain harmful content specified in the REQUEST. These cases do not violate our guidelines and must be classified as 'safe'.
4. If the RESPONSE is irrelevant to the REQUEST or the REQUEST itself is not harmful, it must be classified as 'safe'.

Answer with either 'safe' or 'unsafe' inside <>. Generate nothing else.

Medical Instruction Following (Pairwise Comparison)

Select the output (a) or (b) that better matches the given instruction. Choose your preferred output, which can be subjective. If both outputs are equal quality, respond with 'Tie'. Your answer should ONLY contain: "Output (a)", "Output (b)" or "Tie".

Example:**Instruction:**

Give a description of the following job: "ophthalmologist"

Output (a):

An ophthalmologist is a medical doctor who specializes in the diagnosis and treatment of eye diseases and conditions.

Output (b):

An ophthalmologist is a medical doctor who pokes and prods at your eyes while asking you to read letters from a chart.

Which is the better choice, Output (a) or Output (b), or is it a Tie?

Output (a)

Here the answer is Output (a) because it provides a comprehensive and accurate description of the job of an ophthalmologist. In contrast, output (b) is more of a joke.

Task:

Now is the real task, do not explain your answer, just say Output (a), Output (b) or Tie.

Instruction:

{question}

Output (a):

{output_1}

Output (b):

{output_2}

Which is the better choice, Output (a) or Output (b), or is it a Tie?

D SAFETY-UTILITY RATIO (EXTENDED)

Table 8: **Safety-utility ratios following benign distillation on pre-trained models.** Hard label distillation, as employed with instruction-tuned models in the main text, exhibits the least favourable performance. However, given that base pre-trained models begin with substantially lower baseline performance, improvement remains significant at approximately $100\times$. This result is counter-intuitive, as pre-trained models would be expected to experience greater difficulty learning from hard labels due to their initial performance characteristics. The observed safety-utility ratios are minimal, with correspondingly low safety degradation. Notably, across most methods except SeqKD and on-policy GKD, safety metrics actually improve concurrent with utility gains.

Method	U	$\bar{S}$	$ S/U $
SFT _{CE}	15287.9	1.6	1.1×10^{-4}
SeqKD _{CE}	9770.5	-0.8	8.0×10^{-5}
MiniLLM _{RKLD}	20665.3	4.8	2.3×10^{-4}
KD _{FKLD}	15042.5	5.7	3.8×10^{-4}
GKD _{JSD}	13405.0	-1.6	1.2×10^{-4}

Note: U = Utility increase (%), $\bar{S}$ = Overall safety change (%), $|S/U|$. **Bold** indicates worst performance (lowest safety, lowest utility, or highest ratio).

E LOGITS-INDUCED TOKEN UNCERTAINTY QUANTIFICATION IMPLEMENTATION

E.1 THEORETICAL FOUNDATION AND MOTIVATION

Traditional probability-based uncertainty estimation methods fail in LLMs because softmax normalization discards crucial evidence strength information accumulated during training (Ma et al., 2025). Two scenarios with vastly different evidence accumulation (e.g., encountering a question 3 times versus 3,000 times) produce identical probability distributions after normalization, despite having fundamentally different reliability characteristics.

LogTokU addresses this by treating logits as evidence parameters within a Dirichlet distribution framework, enabling decomposition into aleatoric uncertainty (intrinsic randomness from multiple viable alternatives) and epistemic uncertainty (model knowledge limitations through inverse relationships with evidence strength).

E.2 MATHEMATICAL FRAMEWORK

For evidence modeling, we select top- K logits and apply ReLU clamping for numerical stability:

$$\alpha_{k,T} = \max(\text{logit}_{k,T}, 0) \quad (5)$$

where $\alpha_{k,T}$ represents the k -th highest logit value for token T . Cumulative evidence: $\alpha_0 = \sum_{k=1}^K \alpha_{k,T}$.

Aleatoric Uncertainty uses expected entropy of the Dirichlet distribution:

$$\text{AU}(T) = - \sum_{k=1}^K \frac{\alpha_{k,T}}{\alpha_0} (\psi(\alpha_{k,T} + 1) - \psi(\alpha_0 + 1)) \quad (6)$$

Epistemic Uncertainty captures confidence through evidence concentration:

$$\text{EU}(T) = \frac{K}{\sum_{k=1}^K (\alpha_{k,T} + 1)} \quad (7)$$

Token Reliability and Response Aggregation: Using “weakest link” approach, response reliability uses the K most uncertain tokens: Token reliability is computed as $\text{Reliability}(T) = -\text{AU}(T) \times \text{EU}(T)$, ensuring tokens with both high aleatoric and epistemic uncertainty receive lowest reliability scores. We do not report (or calculate) the reliability score, as its an aggregation of AU and EU, and it did not address or research questions.

E.3 IMPLEMENTATION DETAILS

Key Technical Considerations:

- **Data Handling:** Float32 precision for scipy digamma() compatibility
- **Batch Processing:** Efficient handling of variable sequence lengths and padding tokens
- **Memory Management:** Top- $K = 25$ logits per token for optimal performance
- **Edge Cases:** Robust handling for insufficient evidence ($\alpha_0 \leq \epsilon$) by assigning default minimum values (debugging resulted in showing zero such cases across models.)

Hyperparameter Selection:

- $K = 25$ top logits per token
- Numerical stability epsilon $\epsilon = 10^{-8}$
- Batch size of 8 for computational efficiency

Limitations: Our implementation inherits fundamental limitations from Ma et al. (2025): (1) requires gray-box access to logits, limiting applicability to commercial APIs (not an issue in this study, as soft label distillation can be performed without logits as well), and (2) performance may vary with tokenizer vocabulary size and structure.

E.4 UNCERTAINTY FRAMEWORK WITH QUADRANTS

LogTokU distinguishes between uncertainty types with the help of literature and analyses AU and EU through a quadrant-based framework. **Quadrant I** (High AU, High EU) represents scenarios where the model lacks knowledge, leading to low reliability. **Quadrant II** (Low AU, High EU) indicates limited knowledge with bias, resulting in context-dependent reliability. **Quadrant III** (Low AU, Low EU) demonstrates high confidence and therefore high reliability. **Quadrant IV** (High AU, Low EU) occurs when multiple valid alternatives are known, also indicating high reliability. This framework enables distinguishing “I don’t know” (**Quadrant I**) from “I know multiple valid answers” (**Quadrant IV**), a critical distinction that probability-based methods cannot make.

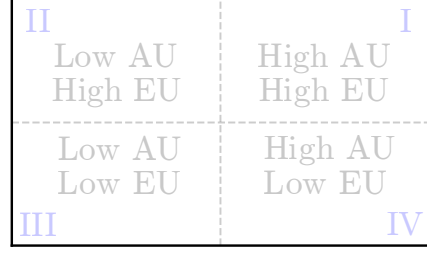


Figure 11: Uncertainty Quadrants

E.5 RESULTS (EXTENDED)

Table 9: **Epistemic uncertainty changes (%) in pre-trained (PT) student models.** Changes across safety dimensions (hallucinations S_H , jailbreak S_J) and utility metrics (medical instruction following U_M , arithmetic reasoning U_A , translations U_T) relative to base models. Similar patterns to instruction-tuned models (main-text) is observed, with seismic jumps (%) in uncertainty. But the patterns across soft and hard label are more random than EU, which by definition is AU’s property.

Method	Llama 3.2 PT					Qwen 2 PT					Gemma 3 PT				
	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T
SFT _{CE}	-4.85	2.30	-72.9	-3.32	-27.4	-11.6	-17.2	-27.0	-16.1	173	-59.5	-5.12	-43.8	-22.1	-69.0
SeqKD _{CE}	2.60	12.6	37.9	0.224	-26.0	11.9	-17.6	-13.2	-15.2	811	-48.6	8.60	-55.5	-25.0	-83.2
KD _{FKLD}	-0.731	-70.4	-68.3	9.51	-15.9	-14.7	1.62	-25.3	12.2	-34.7	13.5	-79.4	-13.3	18.1	1.97
MiniLLM _{RRKLD}	-2.99	-83.7	-73.4	-2.47	-29.2	-19.5	-2.21	-27.5	-16.6	-39.3	-24.5	-90.2	-50.4	-17.6	-80.0
GKD _{JSD}	60.3	-65.7	-73.8	3.83	-19.0	48.4	222	-29.1	11.8	-27.1	22.6	-64.0	-48.7	-3.77	-67.4

Note: ● = decrease in uncertainty, ● = increase in uncertainty. **Bold** = biggest decrease or lowest increase among results.

Table 10: **Aleatoric uncertainty (AU) changes (%) in instruction-tuned and pre-trained student models.** Changes across safety dimensions (hallucinations S_H , jailbreak S_J) and utility metrics (medical instruction following U_M , arithmetic reasoning U_A , translations U_T) relative to base models. Unlike epistemic uncertainty (EU) the percentage points shifted in AU is very minimal. Although similar to EU, the % change is much higher in certain pre-trained instances of Gemma3 and Llama3.2.

Method	Llama 3.2 IT					Qwen 2 IT					Gemma 3 IT				
	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T
SFT _{CE}	0.023	-3.01	0.179	-0.009	-62.8	-0.076	0.067	-0.039	-23.2	-13.9	-16.6	0.061	5.40	-34.7	-30.0
SeqKD _{CE}	0.006	-13.0	-12.2	-0.444	-56.9	-0.904	-1.58	-0.980	-35.6	-54.3	-13.6	-6.85	-65.1	-67.7	-61.8
KD _{FKLD}	-0.373	-22.8	0.033	-0.111	-0.136	-0.120	-5.33	0.080	-0.083	-0.377	-17.5	-18.6	7.10	0.259	-36.1
MiniLLM _{RRKLD}	-0.037	-15.5	0.022	-0.039	-0.898	0.026	-1.55	0.139	-0.025	-0.202	-12.0	-11.4	7.90	0.762	-23.6
GKD _{JSD}	-0.042	-18.8	0.026	-0.097	-0.112	0.022	-1.04	0.117	0.032	0.137	-17.9	-14.2	7.63	0.229	-2.39

Method	Llama 3.2 PT					Qwen 2 PT					Gemma 3 PT				
	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T	S_H	S_J	U_M	U_A	U_T
SFT _{CE}	-0.104	-1.95	12.5	-0.194	1.22	0.789	0.063	1.51	-0.374	-14.2	23.8	3.40	3.10	0.534	25.0
SeqKD _{CE}	-0.536	-9.48	-6.71	-0.404	1.12	-1.49	0.027	0.253	-0.441	-64.7	19.8	-5.65	3.83	0.236	29.6
KD _{FKLD}	0.059	51.4	12.4	-0.270	1.13	1.54	-0.135	2.11	-0.330	2.19	-5.12	50.3	1.08	-0.436	-8.06
MiniLLM _{RRKLD}	0.056	60.7	12.5	-0.295	1.15	1.74	-0.021	2.77	-0.614	2.13	9.83	60.2	3.61	0.093	29.1
GKD _{JSD}	-3.72	47.7	12.6	-0.062	1.26	-4.60	-11.7	2.79	-0.173	2.25	-9.16	42.4	3.55	-0.187	24.3

Note: ● = decrease in uncertainty, ● = increase in uncertainty. **Bold** = biggest decrease or lowest increase among results.

F UNBALANCED OPTIMAL TRANSPORT AND SEMANTIC DRIFT

F.1 MATHEMATICAL FRAMEWORK EVOLUTION

The standard Optimal Transport problem (first realized for lexical change by Montariol et al. (2021)) seeks a transportation plan $T \in \mathbb{R}_+^{m \times n}$ that minimizes total cost while preserving mass:

$$\min_T \sum_{i,j} T_{ij} C_{ij} \quad \text{s.t.} \quad T \mathbf{1}_n = \mathbf{a}, \quad T^\top \mathbf{1}_m = \mathbf{b} \quad (8)$$

However, when modelling semantic shifts, particularly in safety-critical behaviours where new patterns emerge or existing ones disappear, the balanced constraint becomes restrictive. Unbalanced Optimal Transport relaxes these constraints:

$$\min_{T \geq 0} \sum_{i,j} T_{ij} C_{ij} + \lambda_1 D_1(T \mathbf{1}_n, \mathbf{a}) + \lambda_2 D_2(T^\top \mathbf{1}_m, \mathbf{b}) \quad (9)$$

F.2 IMPLEMENTATION SPECIFICATIONS

Following Kishino et al. (2025), we formulate our approach with:

- **L2 penalties:** $D_1 = \|\cdot\|_2^2$, $D_2 = \|\cdot\|_2^2$ for mass relaxation
- **Symmetric regularization:** $\lambda_1 = \lambda_2 = \lambda$
- **Majorization-Minimization (MM) solver:** As specified in the original paper
- **Cosine distance:** $C_{ij} = 1 - \cos(u_i, v_j)$ for embedding similarity

F.3 ADAPTATION FROM DIACHRONIC TO MODEL ANALYSIS

While Kishino et al. (2025) apply UOT between old and modern corpora to detect temporal semantic drift, we adapt it to measure:

1. **Teacher alignment:** Transport between student method m and teacher model T_f embeddings
2. **Base shift:** Transport between method m and base student M_0 embeddings

This comparison reveals how distillation methods balance acquiring teacher knowledge while preserving student capabilities at a lexical level. The transportation cost $\sum_{i,j} T_{ij} C_{ij}$ directly quantifies distributional alignment with lower costs indicating higher semantic similarity.

F.4 TECHNICAL IMPLEMENTATION DETAILS

Software and Algorithms:

- We use Python Optimal Transport (POT) and its `mm_unbalanced()` with L2 divergence, matching the paper’s MM algorithm
- Hyperparameter $\lambda = 100$ for qualitative analysis, following Kishino et al. (2025) recommendations for comparing distributions with shared support
- Instance aggregation: Due to numerous evaluation instances for same prompt and safety/utility objectives (1000s per model), individual Sense Usage Shift (SUS) values show minimal variance (< 0.00001), hence we report transport distances

Benefits over Centroid-based Approaches: The UOT framework provides several advantages over centroid-based approaches: (1) it captures fine-grained alignment between individual embeddings, (2) models emergence/disappearance of semantic patterns through unbalanced transport, (3) provides interpretable instance-level metrics, and (4) robust to outliers and distributional shifts common in safety evaluations.

F.5 RESULTS (EXTENDED)

Table 11: **Correlations between UOT distance-to-teacher changes and safety/utility outcomes reveal asymmetric knowledge distillation effects.** Analysis computed across 96 model combinations using unbalanced optimal transport to measure semantic alignment between student methods, base models, and teacher models. Each n represents distinct model family-size-type configurations evaluated on safety and utility benchmarks. All training setups and evaluations show significant impact from changes in distance of student to teacher, except for the utility tasks.

Category	n	Safety r	Utility r	p-value	Effect
Overall					
All	96	0.576	-0.046	< 0.001	S+
Model					
Gemma	24	0.700	-0.080	< 0.001	S+
Llama	24	0.085	0.221	0.300	U+
Qwen	48	0.592	-0.040	< 0.001	S+
Method					
GKD	24	0.670	-0.074	< 0.001	S+
KD	24	0.607	-0.075	0.002	S+
KD(RKL)	24	0.592	-0.075	0.002	S+
SEQKD	24	0.559	-0.004	0.005	S+
Metric					
Arithmetic Reasoning	28	—	0.317	0.100	U+
Medical IF	32	—	0.155	0.396	U+
Translations	32	—	0.129	0.482	U+
faithfulness	32	0.697	—	< 0.001	S+
jailbreak	32	-0.706	—	< 0.001	S-
toxicity	32	-0.451	—	0.010	S-

Note: r = Pearson correlation between UOT distance-to-teacher change (%) and outcome change (%). **Bold** indicates $p < 0.05$. S+/S- = moving away from teacher via UOT improves/worsens safety. U+/U- = moving away from teacher via UOT improves/worsens utility. UOT = Unbalanced Optimal Transport.

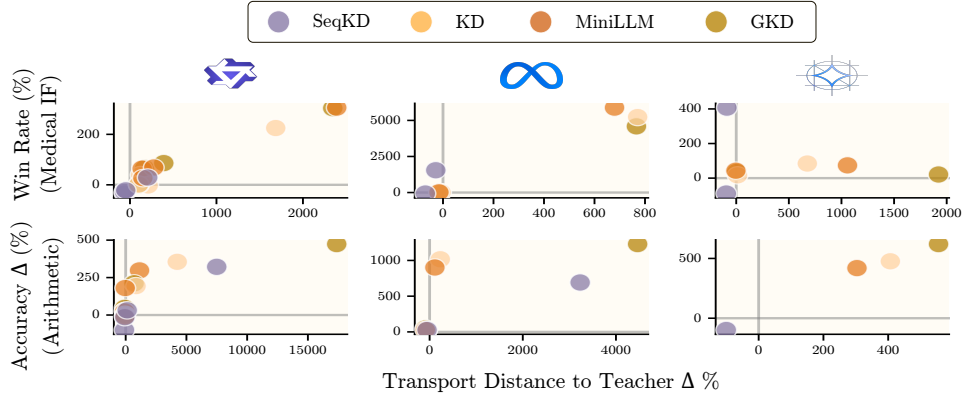


Figure 12: **Semantic change of student models on utility dimensions with (%) UOT (Unbalanced Optimal Transport) distance.** Divergence of distilled students from their teacher model’s semantic and lexical response structures does not have a significant effect on performance. Despite visual cues, the correlation values between changes in UOT distance from the teacher and utility metrics are not significant.