TOFU: FIXING SUPERVISED FINE-TUNING INDUCED DIVERSITY COLLAPSE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised Fine-Tuning (SFT) is essential for aligning Large Language Models (LLMs) with user intent, yet it is believed to suppress generative diversity. Although this reduction is frequently referenced, formal empirical testing of the phenomenon remains limited. The expressiveness of LLMs by itself was addressed by multiple prior methods. Their varying perspectives suggest that deeper investigation could yield further improvements. In this study, we attribute the decline to two primary drivers: the neglect of low-frequency patterns within fine-tuning datasets and the forgetting of preexisting knowledge. Motivated by our theoretical analysis, we develop Tempered Focal (TOFU) loss, a novel objective that addresses both stated challenges simultaneously. Our extensive evaluation confirms at scale that generation breadth narrows after SFT and strengthens the hypothesis explaining this effect. Across multiple models and benchmarks, we demonstrate that TOFU enhances output diversity while preserving high response quality, offering a principled approach to SFT.

1 INTRODUCTION

Autoregressive language models (Bengio et al., 2000) have demonstrated remarkable progress in modeling natural language. Increased data availability and model capacity have allowed Transformer-based (Vaswani et al., 2023) architectures to generate text that closely resembles human-written content. Modern Large Language Models (LLMs) now serve as powerful generative engines that excel across a broad range of specialized tasks, from document summarization (Brown et al., 2020) to complex reasoning (Wei et al., 2023).

Despite their impressive capabilities, pretrained LLMs often produce responses that do not fully capture user intents. This limitation arises because models may generate tokens that are statistically plausible yet semantically misaligned with the actual query, resulting in irrelevant or unhelpful output. To address this issue, researchers increasingly rely on instruction tuning (Wei et al., 2022; Chung et al., 2022; Raffel et al., 2023), also referred to as Supervised Fine-Tuning (SFT) (Ouyang et al., 2022; Bai et al., 2022a). This process refines LLMs on curated collections of high-quality prompt–response pairs, enabling better alignment with user expectations and task-specific objectives. This stage plays a crucial role in preparing models for subsequent reinforcement learning (RL) alignment (Ziegler et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024). By first grounding the model in high-quality demonstrations, SFT provides a stable initialization that enables RL methods to effectively refine task-specific behaviors, rather than struggling with unstructured or misaligned model outputs.

However, SFT also introduces notable challenges. While pretrained LLMs naturally generate diverse outputs, fine-tuning is considered to reduce this variety, as has been observed for specific model families and evaluation sets (O'Mahony et al., 2024). This reduction is problematic, as high diversity offers significant practical advantages: it aids post-training alignment (Ouyang et al., 2022), facilitates RL exploration (Bai et al., 2022b), and enhances complex reasoning (He et al., 2022; Wang et al., 2023). Furthermore, diverse generation benefits multi-model systems through complementary generations (Brown et al., 2005) and provides users with a wider range of perspectives.

In an effort to preserve distributional breadth, prior work has explored various regularization techniques, including weight decay during fine-tuning (Krogh & Hertz, 1991), noise injection into input embeddings (Jain et al., 2023), adaptive sampling strategies (Troshin et al., 2025; Nguyen et al.,

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

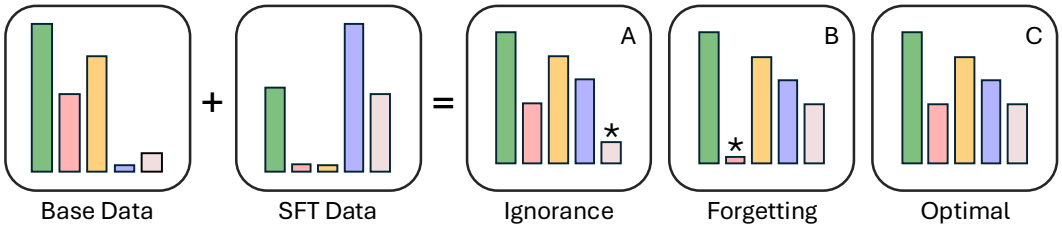

Figure 1: Impact of SFT on Generative Diversity. Comparison of data distributions and model states: Base Data represents the broad pretraining corpus, SFT Data represents the curated instruction set. The bars illustrate the discrete probability distribution over the vocabulary, color denotes token category. (A) **Ignorance** illustrates the failure to capture low-frequency SFT patterns, while (B) **Forgetting** depicts the erosion of the original pretraining knowledge. (C) **Optimal** represents the ideal balance where the model integrates new instructions without sacrificing the generative flexibility or richness of the base distribution.

2025), model ensembling (Hao et al., 2025), and alternative loss formulations (Li et al., 2024; Verine et al., 2025). Although these methods offer various improvements, they are inspired by different analytical paradigms and do not fully resolve the underlying factors limiting generative diversity, leaving room for further refinement.

In this work, we attribute the reduction in output diversity during SFT to the interplay of two factors (Figure 1). First, fine-tuning on small, curated datasets often induces overfitting to dominant patterns (Bethune et al., 2025), causing the model to neglect the long-tail distribution of the data. Second, the process can erode the diverse knowledge acquired during pretraining (Goodfellow et al., 2015), further narrowing the model's available response space. The combination of these two phenomena leads to an overall diminished diversity.

To counter these issues, we focus on loss-based regularization as a principled solution. We investigate existing SFT functions to determine their ability to preserve pretrained knowledge. In addition, we also evaluate Focal Loss (FL) (Lin et al., 2017) as a potential SFT objective for maintaining the balanced treatment of rare samples. Building on these insights, we propose a new Tempered Focal (TOFU) loss function explicitly designed to tackle both forgetting and ignorance simultaneously. Our main contributions can be summarized as follows:

1. We conduct an extensive evaluation of existing diversity-oriented SFT objectives across various model families, datasets, and benchmarks.
2. We investigate Focal Loss as an SFT objective for maintaining balanced treatment of rare samples and demonstrate its promising performance.
3. We propose a novel training objective, TOFU, designed to mitigate both ignorance and forgetting.
4. We show that our method achieves superior generative diversity while maintaining competitive output quality compared to standard approaches.

2 Theoretical Preliminaries and Background

Large Language Models Large Language Models are trained as next-token predictors over a discrete vocabulary $\mathcal{V}$. Given a sequence of tokens $x = (x_1, ..., x_L)$, the model defines a conditional distribution over the next token x_l for each vocabulary element $y \in \mathcal{V}$:

$$p_\theta(y|x_{<l}) \doteq p_\theta(x_l = y|x_{<l}). \tag{1}$$

Training consists of minimizing the Cross-Entropy (CE) between the model's predicted distribution and a target distribution $q \doteq q(y|x_{<l})$ at each position:

$$\mathcal{L}_{\text{CE}}(\theta) = -\mathbb{E}_{(x,l)\in\mathcal{D}}\mathbb{E}_q \log p_\theta(y|x_{<l}) \tag{2}$$

In standard supervised training q is one-hot on the ground-truth token. For simplicity, throughout this paper we focus on the loss corresponding to a single token, without explicitly showing its dependence on the context, as this omission is purely notational and does not change the underlying mathematics.

Supervised Fine-Tuning In Supervised Fine-Tuning, LLMs are adapted to specific tasks using sequences that combine a prompt and a response. The model conditions on the prompt as a fixed prefix but is optimized exclusively on the response. This ensures focus on generating the correct outputs for the task, while leveraging existing pretrained knowledge.

The standard training objective for SFT is the Cross-Entropy loss. It has a significant limitation: on relatively small fine-tuning datasets, it encourages the model to focus narrowly on the few observed responses while ignoring other plausible outputs (Li et al., 2024). Therefore, CE can reduce the expressiveness of the model's generation, a property that is important for downstream exploration and robust alignment. To address this issue, recent work in the community has explored variants of SFT that employ modified loss functions specifically designed to preserve or enhance output diversity.

GEM by Li et al. (2024) Game-theoretic Entropy Maximization (GEM) reframes SFT as a distribution-matching process in which learning is modeled as transferring probability mass from non-target to target tokens. Instead of relying on Cross-Entropy, which forces indiscriminate and unbounded probability flow, GEM introduces a game-theoretic formulation. In this setup, a meta-controller regulates how and where probability mass moves. This selective control prevents collapse of the output distribution and avoids over-penalizing semantically meaningful or rare tokens. Ultimately, this approach yields a practical training algorithm in which the entire framework reduces to optimizing a new loss function replacing Cross-Entropy.

Definition 2.1 (GEM loss (Li et al., 2024))**.** The GEM's objective is defined as follows:

$$\mathcal{L}_{\mathrm{GEM}}(\theta) = -\mathbb{E}_q[\log p_\theta] + \mathbb{E}_{\mathrm{sg}(p_\theta^\beta)}[\log p_\theta], \tag{3}$$

where $p_\theta^\beta \doteq \mathrm{softmax}\left(\beta^{-1} \log p_\theta\right)$ is a temperature-scaled distribution with $\beta \in (0, 1)$ as a temperature parameter. Note that $\mathrm{sg}(p_\theta^\beta)$ is a stop gradient operator and p_θ^β is detached from gradient computation. Here and throughout this work we use this notation for detached gradients.

Taken together, GEM produces sparse, targeted updates that preserve useful pretraining knowledge while still aligning the model to the supervised dataset. The reduced forgetting results in greater output diversity, as the model maintains a broader and more balanced token distribution.

λ-PR by Verine et al. (2025) λ-PR is a training objective that explicitly manages the trade-off between quality and diversity, building on previous approaches that improve generation.

Definition 2.2 (λ-PR loss (Verine et al., 2025))**.** The λ-PR's objective is defined as follows:

$$\mathcal{L}_{\lambda-\mathrm{PR}}(\theta) = -\mathbb{E}_q\left[w(\lambda, \alpha) \log p_\theta\right], \tag{4}$$

where $w(\lambda, \alpha) = \lambda^{\frac{l-1}{L}} \mathbb{I}_{\mathrm{sg}(p_\theta) \leq \delta} \frac{\mathrm{sg}(p_\theta)}{\alpha + (1-\alpha)\mathrm{sg}(p_\theta)}$. Here, $\lambda \in \mathbb{R}^+$ is the main parameter controlling a trade-off, $\alpha \in [0, 1]$ and $\delta = \frac{\alpha \lambda^{1/L}}{1-(1-\alpha)\lambda^{1/L}}$ (L being sequence length and l token's position).

It draws on ideas such as removing high-loss examples to reduce the impact of noisy references (Kang & Hashimoto, 2020), downweighting unlikely sequences to prevent degeneration (Ji et al., 2023), and reweighting gradients to promote high-probability outputs (Pang & He, 2021). Although these techniques were originally intended to enhance quality, λ-PR partially reverses their effects to optimize for diversity. The resulting compound loss function provides a unified framework for fine-tuning the balance between the two competing metrics. λ-PR can be interpreted as a weighted Cross-Entropy loss, where different tokens contribute unequally to the overall objective. The loss is motivated by two complementary goals: removing low log-loss examples to encourage a broader range of outputs, and downweighting less probable tokens to preserve accuracy.

Focal Loss by Lin et al. (2017) Focal Loss, previously widely used in computer vision for unbalanced training (Lin et al., 2017), has been recently explored for enhancing the quality of large language models (Rege Cambrin et al., 2024; Xia et al., 2025), yet its potential remains relatively overlooked. It addresses class imbalance by reshaping the standard Cross-Entropy loss to downweight well-classified examples.

Definition 2.3 (Focal Loss (Lin et al., 2017))**.** The Focal Loss is defined as follows:

$$\mathcal{L}_{\mathrm{FL}}(\theta) = -\mathbb{E}_q \left[(1 - p_\theta)^\gamma \log p_\theta \right], \tag{5}$$

where power coefficient $\gamma \geq 0$ is typically selected from range $[1, 5]$ (Lin et al., 2017; Mukhoti et al., 2020; Charoenphakdee et al., 2020).

We hypothesize that this loss can enhance the model's diversity by placing greater emphasis on underrepresented examples in the SFT dataset.

3 Proposed Method

Forgetting and ignorance arise from different aspects of the training dynamics. We propose to address both by combining the forgetting mitigation of GEM with the weighting of Focal Loss for underrepresented samples. However, combining the objectives is not straightforward, as adding terms or factors can significantly alter the update structure, leading to unstable optimization or poor interpretability. This motivates a careful analysis at the gradient level.

Theorem 3.1 (GEM loss equivalence)**.** *Training with $\mathcal{L}_{\mathrm{GEM}}$ is equivalent to training with temperature-scaled Cross-Entropy loss, as*

$$\nabla_\theta \mathcal{L}_{\mathrm{GEM}}(\theta) = \nabla_\theta \mathcal{L}^\beta_{\mathrm{CE}}(\theta), \tag{6}$$

where $\mathcal{L}^\beta_{\mathrm{CE}}(\theta) \doteq -\beta \mathbb{E}_q \log p^\beta_\theta$.

By examining GEM's gradients, we gain both theoretical insight and practical benefits. Specifically, our analysis reveals that GEM is equivalent to a temperature-scaled Cross-Entropy loss (Theorem 3.1, full proof is given in Appendix A.3). The equivalence clarifies GEM's advantage over methods that explicitly relax the predictive distribution by adding an entropy term. Instead of uniform flattening, GEM exerts softer pressure via adaptation, preventing the model from becoming overly confident while still guiding it toward the target. For practitioners, this equivalence also simplifies the computation required to implement GEM.

Theoretically, the GEM objective shares a global minimum with Cross-Entropy when the latter is scaled by β^{-1} after fine-tuning. In practice, however, the model does not converge to that point due to the limited updates in SFT and inherent optimization noise. This allows GEM to reach a parameter solution distinct from that of standard post-hoc temperature scaling (Ficler & Goldberg, 2017).

We next analyze Focal Loss in a similar manner. Prior work (Mukhoti et al., 2020) has shown that despite not being a proper loss it can be interpreted as a gradient-scaled Cross-Entropy. For completeness, we restate this result and explicitly state the assumptions required for the equivalence in Proposition 3.2 (the full proof is given in Appendix A.4).

Proposition 3.2 (Focal Loss equivalence for one-hot targets)**.** *Assume that the target distribution q is one-hot. Then, for Focal Loss $\mathcal{L}_{\mathrm{FL}}$ and Cross-Entropy $\mathcal{L}_{\mathrm{CE}}$, the gradients satisfy*

$$\nabla_\theta \mathcal{L}_{\mathrm{FL}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}_{\mathrm{CE}}(\theta), \tag{7}$$

where $g(p, \gamma) = (1 - p)^\gamma - \gamma p (1 - p)^{\gamma - 1} \log p$. Here and throughout this work $\hat{p}$ denotes the predicted probability assigned to the ground-truth token.

The scaling function $g(p, \gamma)$ increases the weight corresponding to the underrepresented samples (Figure 2). By peaking at a small but non-zero p, the gradient magnitude prioritizes moderately difficult tokens relative to extremely hard ones. This behavior echoes the idea of downweighting based on predicted probability usually employed for quality enhancement (Kang & Hashimoto, 2020; Pang & He, 2021; Ji et al., 2023).

Comparing FL scaling with that of the λ-PR (Figure 2), one might notice the latter's downside: it cannot account for a previously zeroed-out probability p without assigning it the highest possible weight. Consequently, the γ parameter of FL provides more flexibility in steering the weighting balance than the δ of λ-PR.

Corollary 3.3. *If q is one-hot, then weighting $\mathcal{L}^\beta_{\mathrm{CE}}(\theta)$ by the focal term $(1 - p^\beta_\theta)^\gamma$ results in a gradient of the form $g(\hat{p}^\beta_\theta, \gamma) \nabla_\theta \mathcal{L}^\beta_{\mathrm{CE}}(\theta)$.*

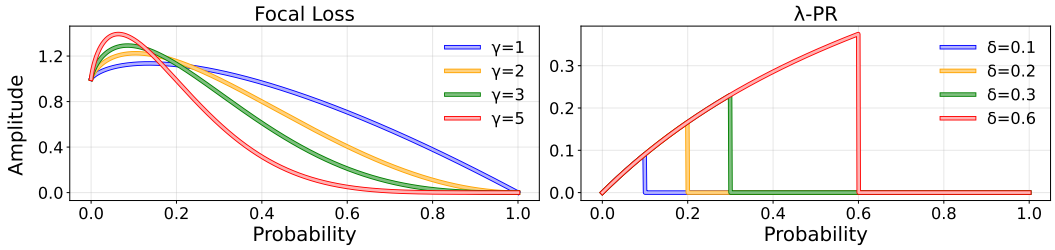

Figure 2: Gradient scaling amplitudes for Focal Loss and λ-PR as functions of various parameters and probabilities p. For λ-PR we omit the dependence on tokens position for simplicity.

The equivalences established by Theorem 3.1 and Proposition 3.2 allow us to merge the core ideas of GEM and Focal Loss by applying the focal term to the temperature-scaled Cross-Entropy. However, as follows from Corollary 3.3 (the complete proof is given in Appendix A.5), a naive multiplication of $\mathcal{L}_{\text{CE}}^{\beta}(\theta)$ by the focal term can excessively amplify the influence of very small probabilities in gradient weighting. This occurs because temperature-scaling reduces these probabilities further, which in turn increases the corresponding focal weight, potentially leading to unintended overemphasis on hard examples. Ideally, one would like to scale the GEM gradients using the focal scaling function $g(p, \gamma)$ evaluated on the unscaled probabilities p. Simply multiplying the temperature-scaled Cross-Entropy loss by a focal term that depends on p does not achieve this, because the gradients would propagate differently through p and p^{β}. The desired effect can be attained by using a detached version of $g(p, \gamma)$, which ensures that the gradient is scaled correctly. Therefore, in this work, we present Tempered Focal (TOFU) training objective, which naturally combines the advantages of Focal Loss and GEM.

Definition 3.4 (TOFU loss)**.** Let us denote p^{β} as a temperature-scaled distribution (2.1), then

$$\mathcal{L}_{\text{TOFU}}(\theta) = -\mathbb{E}_q \left[\text{sg}(g(\hat{p}_\theta, \gamma)) \beta \log p_\theta^{\beta} \right], \tag{8}$$

where $g(\hat{p}_\theta, \gamma)$ is detached from gradient computation.

Due to the inclusion of the focal term, TOFU does not directly optimize a well-defined statistic (unlike Cross-Entropy, which minimizes the Kullback–Leibler divergence).

Corollary 3.5. *If the target distribution q is one-hot, TOFU loss gradient is proportional to the one of the temperature-scaled CE:*

$$\nabla_\theta \mathcal{L}_{\text{TOFU}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}_{\text{CE}}^{\beta}(\theta) \tag{9}$$

Nevertheless, it can be interpreted as a form of adaptive gradient scaling of the temperature-scaled Cross-Entropy loss, emphasizing under-predicted tokens while preventing the learned distribution from becoming overly concentrated (Corollary 3.5, proof is given in Appendix A.6).

4 Experimental Evaluation

4.1 Tested LLMs

To properly evaluate our method and its alternatives, we aim for an experimental setup that is both realistic and diverse in terms of the models tested. In practice, widely used families such as Llama-3 (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024), and Qwen-3 (Yang et al., 2025) suggest an underlying post-training process that results in security guardrails, though their exact development pipelines remain largely opaque. The potential of alignment to obscure the isolated effects of SFT necessitates a more careful evaluation design. Therefore, in addition to Llama-3.1-8B, Qwen-3-8B, and Phi-4-14B, we focus on the models that have completed only the pretraining stage: OLMo-2-13B (Groeneveld et al., 2024), Mistral-12B (Mistral AI & NVIDIA, 2024), and Pythia-12B (Biderman et al., 2023). Although the resulting selection is expected to perform well in simple settings, complex reasoning requires specific training. To meet this requirement, we have added to our setup Qwen-2.5-Math-1.5B, Qwen-2.5-Math-7B (Yang et al., 2024) and DeepSeek-Math-7B (Shao et al., 2024).

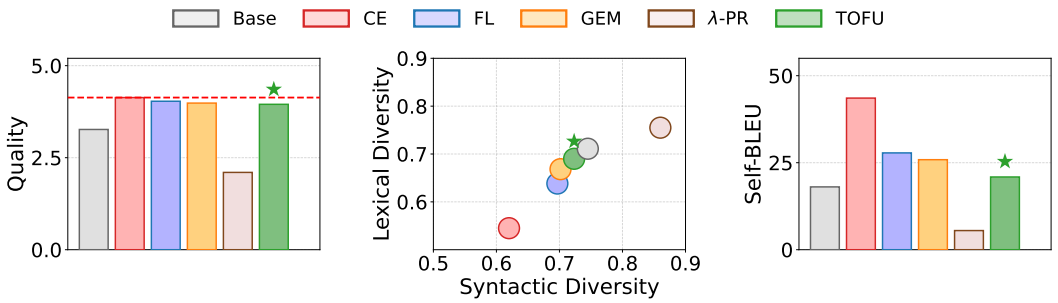

Figure 3: Performance of models across Alpaca SFT objectives on Alpaca Eval. (**Left**) Quality is measured via LLM Judge score (0–5), where higher scores are better. (**Center**) Lexical and Syntactic diversity scores (0-1), where higher scores are better. (**Right**) Self-BLEU scores (0–100), where lower scores are better. The optimal performance achieved by TOFU is highlighted with $\star$.

As such, we are able to evaluate the SFT objectives both in isolation and in combination with other post-training procedures on a wide range of tasks.

4.2 Benchmarks

Instruction Following As our primary goal is to investigate how diversity is affected by SFT under different objectives, we prioritize benchmarks containing open-ended prompts that favor creative synthesis over deterministic accuracy. The more straightforward tasks, such as story continuation and constrained generation (instruction following), are represented by ROC Stories (ROC) and Alpaca Eval (AE), respectively. We composed both datasets from open-source materials, with the curation process detailed in Appendix B. Further challenging the models, we employ NoveltyBench (NB) (Zhang et al., 2025). It is designed to measure the capacity for generating multiple distinct high-quality outputs using specifically curated prompts to elicit diverse responses.

Reasoning We also examine whether improved diversity aids in solving complex reasoning tasks by enabling a broader search for correct solutions across a suite of mathematical benchmarks: MATH500 (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), MinervaMath (Lewkowycz et al., 2022) and OlympiadBench (He et al., 2024). Going beyond mathematical tasks, we also evaluate the coding abilities of fine-tuned models on the HumanEval benchmark (Chen et al., 2021). The tested ability is particularly relevant if the goal of SFT is to enhance exploration without introducing low-quality noise, providing a better foundation for the subsequent Reinforcement Learning phase.

Factuality One might be concerned that prioritizing output diversity could degrade factual accuracy. To monitor this trade-off, we evaluated the models on standard multiple-choice benchmarks for retrieval of professional and scientific knowledge, ARC (Clark et al., 2018) and MMLU (Hendrycks et al., 2020). As our instruction tuning utilizes general-purpose datasets, these tasks primarily measure the retention of information from the initial training phase.

Safety Increased diversity may reduce sensitivity to malicious prompts by broadening the range of possible responses, including unsafe ones. To investigate this topic, we adopted two commonly used red-teaming benchmarks, Malicious Instruct (Huang et al., 2023) and HarmBench (Mazeika et al., 2024), and evaluated the model's robustness to adversarial and harmful instructions.

4.3 Metrics

We evaluate output diversity using several complementary approaches across the benchmarks. On the ROC Stories and Alpaca Eval, we employ the widely adopted Self-BLEU metric (Zhu et al., 2018), which quantifies surface-level variation by measuring word and phrase overlap across responses for a given prompt. However, Self-BLEU may not capture linguistic diversity of the output. To address this limitation, we additionally calculate lexical and syntactic diversity metrics adopted from (Guo

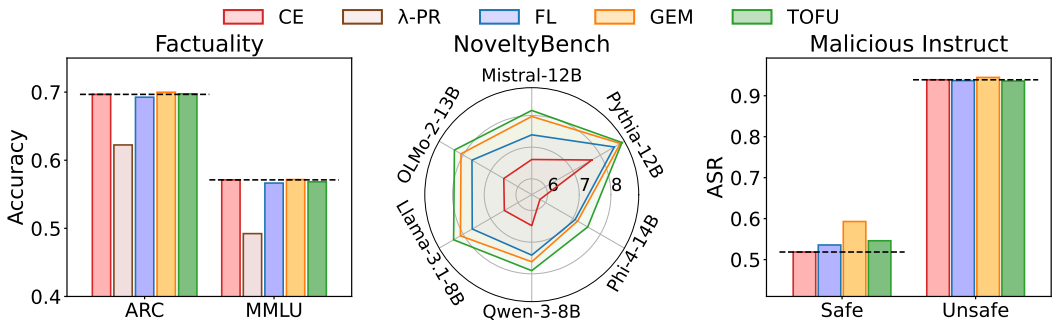

Figure 4: **(Left)** Mean accuracy (0-1) for the ARC and MMLU benchmarks averaged across the tested models. **(Center)** Distinct (1–10) scores on NoveltyBench dataset for different models and SFT methods, higher values indicate superior performance. **(Right)** Mean ASR scores (0-1) for the Malicious Instruct benchmark averaged across the safe and unsafe models, lower scores indicate better safety. Across all plots, error bars are omitted to maintain visual clarity and dotted lines serve as a reference point for CE performance.

et al., 2025). NoveltyBench provides its own diversity metric, *Distinct*, employing a specialized LLM based classifier, that prioritizes semantic variation over surface-level linguistics. Consequently, it captures meaningful distinctions between outputs that standard metrics may overlook.

To assess the quality of the generations for ROC and AE, we use an LLM-as-a-judge approach (Liu et al., 2023), scoring responses based on coherence, consistency, and fluency (Appendix B.6). Similarly, NoveltyBench proposes the *Utility* metric and its associated LLM usefulness judge.

For reasoning tasks, we evaluate exploration capabilities by calculating coverage (Pass@k), defined as the probability that at least one generated solution converges to the correct answer. We also track the mean success rate to distinguish between two behaviors: whether a model simply generates correct solutions more frequently (higher precision) or whether its increased diversity allows it to discover correct solutions for harder problems (Cobbe et al., 2021). In the latter case, the coverage would increase even if the mean success rate remains stagnant. To calculate both coverage and mean success scores, we extracted the final answers provided by the model in bounding box format and compared them against the ground truth.

For factual tasks, we decide on the output correctness directly comparing it with the reference answer. To evaluate safety alignment, we employ an Attack Success Rate (ASR) (Zou et al., 2023), the percentage of instructions that receive misaligned outputs, defined as failing to abstain from responding to a malicious instruction.

5 Results

To investigate the performance of TOFU, we evaluate it along with the standard Cross-Entropy and the diversity-oriented objectives GEM and λ-PR. To isolate the impact of prioritizing infrequent patterns, we also benchmark Focal Loss. Hyperparameter settings for GEM and λ-PR match their original publications, whereas the configurations for Focal Loss and TOFU were derived from ablation studies (see Appendix C). Notably, TOFU's parameters were selected only once, on the ARC dataset, and remained fixed across all subsequent experiments, demonstrating the method's robustness.

Creative Writing & Instruction Following We first explore models diversity on ROC Stories and Alpaca Eval. The results for Alpaca as SFT dataset are pictured by Figure 3. On both benchmarks, Cross-Entropy increases quality but shows a significant reduction in diversity compared to the base model, confirming the problem addressed in this work. The only exception is the Qwen-3-8B model when evaluated on ROC Stories (see Table D.1). Its post-training procedures appear to have a significant residual effect: even in the absence of an instruction template, the model consistently attempts to engage in Chain-of-Thought reasoning. This case highlights the importance of employing unaligned models for SFT-related evaluations to avoid the confounding effects of prior alignment.

Table 1: Performance of math reasoning models fine-tuned with different objectives on math datasets. The best values for the coverage and the average success rate (in parentheses) are highlighted in bold.

Bench	Method	Qwen-2.5-Math-1.5B	Qwen-2.5-Math-7B	DeepSeek-Math-7B	Mathstral-7B
MATH500	CE	78.4 (**53.9**)	84.8 (54.2)	71.4 (32.7)	76.2 (49.7)
	GEM	78.4 (53.4)	83.2 (**59.0**)	72.2 (**36.4**)	78.2 (**50.0**)
	TOFU	**80.6** (50.3)	**86.0** (53.8)	**72.6** (33.0)	**81.0** (48.5)
Minerva	CE	28.3 (12.1)	30.9 (**15.1**)	34.2 (**12.5**)	32.0 (13.3)
	GEM	31.3 (**12.4**)	34.9 (**15.1**)	35.3 (**12.5**)	32.7 (**14.7**)
	TOFU	**33.5** (10.7)	**38.6** (14.7)	**39.3** (12.0)	**36.0** (12.5)
GSM8K	CE	79.2 (54.5)	81.7 (60.1)	76.0 (48.0)	77.7 (**60.6**)
	GEM	73.2 (**55.5**)	76.2 (**61.5**)	72.4 (**51.2**)	78.9 (58.6)
	TOFU	**80.6** (51.8)	**88.0** (46.8)	**79.0** (47.5)	**83.6** (58.5)
OlympiadBench	CE	59.9 (**30.7**)	64.5 (**36.6**)	48.4 (16.5)	58.6 (**27.2**)
	GEM	60.7 (30.5)	64.9 (35.8)	51.8 (**16.8**)	59.9 (26.4)
	TOFU	**62.3** (28.9)	**67.1** (33.8)	**52.2** (13.8)	**61.4** (24.8)

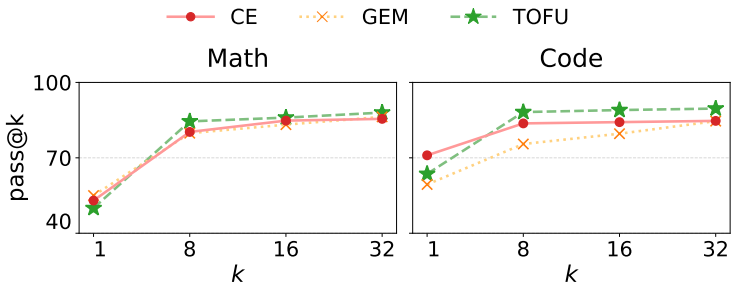

Figure 5: Performance in coverage (pass@k) averaged on reasoning models evaluated on MATH500 for mathematical problems and HumanEval for coding tasks across different k budgets.

Beyond the initial baseline, our evaluation reveals a clear hierarchy of objectives performance. Focal Loss yields noticeably higher diversity than Cross-Entropy, though this gain sometimes comes at a slight cost to quality. This validates our hypothesis that diversity is negatively affected by the ignorance of underrepresented training samples during the SFT process. TOFU maintains quality on par with GEM while achieving superior diversity across all the models and therefore emerges as the prominent choice for these benchmarks.

While λ-PR reaches the absolute highest diversity, it does so at the cost of significant quality degradation, pushing the method beyond the bounds of usability. Consequently, to maintain a focused and computationally efficient analysis, we decided to limit the use of λ-PR in further evaluation.

The performance hierarchy on the NoveltyBench dataset remains largely consistent with our previous findings, as illustrated in Figure 4. Specifically, TOFU consistently outperforms other objectives across all model families by demonstrating a superior Distinct score. Notably, quality remains comparable regardless of the model architecture, given the high Utility variance (Appendix D). Ultimately, the results establish TOFU as the most effective diversity-oriented approach across both structured instruction and open-ended creative tasks. Additional experiments with UltraFeedback as an alternative SFT dataset further solidify this conclusion (Appendix D).

Mathematical and Coding Reasoning Our results on mathematical reasoning, summarized in Table 1, demonstrate that TOFU models achieve a visibly higher coverage than the Cross-Entropy and GEM, while average success rates remain stagnant or slightly decline. This observation suggests that TOFU does not increase per-sample correctness, but instead promotes more diverse exploration, leading to a higher probability of finding a correct solution. Such behavior is further supported by Figure 5, which gathers coverage of reasoning models evaluated on coding tasks. Importantly, we find that TOFU achieves superior coverage at lower k budgets compared to GEM and CE. We anticipate

that this result is particularly relevant for the subsequent Reinforcement Learning stage, which is typically executed by employing GRPO (Shao et al., 2024).

Factuality We observe no evidence of catastrophic forgetting on the factuality tasks, with performance remaining largely comparable to CE models. The sole exception is λ-PR, which exhibits a consistent performance degradation. In Figure 4 we report accuracies after Alpaca SFT averaged across all models, while the full results are available in Appendix D. Overall, TOFU high diversity does not come at the cost of factual knowledge.

Safety To demonstrate that the diversity gains do not compromise models safety, we categorized models into two groups: *safe* models (Qwen-3 and Phi-4), which express inherent security guardrails, and *unsafe* models, which are much less sensitive to malicious prompts. Figure 4 reports the average Attack Success Rate (ASR) for both groups on Malicious Instruct, while detailed results on both Malicious Instruct and HarmBench are provided in Appendix D. Our results indicate that for the unsafe model group, safety metric appears to have reached a saturation point where further degradation is negligible. Conversely, for safe models, TOFU does not lead to an increase in ASR, demonstrating that the method avoids introducing new vulnerabilities to jailbreak attempts.

6 Discussions

In this work, we have presented an extensive evaluation of SFT objectives across a range of model families, datasets, and benchmarks. Through a systematic analysis we provide robust confirmation that standard Cross-Entropy SFT consistently reduces diversity. To our best knowledge, this is the first study to systematically validate this trend across a broad range of model scales and diverse SFT datasets, as previous work of O'Mahony et al. (2024) primarily explored this hypothesis within the Pythia family at smaller sizes. By extending the evaluation to the more substantial architectures, we provide empirical evidence that diversity loss is an inherent characteristic of the standard SFT pipeline rather than an artifact of specific smaller-scale architectures.

We initially attributed the decline to two primary factors: the erosion of pretrained information and the ignorance of low-frequency patterns in the fine-tuning data. This hypothesis is supported by the performance of GEM, designed to mitigate forgetting, and Focal Loss, designed to prioritize underrepresented samples, both of which yield measurable improvements in diversity compared to CE. Through careful theoretical analysis, we created a novel training objective, TOFU, which tackles both challenges simultaneously.

TOFU achieves the highest diversity while maintaining highly competitive quality across our creative writing and instruction following benchmarks. Furthermore, we found that in mathematical Chain-of-Thought reasoning and coding TOFU encourages a higher exploration mode, thereby increasing the probability of capturing a correct solution. This amplified response breadth might provide a superior foundation for the Reinforcement Learning alignment phase. Additionally, we showed that expanded diversity does not come at the cost of factual integrity and safety alignment. Altogether, these results position TOFU as a robust framework for improving model expressivity, enabling greater functional utility across a wide range of downstream applications.

While TOFU demonstrated stable performance across various model sizes and SFT datasets, there remains potential for further validation. Given greater computational resources, future experiments could evaluate these objectives on larger-scale architectures and higher-quality closed-source instruction tuning datasets. Finally, while this study focuses on the supervised phase, the exact influence of the diversity-enhancing objectives on subsequent Reinforcement Learning stages remains an open question for future investigation. We hope that our results encourage further research into specialized training objectives that leverage theoretical insights to refine model performance.

AI use statement

In this work, we used generative AI tools for assisting in writing and text polishing. The core method development in this research does not involve generative AI tools as any important, original, or non-standard components. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work aims to advance the field of Machine Learning by addressing generative diversity and knowledge retention in fine-tuned Large Language Models. By mitigating the forgetting of pretrained information and the underrepresentation of rare patterns, our research supports the development of models that produce more varied and information-rich outputs. There are many potential societal consequences of improving generative variety, none of which we feel must be specifically highlighted here beyond the ethical considerations standard to the advancement of language modeling.

Reproducibility statement

The code and datasets are available at `https://anonymous.4open.science/r/TOFU-DIV`.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report, December 2024. URL `http://arxiv.org/abs/2412.08905`. arXiv:2412.08905 [cs].

Wasi Uddin Ahmad, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Vahid Noroozi, Somshubra Majumdar, and Boris Ginsburg. Opencodeinstruct: A large-scale instruction tuning dataset for code llms, 2025. URL `https://arxiv.org/abs/2504.04030`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022a. URL `http://arxiv.org/abs/2204.05862`. arXiv:2204.05862 [cs].

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, 2022b. URL `https://arxiv.org/abs/2212.08073`.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL `https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html`.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling Laws for Forgetting during Finetuning with Pretraining Data Injection, May 2025. URL `http://arxiv.org/abs/2502.06042`. arXiv:2502.06042 [cs].

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, 2023. URL `https://arxiv.org/abs/2304.01373`.

Gavin Brown, Jeremy L. Wyatt, and Peter Tino. Managing Diversity in Regression Ensembles. *Journal of Machine Learning Research*, 6(55):1621–1650, 2005. ISSN 1533-7928. URL http://jmlr.org/papers/v6/brown05a.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On Focal Loss for Class-Posterior Probability Estimation: A Theoretical Perspective, 2020. URL https://arxiv.org/abs/2011.09172.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, December 2022. URL http://arxiv.org/abs/2210.11416. arXiv:2210.11416 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with Scaled AI Feedback, 2023. URL https://arxiv.org/abs/2310.01377.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL http://arxiv.org/abs/2305.14314. arXiv:2305.14314 [cs].

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback, 2023. URL https://arxiv.org/abs/2305.14387.

Jessica Ficler and Yoav Goldberg. Controlling Linguistic Style Aspects in Neural Language Generation, July 2017. URL http://arxiv.org/abs/1707.02633. arXiv:1707.02633 [cs].

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile:

An 800GB Dataset of Diverse Text for Language Modeling, December 2020. URL http://arxiv.org/abs/2101.00027. arXiv:2101.00027 [cs].

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks, March 2015. URL http://arxiv.org/abs/1312.6211. arXiv:1312.6211 [stat].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers,

Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783. Version Number: 3.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models, June 2024. URL http://arxiv.org/abs/2402.00838. arXiv:2402.00838 [cs].

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models, 2025. URL https://arxiv.org/abs/2412.10271.

Yifan Hao, Xingyuan Pan, Hanning Zhang, Chenlu Ye, Rui Pan, and Tong Zhang. Understanding Overadaptation in Supervised Fine-Tuning: The Role of Ensemble Methods. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 22116–22148. PMLR, October 2025. URL https://proceedings.mlr.press/v267/hao25d.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiad-

Bench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems, 2024. URL https://arxiv.org/abs/2402.14008.

Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with Retrieval: Faithful Large Language Model Inference, December 2022. URL http://arxiv.org/abs/2301.00303. arXiv:2301.00303 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, 2020. URL https://arxiv.org/abs/2009.03300.

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. URL https://arxiv.org/abs/2103.03874.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration, February 2020. URL http://arxiv.org/abs/1904.09751. arXiv:1904.09751 [cs].

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation, 2023. URL https://arxiv.org/abs/2310.06987.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. NEFTune: Noisy Embeddings Improve Instruction Finetuning, October 2023. URL http://arxiv.org/abs/2310.05914. arXiv:2310.05914 [cs].

Haozhe Ji, Pei Ke, Zhipeng Hu, Rongsheng Zhang, and Minlie Huang. Tailoring Language Generation Models under Total Variation Distance, February 2023. URL http://arxiv.org/abs/2302.13344. arXiv:2302.13344 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Daniel Kang and Tatsunori Hashimoto. Improved Natural Language Generation via Loss Truncation, 2020. URL https://arxiv.org/abs/2004.14589.

Anders Krogh and John Hertz. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models, 2022. URL https://arxiv.org/abs/2206.14858.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath, 2024. URL https://huggingface.co/AI-MO/NuminaMath-CoT.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving Diversity in Supervised Fine-Tuning of Large Language Models, 2024. URL https://arxiv.org/abs/2408.16673. Version Number: 2.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection, 2017. URL https://arxiv.org/abs/1708.02002. Version Number: 2.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. URL http://arxiv.org/abs/2303.16634. arXiv:2303.16634 [cs].

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Mistral AI and NVIDIA. Mistral NeMo, 2024. URL https://mistral.ai/news/mistral-nemo.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories, 2016. URL https://arxiv.org/abs/1604.01696.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating Deep Neural Networks using Focal Loss, 2020. URL https://arxiv.org/abs/2002.09437. Version Number: 2.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs, November 2025. URL http://arxiv.org/abs/2407.01082. arXiv:2407.01082 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, volume 2, 2024. URL https://openreview.net/forum?id=3pDMYjpOxk.

Richard Yuanzhe Pang and He He. Text Generation by Learning from Demonstrations, March 2021. URL http://arxiv.org/abs/2009.07839. arXiv:2009.07839 [cs].

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, 2023. URL https://arxiv.org/abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023. URL http://arxiv.org/abs/1910.10683. arXiv:1910.10683 [cs].

Daniele Rege Cambrin, Giuseppe Gallipoli, Irene Benedetto, Luca Cagliero, and Paolo Garza. Beyond Accuracy Optimization: Computer Vision Losses for Large Language Model Fine-Tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12060–12079, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.704. URL https://aclanthology.org/2024.findings-emnlp.704/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle

Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research, June 2024. URL http://arxiv.org/abs/2402.00159. arXiv:2402.00159 [cs].

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An Instruction-following LLaMA model, 2023. URL https://github.com/tatsu-lab/stanford_alpaca. Publication Title: GitHub repository.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL https://arxiv.org/abs/2302.13971. Version Number: 1.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs, September 2025. URL http://arxiv.org/abs/2510.01218. arXiv:2510.01218 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Alexandre Verine, Florian Le Bronnec, Kunhao Zheng, Alexandre Allauzen, Yann Chevaleyre, and Benjamin Negrevergne. Improving Diversity in Language Models: When Temperature Fails, Change the Loss, August 2025. URL http://arxiv.org/abs/2508.09654. arXiv:2508.09654 [cs].

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models, March 2023. URL http://arxiv.org/abs/2203.11171. arXiv:2203.11171 [cs].

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners, February 2022. URL http://arxiv.org/abs/2109.01652. arXiv:2109.01652 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, and Benjamin Roth. Influences on LLM Calibration: A Study of Response Agreement, Loss Functions, and Prompt Styles, 2025. URL https://arxiv.org/abs/2501.03991. Version Number: 1.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating Language Models for Humanlike Diversity, 2025. URL https://arxiv.org/abs/2504.05228.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A Benchmarking Platform for Text Generation Models, February 2018. URL http://arxiv.org/abs/1802.01886. arXiv:1802.01886 [cs].

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences, January 2020. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs].

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, 2023. URL https://arxiv.org/abs/2307.15043.

APPENDIX

A OMITTED PROOFS

Proposition A.1 (Notation and Logit Gradients)**.** *We introduce the notation used throughout the following proofs and derivations. All expressions below are written for a single token, with indices ranging over the vocabulary (i.e., number of classes). Consider a temperature-scaled distribution p^β from Definition (2.1), a Kronecker's delta matrix δ_{ij}, and z as model's output logits. Let us denote the following log probabilities*

$$l_i = \log p_i = z_i - \log \sum_k e^{z_k} \tag{A.1}$$

$$l_i^\beta = \log p_i^\beta = \frac{z_i}{\beta} - \log \sum_k e^{\frac{z_k}{\beta}} \tag{A.2}$$

Then their derivatives with respect to the logits are given by

$$\frac{\partial l_i}{\partial z_j} = \delta_{ij} - p_j = \delta_{ij} - e^{l_j} \tag{A.3}$$

$$\frac{\partial l_i^\beta}{\partial z_j} = \frac{1}{\beta}(\delta_{ij} - p_j^\beta) \tag{A.4}$$

Proof. The computation is straightforward, partial derivatives of log probabilities are computed as

$$\frac{\partial l_i}{\partial z_j} = \delta_{ij} - \frac{e^{z_j}}{\sum_k e^{z_k}} = \delta_{ij} - e^{l_j} = \delta_{ij} - p_j, \tag{A.5}$$

$$\frac{\partial l_i^\beta}{\partial z_j} = \frac{1}{\beta}\delta_{ij} - \frac{1}{\beta}\frac{e^{\frac{z_j}{\beta}}}{\sum_k e^{\frac{z_k}{\beta}}} = \frac{1}{\beta}(\delta_{ij} - p_j^\beta). \tag{A.6}$$

□

While the final loss gradient with respect to the model parameters is independent of whether intermediate derivatives are computed via l or p, practical implementations that omit the explicit calculation of p make derivatives with respect to l more informative. We illustrate this on the Shannon entropy example:

$$\mathcal{H} \doteq -p_i \log p_i = -p_i l_i. \tag{A.7}$$

Its gradients with respect to p are

$$\frac{\partial \mathcal{H}}{\partial p_i} = \frac{\partial}{\partial p_i}(-p_i \log p_i) = -\log p_i - 1. \tag{A.8}$$

Based on the eq. (A.8) in low probability setting when $p \to 0$ the Shannon entropy gradient diverges $\frac{\partial \mathcal{H}}{\partial p_i} \to \infty$. We now examine the gradients with respect to log probabilities, as these are the values calculated in practical implementations:

$$\frac{\partial \mathcal{H}}{\partial l_i} = \frac{\partial}{\partial l_i}(-p_i l_i) = -l_i \frac{\partial p_i}{\partial l_i} - p_i = -l_i p_i - p_i = -p_i \log p_i - p_i. \tag{A.9}$$

From the eq. (A.9) it is clear that there is no explosion as when $p \to 0$, the gradients are $\frac{\partial \mathcal{H}}{\partial l_i} \to 0$ due to the following limit:

$$\lim_{p \to 0}(lp - p) = \lim_{p \to 0}(p \log p - p) = 0. \tag{A.10}$$

The gradients remain robust throughout the rest of backpropagation chain:

$$\frac{\partial \mathcal{H}}{\partial z_j} = \sum_i \frac{\partial \mathcal{H}}{\partial l_i}\frac{\partial l_i}{\partial z_j} = \sum_i -(l_i p_i + p_i)(\delta_{ij} - p_j) = -l_j p_j + \sum_i l_i p_i p_j = -p_j\big(l_j - \sum_i l_i p_i\big) \tag{A.11}$$

This example demonstrates how using l rather than p helps avoid false assumptions regarding the method's behavior.

Proposition A.2 derives the formulation of the GEM loss used in Definition 2.1, translating the original objective into our notation and providing a simplified expression that is more comprehensible on sight.

Proposition A.2 (Reformulation of the original GEM from (Li et al., 2024) with our notation)**.** *Consider a model distribution p_θ, data distribution q and a temperature-scaled distribution $p_\theta^\beta = softmax(\beta^{-1} \log p_\theta)$. Consider y^{real} is the supervised label in the dataset and y^{gene} is the model-generated output. The original GEM formulation is provided in the following equation:*

$$\mathcal{L}_{\mathrm{GEM}}(\theta) = \sum_{y^{real} \sim q} \sum_{y^{gene} \sim p_\theta} \mathrm{sg}(p_\theta^\beta(y^{gene}))\left[\log p_\theta(y^{gene}) - \log p_\theta(y^{real})\right], \tag{A.12}$$

where $\beta \in (0, 1)$ is a temperature parameter and $\mathrm{sg}(p_\theta^\beta(y^{gene}))$ is detached from gradient computational graph. The loss (A.12) in our notation can be reformulated as

$$\mathcal{L}_{\mathrm{GEM}}(\theta) = -\mathbb{E}_q[\log p_\theta] + \mathbb{E}_{\mathrm{sg}(p_\theta^\beta(y^{gene}))}[\log p_\theta]. \tag{A.13}$$

Proof. Let us rewrite the eq. (A.12) with respect to the detached term:

$$\mathcal{L}_{\mathrm{GEM}}(\theta) = \mathbb{E}_{y^{\text{real}} \sim q}\mathbb{E}_{y^{\text{gene}} \sim \mathrm{sg}(p_\theta^\beta(y^{\text{gene}}))}\left[\log p_\theta(y^{\text{gene}}) - \log p_\theta(y^{\text{real}})\right] = \tag{A.14}$$

$$= \mathbb{E}_{y^{\text{real}} \sim q}\mathbb{E}_{y^{\text{gene}} \sim \mathrm{sg}(p_\theta^\beta(y^{\text{gene}}))}\left[\log p_\theta(y^{\text{gene}})\right] - \mathbb{E}_{y^{\text{real}} \sim q}\mathbb{E}_{y^{\text{gene}} \sim \mathrm{sg}(p_\theta^\beta(y^{\text{gene}}))}\left[\log p_\theta(y^{\text{real}})\right] = \tag{A.15}$$

$$= \mathbb{E}_{y^{\text{gene}} \sim \mathrm{sg}(p_\theta^\beta(y^{\text{gene}}))}\left[\log p_\theta(y^{\text{gene}})\right] - \mathbb{E}_{y^{\text{real}} \sim q}\left[\log p_\theta(y^{\text{real}})\right]. \tag{A.16}$$

Now, by disregarding notational y^{gene} and y^{real}, it is clear that GEM exactly matches the following:

$$\mathcal{L}_{\mathrm{GEM}}(\theta) = -\mathbb{E}_q[\log p_\theta] + \mathbb{E}_{\mathrm{sg}(p_\theta^\beta(y^{\text{gene}}))}[\log p_\theta]. \tag{A.17}$$

□

Theorem A.3 (Theorem 3.1)**.** *Training with $\mathcal{L}_{\mathrm{GEM}}$ is equivalent to training with temperature-scaled Cross-Entropy loss, as*

$$\nabla_\theta \mathcal{L}_{\mathrm{GEM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CE}}^\beta(\theta), \tag{A.18}$$

where $\mathcal{L}_{\mathrm{CE}}^\beta(\theta) \doteq -\beta \mathbb{E}_q \log p_\theta^\beta$.

Proof. The Cross-Entropy loss is defined as $\mathcal{L}_{\mathrm{CE}} = -q_i \log p_i = -q_i l_i$. Using the chain rule and Proposition A.1, Cross-Entropy gradients with respect to log probabilities and logits are

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial l_i} = -q_i, \tag{A.19}$$

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial z_j} = \sum_i \frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial l_i}\frac{\partial l_i}{\partial z_j} = -\sum_i q_i(\delta_{ij} - p_j) = p_j - q_j. \tag{A.20}$$

Similarly, gradients for $\mathcal{L}_{\text{GEM}} = -q_i \log p_i + p_i^\beta \log p_i$ are

$$\frac{\partial \mathcal{L}_{\text{GEM}}}{\partial z_j} = \sum_i \frac{\partial \mathcal{L}_{\text{GEM}}}{\partial l_i} \frac{\partial l_i}{\partial z_j} = \sum_i (p_i^\beta - q_i)(\delta_{ij} - p_j) = p_j^\beta - q_j. \tag{A.21}$$

Let us calculate the gradients for scaled Cross-Entropy $\mathcal{L}_{\text{CE}}^\beta = -\beta q_i \log p_i^\beta = -\beta q_i l_i^\beta$.

$$\frac{\partial \mathcal{L}_{\text{CE}}^\beta}{\partial z_j} = \sum_i \frac{\partial \mathcal{L}_{\text{CE}}^\beta}{\partial l_i^\beta} \frac{\partial l_i^\beta}{\partial z_j} = -\sum_i \beta q_i \cdot \frac{1}{\beta}(\delta_{ij} - p_j^\beta) = p_j^\beta - q_j. \tag{A.22}$$

Now, as $\nabla_\theta z$ does not depend on the loss function, we have

$$\nabla_\theta \mathcal{L}_{\text{GEM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{CE}}^\beta(\theta). \tag{A.23}$$

□

Proposition A.4 (Proposition 3.2)**.** *Assume that the target distribution q is one-hot. Then, for Focal Loss $\mathcal{L}_{\text{FL}}$ and Cross-Entropy $\mathcal{L}_{\text{CE}}$, the gradients satisfy*

$$\nabla_\theta \mathcal{L}_{\text{FL}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}_{\text{CE}}(\theta), \tag{A.24}$$

where $g(p, \gamma) = (1-p)^\gamma - \gamma p (1-p)^{\gamma-1} \log p$. Here and further $\hat{p}$ denotes the predicted probability assigned to the ground-truth token.

Proof. First let's consider q to be arbitrary. Then by the chain rule and Proposition A.1 the gradients of Focal Loss are

$$\frac{\partial \mathcal{L}_{\text{FL}}}{\partial l_i} = -(1-p_i)^\gamma q_i + \gamma (1-p_i)^{\gamma-1} q_i l_i p_i, \tag{A.25}$$

$$\frac{\partial \mathcal{L}_{\text{FL}}}{\partial z_j} = \sum_i \frac{\partial \mathcal{L}_{\text{FL}}}{\partial l_i} \frac{\partial l_i}{\partial z_j} = -(1-p_j)^\gamma q_j + \gamma (1-p_j)^{\gamma-1} q_j l_j p_j + \tag{A.26}$$

$$+ \sum_i ((1-p_i)^\gamma q_i p_j - \gamma (1-p_i)^{\gamma-1} q_i l_i p_i p_j) = \tag{A.27}$$

$$= -q_j ((1-p_j)^\gamma - \gamma (1-p_j)^{\gamma-1} l_j p_j) + p_j \sum_i ((1-p_i)^\gamma q_i - \gamma (1-p_i)^{\gamma-1} q_i l_i p_i). \tag{A.28}$$

In this case, the resulting gradients are not proportional to the gradients of the Cross-Entropy loss. However, if q is one-hot, meaning $q_k = 1$ and $q_{i \neq k} = 0$ for some k, then

$$\frac{\partial \mathcal{L}_{\text{FL}}}{\partial z_j} = \underbrace{((1-p_k)^\gamma - \gamma (1-p_k)^{\gamma-1} l_k p_k)}_{g(p_k, \gamma)} (p_j - q_j), \tag{A.29}$$

and therefore, as $\nabla_\theta z$ does not depend on the loss function and denoting $(1-p)^\gamma - \gamma p (1-p)^{\gamma-1} \log p$ as $g(p, \gamma)$, we have

$$\nabla_\theta \mathcal{L}_{\text{FL}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}_{\text{CE}}(\theta). \tag{A.30}$$

□

Corollary A.5 (Corollary 3.3)**.** *If the target distribution q is one-hot*

$$-\beta \nabla_\theta \mathbb{E}_q \left[(1 - p_\theta^\beta)^\gamma \log p_\theta^\beta \right] = g(\hat{p}_\theta^\beta, \gamma) \nabla_\theta \mathcal{L}_{\text{CE}}^\beta(\theta) \tag{A.31}$$

Proof. By the chain rule, Proposition A.1, and analogously to the proof of Proposition 3.2, if the target distribution q is one-hot

$$\frac{\partial}{\partial z_j} \sum_i (1 - p_i^\beta)^\gamma q_i \log p_i^\beta = \frac{1}{\beta} \underbrace{((1-p_k^\beta)^\gamma - \gamma (1-p_k^\beta)^{\gamma-1} l_k^\beta p_k^\beta)}_{g(\hat{p}_k^\beta, \gamma)} (q_j - p_j^\beta). \tag{A.32}$$

Therefore, as $\nabla_\theta z$ does not depend on the loss function

$$-\beta \nabla_\theta \mathbb{E}_q \left[(1 - p_\theta^\beta)^\gamma \log p_\theta^\beta \right] = g(\hat{p}_\theta^\beta, \gamma) \nabla_\theta \mathcal{L}_{\text{CE}}^\beta(\theta). \tag{A.33}$$

□

Corollary A.6 (Corollary 3.5)**.** *If the target distribution q is one-hot, TOFU gradients are proportional to the ones of the temperature-scaled CE:*

$$\nabla_\theta \mathcal{L}_{\text{TOFU}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}^{\beta}_{\text{CE}}(\theta) \tag{A.34}$$

Proof. As $g(\hat{p}_\theta, \gamma)$ is detached from gradients computation

$$\nabla_\theta \mathcal{L}_{\text{TOFU}}(\theta) = \nabla_\theta \text{sg}(g(\hat{p}_\theta, \gamma)) \mathcal{L}^{\beta}_{\text{CE}}(\theta) = g(\hat{p}_\theta, \gamma) \nabla_\theta \mathcal{L}^{\beta}_{\text{CE}}(\theta) \tag{A.35}$$

□

B Experimental details

In this section we provide comprehensive descriptions for the models, benchmarks and datasets used in our study. All the experiments were performed on a single A100 GPU.

B.1 Models

OLMo-2-13B OLMo-2-1124-13B (Groeneveld et al., 2024) from the Allen Institute for AI, trained on the Dolma (Soldaini et al., 2024) dataset for improved performance on tasks such as text generation and instruction following. This model is released under the Apache 2.0 license.

Mistral-12B Mistral NeMo (Mistral AI & NVIDIA, 2024) is trained jointly by Mistral AI and NVIDIA. It is designed for diverse tasks including text generation and instruction following. This model is released under the Apache 2.0 license.

Pythia-12B Pythia 12B (Biderman et al., 2023) is trained on the Pile (Gao et al., 2020) as a scientific tool for studying model functionality and interpretability rather than for deployment or human-facing interactions. This model is released under the Apache 2.0 license.

Llama-3.1-8B Llama-3.1 (Touvron et al., 2023) is released by Meta AI as an extension of the Llama-3 series. The model serves as a strong foundation for downstream fine-tuning and alignment methods, making it widely adopted in both research and applied settings. This model contains custom Llama-3.1 license [1].

Qwen-3-8B Qwen3 (Yang et al., 2025) is a family of large language models developed by Alibaba Cloud, designed to support general-purpose language understanding, reasoning, and instruction-following tasks. The model is trained on a diverse mixture of web, code, and domain-specific data. This model is released under the Apache 2.0 license.

Phi-4-14B Phi-4 (Abdin et al., 2024) is a model family developed by Microsoft, focusing on efficiency and strong reasoning capabilities under limited parameter budgets. It is trained using a carefully curated dataset that emphasizes high-quality, synthetic, and textbook-style data. This model is released under the MIT license.

Qwen2.5-Math Qwen2.5-Math (Yang et al., 2024) is a branch of math-specific large language models from Qwen family. The models in this series possess advanced mathematical reasoning capabilities, including Chain-of-Thought (CoT). This model is released under the Apache 2.0 license.

DeepSeek-Math-7B DeepSeek-Math (Shao et al., 2024) is a collection of models pre-trained on math-related tokens sourced from Common Crawl, together with natural language and code data for 500B tokens. This model is released under the MIT license.

Mathstral-7B Mathstral is an surrogate language model, specializing in STEM subjects and derived from Mistral family (Jiang et al., 2023). This model is released under the Apache 2.0 license.

[1] https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE

B.2 SFT DATASETS

Alpaca Alpaca[2] (Taori et al., 2023) is a widely used instruction-following dataset consisting of approximately 52K instruction–response pairs generated using a self-instruct framework. The dataset covers a broad range of tasks, including question answering, summarization, reasoning, and creative writing. To preprocess the Alpaca dataset, we filter and format each example into prompt–completion pairs. We use explicit delimiters for the instruction, input, and response to provide structural context for the sequence. The model is trained to generate the response following an opening delimiter and is explicitly required to produce a matching delimiter to signal completion. Alpaca is available under the CC-BY-NC-4.0 license.

UltraFeedback UltraFeedback (Cui et al., 2023) is a preference-annotated dataset hosted on Hugging Face by openbmb, derived from the UltraFeedback corpus and adapted for supervised fine-tuning and instruction tuning. The dataset comprises 64K samples, each containing an AI-generated judgment that identifies a preferred response. Our pipeline filters the UltraFeedback dataset for top-quality examples, retaining only the highest-scoring completion per prompt, provided that it meets a minimum score of 7, resulting in 57,400 samples. We format these pairs using fixed templates the same as for the Alpaca dataset. UltraFeedback is available under the MIT license.

NuminaMath CoT NuminaMath (LI et al., 2024) is a dataset of approximately 860,000 mathematical problems, designed to examine the model's reasoning in Chain-of-Thought (CoT) manner. The dataset covers a wide range of tasks, from Chinese high school math exercises to US and international mathematics olympiad competition problems. We filtered 100,000 examples including all sources, ensuring that the solution is given in a bounding box format. This dataset is available under the Apache 2.0 license.

OpenCode-Instruct OpenCode (Ahmad et al., 2025) is the largest open-access instruction tuning dataset, comprising 5 million diverse samples. OpenCode-Instruct is designed for supervised fine-tuning (SFT). For our pipeline we filtered 50,000 Python problems, ensuring that the solution is given in a default format. This dataset is available under the CC-BY-4.0 license.

B.3 SFT DETAILS

Given the limitation of our computational resources, we performed 4-bit NormalFloat quantization of selected models and utilized the Quantized Low Rank Adaptation (Dettmers et al., 2023) technique to optimize our workflow. On top of that, we used gradient accumulation to increase the total batch size.

All models were trained for a single epoch using a linear learning rate schedule with a peak of 2×10^{-4} and 50 warmup steps. We employed a batch size of 2 with 4 gradient accumulation steps and a weight decay of 0.01. For the LoRA adapter, we set $r = 16$ and $\alpha = 16$.

B.4 EVALUATION DATASETS

ROC Stories is a continuation benchmark where a model is given a story beginning and must generate a coherent conclusion. We constructed this dataset by randomly selecting 100 different stories from the ROCStories corpus (Mostafazadeh et al., 2016), each containing exactly five sentences. For our evaluation, the first four sentences serve as the beginning of the story, providing sufficient context for a logical continuation.

Alpaca Eval benchmark is a collection of short questions from the *helpful_base* subset of the AlpacaFarm (Dubois et al., 2023) Hugging Face repository, comprising 129 prompts. We take only this portion of the original dataset to ensure that the evaluation remains focused on standard natural language (in opposition to code or ASCII symbol drawings), as the Self-BLEU metric does not function reliably outside of this domain. Alpaca Eval is available under the CC-BY-NC-4.0.

NoveltyBench (Zhang et al., 2025) is a benchmark designed to evaluate language models' ability to generate multiple distinct and high-quality outputs for the same prompt, removing the traditional focus from a single best response. For the evaluation, we selected its *NB-curated* subset, which

[2] https://huggingface.co/datasets/tatsu-lab/alpaca

contains 100 manually curated prompts. We utilized the original NoveltyBench framework and source code, including the default parameters for their proprietary quality and diversity metrics, Utility-k and Distinct-k. The code is available under the MIT license.

Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020) is a benchmark designed to evaluate the knowledge and reasoning capabilities of language models across multiple subject areas, spanning STEM disciplines, humanities, and social sciences. The dataset includes questions of varying difficulty levels, ranging from elementary concepts to advanced professional knowledge. For the benchmarking, we used its test subset, comprising 14042 questions. This dataset is available under the MIT license.

ARC-Challenge (ARC) (Clark et al., 2018) is a benchmark dataset of multiple-choice science questions curated to evaluate advanced reasoning and scientific understanding. The questions are sourced from standardized science examinations for grades 3 through 9 and are intentionally selected to be challenging for both humans and AI systems. For the benchmarking, we used its test subset, comprising 1172 questions. This dataset is available under the CC-BY-SA-4.0 license.

MATH500 is an open-source test subsample of the original MATH dataset (Hendrycks et al., 2021). It comprises 500 problems alongside their solutions in various subjects, such as algebra, geometry, calculus, and probability. The code accompanying this dataset is available under the MIT license.

MinervaMath is a publicly available subset of Minerva corpus (Lewkowycz et al., 2022), consisting of 272 mathematical problems related to natural sciences. This dataset is available under the MIT license.

GSM8K is a collection of 8,800 high quality linguistically diverse grade school math word problems (Cobbe et al., 2021). For the evaluation we used a main test set of 1,320 tasks. This dataset is available under the MIT license.

OlympiadBench is a dataset of 8,476 challenging olympiad-level problems in math and physics (He et al., 2024). For the evaluation we used a main test set of 674 tasks. This dataset is available under the Apache-2.0 license.

HumanEval is a compilation of 164 programming problems, released by OpenAI (Chen et al., 2021), with several unit tests curated by humans. This dataset is available under the MIT license.

Malicious Instruct is a human crafted dataset (Huang et al., 2023), consisting of 100 prompts. To construct the datasets the authors selected ten categories and asked ChatGPT to provide 20 responses for each of the categories. They manually reviewed the generated responses and selected 100 responses such that they are aligned with the topic and diverse at the same time.

HarmBench is a standardized evaluation framework for automated red teaming (Mazeika et al., 2024), consisting of 400 malicious prompts in different categories: copyright, contextual, and standard. For our main experiments, we randomly selected 100 prompts from the standard subset. This dataset is available under the MIT license.

B.5 Inference details

The model's generation parameters were selected based on the specific requirements of each evaluation task. Following standard empirical practices in the field, we employed a stochastic sampling strategy for the creative writing & instruction following, mathematical reasoning and safety benchmarks. Specifically, we used nucleus sampling (Holtzman et al., 2020) with a cumulative probability threshold of $p = 0.9$ and different unit temperatures, depending on the task.

For instruction following and safety we employed $T = 1.0$ for all the training objectives, while in mathematical reasoning experiments each loss utilizes a specific temperature. While we maintained $T = 1.0$ for CE, GEM and TOFU exhibited high performance variance at this default, leading to non-robust evaluations where per-run fluctuations determined the top-performing method. This instability is inherent to the reasoning task, where a single incorrect token can derail the entire Chain-of-Thought, and minor deviations from the required format result in the response being classified as incorrect. Therefore, we empirically derived $T = 0.3$ such that the variance of the coverage across the inference runs for each objective stays in range of 1%. For creative writing, instruction following, and safety benchmarks we generated 10 responses per prompt, and for mathematical reasoning we

employed 16 decoding runs. In contrast, for the ARC and MMLU, we used deterministic greedy decoding to ensure objective and reproducible outputs, generating a single completion per prompt.

Due to computational constraints, the maximum response length was restricted across all the experiments. We allocated a limit of 64 tokens for creative benchmarks (ROC Stories, Alpaca Eval, Novelty Bench) and 8 tokens for multiple-choice ones (ARC, MMLU). For Malicious Instruct and HarmBench datasets we used limit of 128 tokens. For experiments with mathematical reasoning in a Chain-of-Thought (CoT) manner, we set the maximal generation length of 4096 tokens.

For the inference of SFT models, we adhere to the standard instruction template, incorporating an additional prompt if necessary, depending on the benchmark. Regarding the base models, we observed that they successfully continue the narrative on the ROC benchmark when provided with substantial initial context, even without explicit instructions. This enables a direct comparison with their fine-tuned counterparts. For the AE benchmark, we follow the protocol established by O'Mahony et al. (2024). All prompt templates are provided in the accompanying code repository.

B.6 LLM JUDGE

To assess the quality of the responses, we employ large language model as a judge to score and compare generated responses on ROC Stories and Alpaca Eval datasets. Specifically, we use **Llama-3.1-70B-Instruct** (Grattafiori et al., 2024), an instruction-aligned model. To reduce memory footprint, we quantized the Judge with 4-bit NormalFloat. We prompted the Judge with comprehensive instructions to provide a score from 0 to 5, where 0 corresponds to an incoherent, off-topic, or nonsensical response, and 5 corresponds to a seamless, natural, and stylistically consistent one. Given the restriction in tokens that we applied in the inference stage, we explicitly state in the Judge instruction not to penalize the response if it ends abruptly due to the aforementioned limit. However, it is required to penalize a logically or stylistically flawed ending. To ensure a robust evaluation, we used greedy decoding and strict response template.

C ABLATIONS

We find best parameters for Focal Loss and TOFU objectives by evaluating quality and diversity on NoveltyBench and accuracy on ARC. For these experiments we employ Mistral-12B model. We tested the following ranges of parameters: $\gamma \in [2, 5]$ and $\beta \in [0.6, 0.9]$. According to the results, gathered in Table C.1, the best hyperparameters are $\gamma = 3$ for Focal Loss and $\gamma = 3, \beta = 0.8$ for TOFU. While we selected the optimal values primarily based on ARC accuracy, they also coincide with the top-performing configurations for NoveltyBench Utility. In contrast with the rest of the experiments on NoveltyBench, here we generated only 5 responses per prompt.

Table C.1: ARC Accuracy, NoveltyBench Distinct (1-5) and Utility (1-5) results for the Mistral-12B model fine-tuned with Focal Loss and TOFU with different (γ, β) configurations.

	Focal Loss				TOFU											
$\boldsymbol{\beta}$	-	-	-	-	0.6	0.6	0.6	0.7	0.7	0.7	0.8	**0.8**	0.8	0.9	0.9	0.9
$\boldsymbol{\gamma}$	2	**3**	4	5	2	3	5	2	3	5	2	**3**	5	2	3	5
Accuracy ↑	72.9	**75.0**	74.3	74.2	72.0	74.4	75.4	73.1	73.8	74.6	75.3	**75.6**	74.5	71.6	74.7	74.6
Distinct ↑	3.97	**4.23**	4.44	4.37	4.64	4.69	4.76	4.46	4.40	4.69	4.25	**4.41**	4.54	4.18	4.36	4.45
Utility ↑	4.61	**4.90**	4.81	4.29	4.30	4.13	3.97	4.53	4.37	4.17	4.72	**4.71**	4.22	4.76	4.65	4.16

D ADDITIONAL RESULTS

To verify the robustness of our results across different SFT datasets, we replicated the experiments conducted on Alpaca using UltraFeedback. While the lower overall quality of UltraFeedback negatively affects performance, the trends observed in the Alpaca experiments remain preserved. The corresponding results for ROC Stories and Alpaca Eval are provided in Table D.2. For diversity metrics, we report mean values and standard deviations calculated across the prompts. For quality,

Table D.1: Performance of models across Alpaca SFT objectives on ROC Stories and Alpaca Eval. Diversity (D) is measured via Self-BLEU (0–100), where lower scores are better. Quality (Q) is measured via LLM Judge score (0–5), where higher scores are better.

Bench	Method	Mistral-12B		OLMo-2-13B		Pythia-12B		Llama-3.1-8B		Qwen-3-8B		Phi-4-14B	
		D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑
ROC	Base	$11.4_{\pm 4.7}$	$3.9_{\pm 0.5}$	$12.3_{\pm 4.6}$	$3.7_{\pm 0.7}$	$9.3_{\pm 3.1}$	$2.9_{\pm 0.6}$	$11.5_{\pm 6.1}$	$3.5_{\pm 0.7}$	$24.9_{\pm 8.9}$	$2.9_{\pm 1.0}$	$13.8_{\pm 6.8}$	$3.4_{\pm 0.9}$
	CE	$22.5_{\pm 8.3}$	$4.8_{\pm 0.2}$	$24.5_{\pm 9.1}$	$4.8_{\pm 0.2}$	$23.8_{\pm 8.4}$	$3.7_{\pm 0.6}$	$22.6_{\pm 8.5}$	$4.7_{\pm 0.2}$	$22.3_{\pm 7.7}$	$4.6_{\pm 0.3}$	$24.8_{\pm 9.0}$	$4.8_{\pm 0.2}$
	λ-PR	$3.7_{\pm 0.7}$	$2.5_{\pm 0.5}$	$3.9_{\pm 0.6}$	$3.0_{\pm 0.4}$	$4.1_{\pm 0.7}$	$2.1_{\pm 0.5}$	$3.8_{\pm 0.7}$	$2.7_{\pm 0.5}$	$5.0_{\pm 1.2}$	$3.4_{\pm 0.4}$	$4.0_{\pm 0.6}$	$3.1_{\pm 0.4}$
	FL	$16.8_{\pm 5.4}$	$4.7_{\pm 0.2}$	$15.9_{\pm 5.4}$	$4.7_{\pm 0.2}$	$14.6_{\pm 4.3}$	$3.8_{\pm 0.6}$	$14.3_{\pm 4.7}$	$4.5_{\pm 0.3}$	$16.4_{\pm 6.2}$	$4.6_{\pm 0.3}$	$16.2_{\pm 5.4}$	$4.7_{\pm 0.3}$
	GEM	$14.3_{\pm 6.0}$	$4.6_{\pm 0.3}$	$13.2_{\pm 5.0}$	$4.6_{\pm 0.3}$	$13.0_{\pm 6.2}$	$3.4_{\pm 0.6}$	$11.5_{\pm 4.0}$	$4.5_{\pm 0.3}$	$13.9_{\pm 4.6}$	$4.5_{\pm 0.3}$	$13.4_{\pm 5.1}$	$4.7_{\pm 0.3}$
	TOFU	$\mathbf{12.7}_{\pm 4.7}$	$4.6_{\pm 0.3}$	$\mathbf{11.9}_{\pm 4.1}$	$4.6_{\pm 0.3}$	$\mathbf{10.8}_{\pm 3.7}$	$3.6_{\pm 0.5}$	$\mathbf{11.2}_{\pm 4.2}$	$4.5_{\pm 0.3}$	$\mathbf{12.7}_{\pm 3.8}$	$4.5_{\pm 0.3}$	$\mathbf{12.4}_{\pm 4.1}$	$4.6_{\pm 0.3}$
AE	Base	$12.7_{\pm 7.0}$	$3.8_{\pm 0.9}$	$13.9_{\pm 7.2}$	$3.8_{\pm 1.0}$	$8.4_{\pm 3.0}$	$2.6_{\pm 1.0}$	$12.2_{\pm 6.0}$	$3.5_{\pm 0.9}$	$31.5_{\pm 12.5}$	$3.8_{\pm 0.9}$	$17.6_{\pm 9.3}$	$3.9_{\pm 0.9}$
	CE	$44.5_{\pm 14.5}$	$4.2_{\pm 0.7}$	$45.0_{\pm 14.6}$	$4.3_{\pm 0.7}$	$35.9_{\pm 14.4}$	$3.8_{\pm 0.8}$	$44.5_{\pm 14.4}$	$4.2_{\pm 0.7}$	$44.9_{\pm 13.4}$	$4.1_{\pm 0.7}$	$46.6_{\pm 15.0}$	$4.2_{\pm 0.7}$
	λ-PR	$2.9_{\pm 0.8}$	$2.0_{\pm 0.7}$	$3.2_{\pm 0.9}$	$2.3_{\pm 0.7}$	$3.0_{\pm 1.1}$	$2.0_{\pm 0.6}$	$2.8_{\pm 0.8}$	$2.1_{\pm 0.7}$	$4.6_{\pm 2.1}$	$2.6_{\pm 0.7}$	$3.3_{\pm 1.0}$	$2.4_{\pm 0.6}$
	FL	$29.3_{\pm 11.1}$	$4.1_{\pm 0.7}$	$29.0_{\pm 9.8}$	$4.2_{\pm 0.7}$	$22.2_{\pm 9.8}$	$3.6_{\pm 0.8}$	$28.3_{\pm 10.5}$	$4.1_{\pm 0.7}$	$29.2_{\pm 10.6}$	$4.0_{\pm 0.7}$	$28.9_{\pm 10.4}$	$4.2_{\pm 0.7}$
	GEM	$27.7_{\pm 12.6}$	$4.1_{\pm 0.7}$	$27.0_{\pm 11.6}$	$4.1_{\pm 0.7}$	$17.6_{\pm 7.6}$	$3.5_{\pm 0.7}$	$25.9_{\pm 12.3}$	$4.1_{\pm 0.7}$	$29.5_{\pm 11.5}$	$4.0_{\pm 0.8}$	$27.5_{\pm 12.6}$	$4.1_{\pm 0.7}$
	TOFU	$\mathbf{21.2}_{\pm 8.8}$	$4.0_{\pm 0.6}$	$\mathbf{21.3}_{\pm 8.6}$	$4.1_{\pm 0.6}$	$\mathbf{16.1}_{\pm 7.6}$	$3.5_{\pm 0.8}$	$\mathbf{20.9}_{\pm 8.4}$	$4.0_{\pm 0.7}$	$\mathbf{24.1}_{\pm 8.9}$	$4.0_{\pm 0.7}$	$\mathbf{21.8}_{\pm 8.7}$	$4.1_{\pm 0.6}$

we first compute the average score per prompt, then report the global mean and the standard deviation of those per-prompt averages.

Table D.2: Performance of models across UltraFeedback SFT objectives on ROC Stories and Alpaca Eval. Diversity (D) is measured via Self-BLEU (0–100), where lower scores are better. Quality (Q) is measured via LLM Judge score (0–5), where higher scores are better.

Bench	Method	Mistral-12B		OLMo-2-13B		Pythia-12B		Llama-3.1-8B		Qwen-3-8B		Phi-4-14B	
		D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑
ROC	Base	$11.4_{\pm 4.7}$	$3.9_{\pm 0.5}$	$12.3_{\pm 4.6}$	$3.7_{\pm 0.7}$	$9.3_{\pm 3.1}$	$2.9_{\pm 0.6}$	$11.5_{\pm 6.1}$	$3.5_{\pm 0.7}$	$24.9_{\pm 8.9}$	$2.9_{\pm 1.0}$	$13.8_{\pm 6.8}$	$3.4_{\pm 0.9}$
	CE	$26.9_{\pm 9.8}$	$4.4_{\pm 0.4}$	$24.0_{\pm 8.7}$	$4.4_{\pm 0.5}$	$19.7_{\pm 7.4}$	$3.1_{\pm 0.7}$	$23.9_{\pm 8.8}$	$4.2_{\pm 0.5}$	$25.6_{\pm 12.0}$	$4.2_{\pm 0.5}$	$24.4_{\pm 7.9}$	$4.5_{\pm 0.3}$
	λ-PR	$3.3_{\pm 0.5}$	$1.8_{\pm 0.5}$	$3.2_{\pm 0.4}$	$2.1_{\pm 0.5}$	$3.6_{\pm 0.7}$	$1.2_{\pm 0.5}$	$3.3_{\pm 0.5}$	$1.9_{\pm 0.5}$	$4.0_{\pm 0.8}$	$2.6_{\pm 0.5}$	$3.4_{\pm 0.5}$	$2.2_{\pm 0.5}$
	FL	$15.3_{\pm 4.7}$	$4.4_{\pm 0.3}$	$14.3_{\pm 4.6}$	$4.3_{\pm 0.4}$	$14.9_{\pm 4.7}$	$2.7_{\pm 0.6}$	$14.9_{\pm 5.4}$	$4.1_{\pm 0.4}$	$15.1_{\pm 5.7}$	$4.2_{\pm 0.4}$	$15.5_{\pm 5.4}$	$4.4_{\pm 0.3}$
	GEM	$13.1_{\pm 5.3}$	$4.2_{\pm 0.4}$	$12.1_{\pm 4.3}$	$4.1_{\pm 0.5}$	$9.8_{\pm 3.6}$	$2.6_{\pm 0.7}$	$12.0_{\pm 4.8}$	$4.0_{\pm 0.4}$	$13.4_{\pm 6.5}$	$4.1_{\pm 0.4}$	$13.1_{\pm 5.8}$	$4.3_{\pm 0.4}$
	TOFU	$\mathbf{12.1}_{\pm 4.6}$	$4.2_{\pm 0.3}$	$\mathbf{11.2}_{\pm 3.5}$	$4.2_{\pm 0.4}$	$\mathbf{9.5}_{\pm 3.0}$	$2.9_{\pm 0.6}$	$\mathbf{11.4}_{\pm 3.7}$	$4.0_{\pm 0.4}$	$\mathbf{12.3}_{\pm 4.7}$	$4.1_{\pm 0.4}$	$\mathbf{11.8}_{\pm 4.3}$	$4.3_{\pm 0.4}$
AE	Base	$12.7_{\pm 7.0}$	$3.8_{\pm 0.9}$	$13.9_{\pm 7.2}$	$3.8_{\pm 1.0}$	$8.4_{\pm 3.0}$	$2.6_{\pm 1.0}$	$12.2_{\pm 6.0}$	$3.5_{\pm 0.9}$	$31.5_{\pm 12.5}$	$3.8_{\pm 0.9}$	$17.6_{\pm 9.3}$	$3.9_{\pm 0.9}$
	CE	$45.1_{\pm 10.3}$	$3.8_{\pm 0.5}$	$43.9_{\pm 9.6}$	$3.8_{\pm 0.5}$	$32.0_{\pm 9.0}$	$3.2_{\pm 0.6}$	$42.8_{\pm 10.0}$	$3.8_{\pm 0.5}$	$44.3_{\pm 9.8}$	$3.7_{\pm 0.6}$	$48.0_{\pm 9.9}$	$3.8_{\pm 0.5}$
	λ-PR	$2.4_{\pm 0.5}$	$1.3_{\pm 0.6}$	$2.5_{\pm 0.6}$	$1.4_{\pm 0.6}$	$2.5_{\pm 0.4}$	$1.3_{\pm 0.5}$	$2.6_{\pm 0.5}$	$1.3_{\pm 0.5}$	$3.4_{\pm 0.8}$	$1.7_{\pm 0.6}$	$2.8_{\pm 0.7}$	$1.4_{\pm 0.5}$
	FL	$27.8_{\pm 7.1}$	$3.6_{\pm 0.5}$	$26.8_{\pm 7.0}$	$3.6_{\pm 0.5}$	$20.6_{\pm 5.8}$	$3.0_{\pm 0.6}$	$26.0_{\pm 7.0}$	$3.6_{\pm 0.5}$	$28.3_{\pm 7.5}$	$3.5_{\pm 0.5}$	$27.9_{\pm 7.3}$	$3.6_{\pm 0.5}$
	GEM	$26.3_{\pm 8.0}$	$3.6_{\pm 0.5}$	$25.1_{\pm 7.3}$	$3.5_{\pm 0.5}$	$16.7_{\pm 5.7}$	$2.8_{\pm 0.6}$	$24.2_{\pm 6.9}$	$3.5_{\pm 0.5}$	$28.4_{\pm 8.5}$	$3.5_{\pm 0.5}$	$28.1_{\pm 8.0}$	$3.6_{\pm 0.5}$
	TOFU	$\mathbf{20.9}_{\pm 5.7}$	$3.5_{\pm 0.6}$	$\mathbf{20.2}_{\pm 5.7}$	$3.5_{\pm 0.5}$	$\mathbf{14.6}_{\pm 4.9}$	$2.8_{\pm 0.6}$	$\mathbf{19.8}_{\pm 5.7}$	$3.5_{\pm 0.5}$	$\mathbf{22.5}_{\pm 6.2}$	$3.5_{\pm 0.5}$	$\mathbf{21.1}_{\pm 6.2}$	$3.5_{\pm 0.5}$

While the primary figures for NoveltyBench are presented in the main text, the corresponding raw values are provided in Table D.3. We report the mean values and standard deviations calculated across the prompts. We note that the Utility metric exhibits a high standard deviation, in some cases exceeding half of the mean value. This significant variance may stem from the diverse difficulty levels of the prompts within the benchmark, or potentially from a lack of robustness in the judge used for evaluation.

To explore the potential longer-horizon collapse that may occur in longer generations, we conducted additional experiments on ROC Stories and Alpaca Eval by employing Mistral-12B, OLMo-2-13B, and Llama-3.1-8B with a limit of 256 and 512 tokens (see Table D.7 and Table D.8). Our results are in line with Table D.1 showing that models fine-tuned with TOFU consistently outperform other approaches across all tested lengths.

We additionally performed SFT on a non-quantized Llama-3.1-8B and Qwen3-8B and evaluated the model's performance on ROC Stories and Alpaca Eval to ensure that employed quantization does not affect the consistency of our results. Then we compared a quantized model against non-quantized

Table D.3: Performance of models across UltraFeedback and Alpaca SFT objectives on NoveltyBench. (D) Distinct (1–10) measures responses diversity, while (U) Utility (1–10) represents quality. For both metrics, higher values indicate superior performance.

SFT	Method	Mistral-12B		OLMo-2-13B		Pythia-12B		Llama-3.1-8B		Qwen-3-8B		Phi-4-14B	
		D↑	U↑	D↑	U↑	D↑	U↑	D↑	U↑	D↑	U↑	D↑	U↑
Alpaca	CE	$6.6_{\pm 2.7}$	$4.3_{\pm 2.4}$	$6.5_{\pm 2.8}$	$4.2_{\pm 2.5}$	$7.7_{\pm 2.3}$	$3.2_{\pm 2.2}$	$6.5_{\pm 2.7}$	$4.2_{\pm 2.4}$	$6.5_{\pm 2.7}$	$4.2_{\pm 2.4}$	$5.8_{\pm 2.8}$	$1.7_{\pm 1.2}$
	λ-PR	$9.5_{\pm 1.2}$	$1.3_{\pm 0.5}$	$9.4_{\pm 0.4}$	$1.3_{\pm 0.6}$	$9.3_{\pm 1.4}$	$1.1_{\pm 0.4}$	$9.5_{\pm 1.4}$	$1.3_{\pm 0.6}$	$9.1_{\pm 1.7}$	$1.2_{\pm 0.5}$	$9.6_{\pm 1.2}$	$1.2_{\pm 0.5}$
	FL	$7.4_{\pm 2.4}$	$4.3_{\pm 2.3}$	$7.7_{\pm 2.6}$	$4.5_{\pm 2.5}$	$8.5_{\pm 2.0}$	$3.2_{\pm 2.0}$	$7.7_{\pm 2.5}$	$4.2_{\pm 2.4}$	$7.4_{\pm 2.5}$	$4.3_{\pm 2.3}$	$7.1_{\pm 2.5}$	$2.0_{\pm 1.3}$
	GEM	$8.0_{\pm 2.4}$	$4.5_{\pm 2.5}$	$8.1_{\pm 2.2}$	$4.6_{\pm 2.5}$	$8.7_{\pm 1.8}$	$3.0_{\pm 2.0}$	$8.1_{\pm 2.2}$	$4.7_{\pm 2.6}$	$7.6_{\pm 2.4}$	$4.3_{\pm 2.5}$	$7.2_{\pm 2.6}$	$1.8_{\pm 1.2}$
	TOFU	**8.2**$_{\pm 2.2}$	$4.4_{\pm 2.4}$	**8.3**$_{\pm 2.1}$	$4.4_{\pm 2.4}$	**8.8**$_{\pm 1.9}$	$3.0_{\pm 2.0}$	**8.3**$_{\pm 2.1}$	$4.3_{\pm 2.4}$	**7.9**$_{\pm 2.5}$	$4.2_{\pm 2.4}$	**7.5**$_{\pm 2.5}$	$2.0_{\pm 1.3}$
UF	**CE**	$7.3_{\pm 2.4}$	$3.2_{\pm 2.2}$	$7.3_{\pm 2.5}$	$3.0_{\pm 2.1}$	$8.5_{\pm 1.8}$	$1.7_{\pm 1.1}$	$7.6_{\pm 2.3}$	$2.9_{\pm 1.9}$	$7.0_{\pm 2.5}$	$2.9_{\pm 2.0}$	$6.9_{\pm 2.4}$	$2.2_{\pm 1.3}$
	λ-PR	$10.0_{\pm 0.2}$	$1.1_{\pm 0.3}$	$9.9_{\pm 0.5}$	$1.1_{\pm 0.2}$	$9.9_{\pm 0.5}$	$1.0_{\pm 0.2}$	$10.0_{\pm 0.2}$	$1.1_{\pm 0.3}$	$9.9_{\pm 0.5}$	$1.1_{\pm 0.4}$	$9.9_{\pm 0.3}$	$1.1_{\pm 0.3}$
	FL	$8.5_{\pm 1.9}$	$2.9_{\pm 1.9}$	$8.8_{\pm 1.5}$	$3.0_{\pm 1.9}$	$9.3_{\pm 1.1}$	$1.6_{\pm 0.8}$	$8.9_{\pm 1.6}$	$2.8_{\pm 1.8}$	$8.2_{\pm 1.6}$	$2.9_{\pm 1.9}$	$8.4_{\pm 1.8}$	$2.3_{\pm 1.2}$
	GEM	$9.0_{\pm 1.6}$	$2.8_{\pm 1.8}$	$9.1_{\pm 1.3}$	$2.8_{\pm 1.9}$	$9.5_{\pm 0.8}$	$1.4_{\pm 0.7}$	$9.1_{\pm 1.2}$	$2.6_{\pm 1.6}$	$8.8_{\pm 1.8}$	$2.8_{\pm 1.9}$	$8.6_{\pm 1.6}$	$2.0_{\pm 1.2}$
	TOFU	**9.1**$_{\pm 1.2}$	$2.6_{\pm 1.7}$	**9.1**$_{\pm 1.5}$	$2.9_{\pm 1.9}$	**9.5**$_{\pm 0.9}$	$1.4_{\pm 0.7}$	**9.3**$_{\pm 1.3}$	$2.6_{\pm 1.8}$	**8.8**$_{\pm 1.5}$	$2.8_{\pm 1.9}$	**8.9**$_{\pm 1.4}$	$2.0_{\pm 1.2}$

Table D.4: Performance of quantized SFT models vs non-quantized across Alpaca SFT objectives on ROC Stories (ROC) and Alpaca Eval (AE).

Bench	Method	Llama-3.1-8B		Llama-3.1-8B-4bit		Qwen3-8B		Qwen3-8B-4bit	
		D↓	Q↑	D↓	Q↑	D↓	Q↑	D↓	Q↑
ROC	CE	$23.2_{\pm 8.2}$	$4.8_{\pm 0.2}$	$22.6_{\pm 8.5}$	$4.7_{\pm 0.3}$	$22.1_{\pm 6.5}$	$4.6_{\pm 0.2}$	$22.3_{\pm 7.7}$	$4.6_{\pm 0.3}$
	FL	$15.5_{\pm 5.0}$	$4.6_{\pm 0.3}$	$14.3_{\pm 4.7}$	$4.5_{\pm 0.3}$	$16.0_{\pm 5.5}$	$4.6_{\pm 0.3}$	$16.4_{\pm 6.2}$	$4.6_{\pm 0.3}$
	GEM	$12.0_{\pm 4.4}$	$4.5_{\pm 0.3}$	$11.5_{\pm 4.0}$	$4.5_{\pm 0.2}$	$13.4_{\pm 4.2}$	$4.5_{\pm 0.2}$	$13.9_{\pm 4.6}$	$4.5_{\pm 0.3}$
	TOFU	**11.5**$_{\pm 4.3}$	$4.5_{\pm 0.3}$	**11.2**$_{\pm 4.2}$	$4.5_{\pm 0.3}$	**12.5**$_{\pm 3.5}$	$4.5_{\pm 0.3}$	**12.7**$_{\pm 3.8}$	$4.5_{\pm 0.3}$
AE	CE	$52.7_{\pm 15.8}$	$4.2_{\pm 0.7}$	$44.5_{\pm 14.3}$	$4.2_{\pm 0.7}$	$44.5_{\pm 13.0}$	$4.2_{\pm 0.5}$	$44.9_{\pm 13.4}$	$4.1_{\pm 0.7}$
	FL	$32.8_{\pm 13.1}$	$4.2_{\pm 0.7}$	$28.3_{\pm 8.8}$	$4.1_{\pm 0.7}$	$28.0_{\pm 10.0}$	$4.0_{\pm 0.6}$	$29.2_{\pm 10.6}$	$4.0_{\pm 0.7}$
	GEM	$33.6_{\pm 14.7}$	$4.1_{\pm 0.7}$	$25.9_{\pm 12.3}$	$4.1_{\pm 0.7}$	$27.7_{\pm 9.5}$	$4.0_{\pm 0.7}$	$29.5_{\pm 11.5}$	$4.0_{\pm 0.8}$
	TOFU	**26.2**$_{\pm 11.2}$	$4.1_{\pm 0.6}$	**20.8**$_{\pm 8.4}$	$4.0_{\pm 0.7}$	**22.2**$_{\pm 8.0}$	$4.0_{\pm 0.5}$	**24.1**$_{\pm 8.9}$	$4.0_{\pm 0.7}$

across different SFT objectives and gathered the results in Table D.4. Ultimately, we observe the same pattern as in the main results across all benchmarks — models fine-tuned with TOFU consistently outperform other SFT approaches.

As a qualitative illustration of induced diversity in creative writing, we compare the outputs of Mistral-12B fine-tuned on Alpaca using Cross-Entropy versus TOFU loss, the results are gathered in Figure D.1. When prompted to tell a funny joke, the model trained with CE tends toward redundancy, often generating very similar or nearly identical responses. In contrast, the TOFU-tuned version maintains significantly more variety between its completions.

Additionally, we employ mechanistic validation, demonstrating TOFU's successful impact on simultaneously reducing ignorance and forgetting (see Figure D.2). Specifically, we ask the model to generate an integer between 1 and 5 and track the output probabilities of the top-ranked tokens. The critical comparison is between the resulting distributions of CE and TOFU, where TOFU improves the chances of correct answers without suppressing legitimate generative breadth.

Table D.5: The performance of Alpaca SFT and UltraFeedback SFT models across different objectives on ARC (first column) and MMLU (second column). Values are measured as Accuracy scores (0-100).

Bench	Method	Mistral-12B		OLMo-2-13B		Pythia-12B		Llama-3.1-8B		Qwen-3-8B		Phi-4-14B	
		ARC	MMLU	ARC	MMLU	ARC	MMLU	ARC	MMLU	ARC	MMLU	ARC	MMLU
Alpaca	CE	74.6	58.3	73.5	58.9	25.5	25.9	71.4	57.8	86.9	69.9	86.1	72.0
	FL	76.8	59.4	74.0	58.1	25.1	26.1	71.2	58.0	86.9	69.4	85.7	71.0
	GEM	76.8	59.4	74.2	58.6	23.5	25.2	72.1	58.3	87.0	69.7	86.2	71.9
	λ-PR	61.6	50.4	62.4	39.2	21.3	18.3	55.4	47.6	86.1	68.2	86.7	71.7
	TOFU	73.9	57.8	74.6	58.7	24.9	26.0	72.5	57.3	86.9	69.2	85.8	71.9
UF	CE	76.3	59.8	74.4	59.8	23.6	21.8	71.8	58.3	88.1	70.5	88.7	74.8
	FL	74.3	58.7	73.7	59.3	25.2	22.7	71.1	58.5	88.3	70.3	89.0	75.2
	GEM	77.3	59.8	74.1	59.6	27.1	21.5	71.8	58.6	87.9	70.2	88.9	74.9
	λ-PR	37.7	32.1	14.4	0.3	20.1	13.4	48.4	38.2	78.8	56.3	77.2	62.4
	TOFU	76.1	58.8	74.1	59.4	25.3	24.6	72.3	58.2	88.2	70.3	88.7	74.9

Table D.6: SFT models across different SFT objectives on Malicious Instruct (MI) and HarmBench (HB) datasets. Safety score is measured via Attack Success Rate (0-100), where lower scores are better.

Bench	Method	Mistral-12B	OLMo-2-13B	Pythia-12B	Llama-3.1-8B	Qwen-3-8B	Phi-4-14B
MI	CE	86.0	96.8	98.1	94.7	54.4	49.3
	FL	88.4	95.3	97.1	94.0	54.3	52.8
	GEM	90.0	96.6	97.2	95.2	61.9	56.7
	TOFU	87.2	94.9	97.0	95.3	56.8	52.4
HB	CE	93.2	96.1	93.8	95.4	82.2	79.4
	FL	92.6	93.8	91.9	91.3	79.2	81.2
	GEM	93.9	94.8	91.4	93.5	83.4	82.2
	TOFU	92.6	91.8	91.3	93.0	79.5	81.7

Table D.7: Performance of models across Alpaca SFT objectives on ROC Stories and Alpaca Eval with extended generation length to 256 tokens. Diversity (D) is measured via Self-BLEU (0–100), where lower scores are better. Quality (Q) is measured via LLM Judge score (0–5), where higher scores are better.

Bench	Method	Mistral-12B		OLMo-2-13B		Llama-3.1-8B	
		D↓	Q↑	D↓	Q↑	D↓	Q↑
ROC	CE	$25.1_{\pm 8.5}$	$5.0_{\pm 0.1}$	$24.5_{\pm 9.8}$	$5.0_{\pm 0.1}$	$22.6_{\pm 8.1}$	$4.9_{\pm 0.1}$
	FL	$15.7_{\pm 4.9}$	$5.0_{\pm 0.1}$	$16.1_{\pm 5.7}$	$4.9_{\pm 0.1}$	$14.4_{\pm 5.3}$	$4.9_{\pm 0.2}$
	GEM	$14.4_{\pm 5.6}$	$5.0_{\pm 0.1}$	$13.4_{\pm 5.0}$	$5.0_{\pm 0.1}$	$11.6_{\pm 4.0}$	$4.9_{\pm 0.1}$
	TOFU	$\mathbf{12.5}_{\pm 4.5}$	$5.0_{\pm 0.1}$	$\mathbf{12.3}_{\pm 4.5}$	$4.9_{\pm 0.1}$	$\mathbf{11.4}_{\pm 3.6}$	$4.9_{\pm 0.1}$
AE	CE	$50.9_{\pm 15.4}$	$4.9_{\pm 0.3}$	$51.6_{\pm 15.9}$	$4.9_{\pm 0.2}$	$50.3_{\pm 15.2}$	$4.9_{\pm 0.3}$
	FL	$33.7_{\pm 12.2}$	$4.8_{\pm 0.3}$	$34.3_{\pm 12.2}$	$4.9_{\pm 0.2}$	$33.2_{\pm 10.6}$	$4.8_{\pm 0.3}$
	GEM	$32.9_{\pm 14.2}$	$4.9_{\pm 0.3}$	$32.0_{\pm 13.5}$	$4.8_{\pm 0.2}$	$31.8_{\pm 13.6}$	$4.8_{\pm 0.3}$
	TOFU	$\mathbf{26.8}_{\pm 11.5}$	$4.8_{\pm 0.3}$	$\mathbf{26.1}_{\pm 10.8}$	$4.8_{\pm 0.3}$	$\mathbf{25.6}_{\pm 10.9}$	$4.8_{\pm 0.3}$

Table D.8: Performance of models across Alpaca SFT objectives on ROC Stories and Alpaca Eval with extended generation length to 512 tokens. Diversity (D) is measured via Self-BLEU (0–100), where lower scores are better. Quality (Q) is measured via LLM Judge score (0–5), where higher scores are better.

Bench	Method	Mistral-12B		OLMo-2-13B		Llama-3.1-8B	
		D↓	Q↑	D↓	Q↑	D↓	Q↑
ROC	CE	$23.1_{\pm 8.1}$	$5.0_{\pm 0.1}$	$23.6_{\pm 9.4}$	$5.0_{\pm 0.1}$	$21.6_{\pm 7.6}$	$4.9_{\pm 0.1}$
	FL	$15.8_{\pm 5.0}$	$5.0_{\pm 0.1}$	$15.7_{\pm 5.7}$	$5.0_{\pm 0.1}$	$14.8_{\pm 4.7}$	$4.9_{\pm 0.1}$
	GEM	$14.0_{\pm 5.4}$	$5.0_{\pm 0.1}$	$13.0_{\pm 4.6}$	$5.0_{\pm 0.1}$	$11.8_{\pm 4.2}$	$4.9_{\pm 0.1}$
	TOFU	$\mathbf{13.0}_{\pm 4.7}$	$5.0_{\pm 0.1}$	$\mathbf{11.6}_{\pm 3.4}$	$4.9_{\pm 0.1}$	$\mathbf{11.6}_{\pm 4.1}$	$4.9_{\pm 0.1}$
AE	CE	$50.9_{\pm 15.3}$	$4.9_{\pm 0.3}$	$51.4_{\pm 16.2}$	$4.9_{\pm 0.2}$	$50.5_{\pm 15.5}$	$4.9_{\pm 0.3}$
	FL	$33.4_{\pm 12.2}$	$4.8_{\pm 0.2}$	$34.3_{\pm 12.5}$	$4.9_{\pm 0.2}$	$31.2_{\pm 10.8}$	$4.9_{\pm 0.2}$
	GEM	$32.9_{\pm 14.2}$	$4.9_{\pm 0.2}$	$33.4_{\pm 13.9}$	$4.9_{\pm 0.2}$	$32.8_{\pm 13.8}$	$4.8_{\pm 0.3}$
	TOFU	$\mathbf{26.6}_{\pm 11.5}$	$4.8_{\pm 0.3}$	$\mathbf{26.3}_{\pm 10.8}$	$4.9_{\pm 0.2}$	$\mathbf{25.8}_{\pm 10.3}$	$4.8_{\pm 0.3}$

Table D.9: Performance of code reasoning models evaluated on Human Eval coding problems. Across different SFT methods, coverage measured via pass@k score for different k budgets. Best performance is highlighted in **bold**.

Model	Method	pass@1	pass@8	pass@16	pass@32
Qwen2.5-Coder-7B	CE	**77.5**	88.2	89.4	89.5
	GEM	76.8	89.3	89.9	90.0
	TOFU	68.3	**93.5**	**94.1**	**94.2**
Qwen2.5-Coder-3B	CE	**68.2**	83.0	83.3	83.6
	GEM	33.8	57.5	69.5	83.1
	TOFU	62.5	**87.7**	**88.8**	**89.1**
DeepSeek-Coder-7B	CE	67.2	79.8	79.9	81.0
	GEM	**67.6**	79.8	79.4	80.8
	TOFU	60.0	**83.5**	**84.1**	**85.4**

Prompt:
Tell me a funny joke.

Cross Entropy Response:

- *Why don't scientists trust atoms? Because they make up everything!*
- *Why don't scientists trust atoms? Because they make up everything!*
- *Q: Why did the chicken go to the séance? A: To get to the other side!*
- *Why did the chicken go to the séance? To get to the other side.*
- *Why don't scientists trust atoms? Because they make up everything!*
- *Why did the chicken go to the séance? To get to the other side!*
- *Why did the chicken go to the séance? To get to the other side!*
- Q: What do you call a bear with no teeth? A: A gummy bear.
- *Why did the chicken cross the playground? To get to the other slide!*
- Q: What did the fish say when it hit the wall? A: Dam!

TOFU Response:

- A duck walks into a bar and orders a beer. The bartender says, ''We don't serve ducks.'' The duck says, ''Do you think I ordered a sandwich?!''
- Q. What do you call an alligator in a vest? A. An investigator!
- What did the fisherman do when he noticed he had been bait-and-switched? He looked for his tackle box, then he took a look in his trunk.
- Why don't you put your clothes under the trees? It's not safe from the squirrels!
- Why did the chicken go to the séance? To get to the ouff-side!
- Q. What did one snowman say to the other snowman? A. Do you smell carrots?
- Why did the chicken go to the séance? To get to the other side! Q. What does a cloud wear when it rains? A. An InCloud-a. Q. Why did the cow go in the movie theater? A. To mooo-ve!
- Q. Why do fire fighters fight? A. Because they have nine lives.
- What did the snowman say to the snowwoman? Don't snow around.
- Q. Why don't scientists trust atoms? A. Because they make up everything.
- Q. What did the baby corn say to its dad? A. Where's my Popperma?

Figure D.1: The comparison of Mistral-12B responses fine-tuned with TOFU against CE following the prompt "Tell me a funny joke". TOFU produces 10/10 unique jokes, whereas the CE model generates only 4 unique outputs. We highlight repeated CE responses in italics.

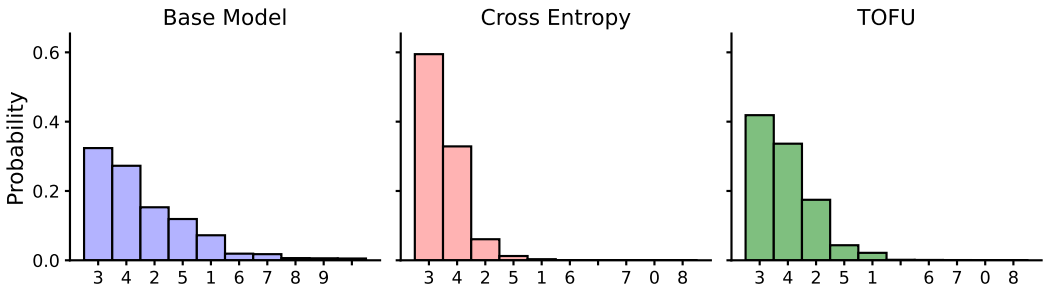

Figure D.2: Predicted probability distribution for the first generated token (ignoring spaces) following the prompt "Generate an integer between 1 and 5". The comparison across base model, SFT with CE, and SFT with TOFU demonstrates that TOFU prevents knowledge forgetting while simultaneously reducing the probability mass of the tail for incorrect tokens.