Entropy-Adaptive Fine-Tuning: Resolving Confident Conflicts to Mitigate Forgetting

Muxi Diao[1,2*], **Lele Yang**[1*], **Wuxuan Gong**[1*], **Yutong Zhang**[1], **Zhonghao Yan**[1], **Yufei Han**[1], **Kongming Liang**[1], **Weiran Xu**[1], **Zhanyu Ma**[1†]
[1]Beijing University of Posts and Telecommunications, [2]Zhongguancun Academy
{dmx, yang_happy, mazhanyu}@bupt.edu.cn

GitHub: https://github.com/PRIS-CV/EAFT

Abstract

Supervised Fine-Tuning (SFT) is the standard paradigm for domain adaptation, yet it frequently incurs the cost of catastrophic forgetting. In sharp contrast, on-policy Reinforcement Learning (RL) effectively preserves general capabilities. We investigate this discrepancy and identify a fundamental distributional gap: while RL aligns with the model's internal belief, SFT forces the model to fit external supervision. This mismatch often manifests as **"Confident Conflicts"**—tokens characterized by low probability but low entropy. In these instances, the model is highly confident in its own prediction but is forced to learn a divergent ground truth, triggering destructive gradient updates. To address this, we propose **Entropy-Adaptive Fine-Tuning (EAFT)**. Unlike methods relying solely on prediction probability, EAFT utilizes token-level entropy as a gating mechanism to distinguish between epistemic uncertainty and knowledge conflict. This allows the model to learn from uncertain samples while suppressing gradients on conflicting data. Extensive experiments on Qwen and GLM series (ranging from 4B to 32B parameters) across mathematical, medical, and agentic domains confirm our hypothesis. EAFT consistently matches the downstream performance of standard SFT while significantly mitigating the degradation of general capabilities.

1 Introduction

Supervised Fine-Tuning (SFT) is the standard method for adapting Large Language Models (LLMs) to specific domains (e.g., mathematics or agentic tool-use) (Yang et al., 2024; Shao et al., 2024; Team et al., 2025). However, this paradigm often comes with a significant cost known as catastrophic forgetting (Kirkpatrick et al., 2017; Ouyang et al., 2022). Previous studies have extensively documented that while fitting specific target distributions, models frequently suffer from degradation in general capabilities (Ouyang et al., 2022; Luo et al., 2023). In contrast, on-policy Reinforcement Learning (RL) has shown a remarkable ability to improve domain-specific performance while effectively preserving the robustness of the base model (Chen et al., 2025; Shenfeld et al., 2025). This sharp contrast raises a fundamental question:

Why does SFT frequently degrade general abilities, while on-policy RL preserves them?

To investigate the mechanisms behind this phenomenon, we systematically analyze the token-level probability and entropy of the training data. As visualized in Fig. 1, this analysis reveals a distinct distributional gap arising from different data sources. In on-policy RL, training sequences are generated via self-rollout; consequently, the tokens inherently align with the model's current probability landscape, falling into either high-probability confidence zones or high-entropy exploration regions. Conversely, SFT relies on external supervision (e.g., humans or strong teacher models), introducing a mismatch manifested as low-probability, low-entropy tokens. Crucially, this mismatch manifests as tokens characterized by **low probability yet low entropy**. This specific region corresponds to scenarios where the model is highly confident in its own prediction (low entropy) but is forced to fit a divergent ground-truth label (low probability). We term these instances **"Confident Conflicts"**. See App. A for representative word clouds.

To verify whether these conflicts are indeed the drivers of forgetting, we conducted a **pilot experiment**. By simply masking out these "Confident Conflict" tokens during training (Fig. 2). We observed that catastrophic forgetting was significantly mitigated compared to standard SFT. This confirms that enforcing updates on these conflicting samples is the primary driver of capability degradation.

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14751–14771
October 24-29, 2026 ©2026 Association for Computational Linguistics

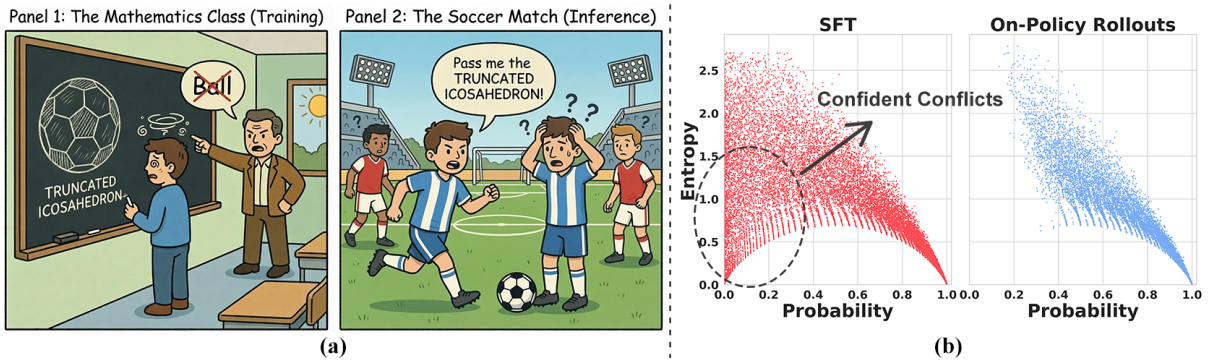


Figure 1: **(a)** Conceptual illustration. When SFT forces the model to override its strong priors (e.g., labeling a "ball" as a "truncated icosahedron"), it creates a ***Confident Conflict***. Fitting these conflicts distorts the model's existing representations, leading to catastrophic forgetting. **(b)** Token-level entropy–probability landscape. Compared to on-policy rollouts (right), the SFT data (left) exhibits a prominent cluster of ***Low Entropy, Low Probability*** tokens.

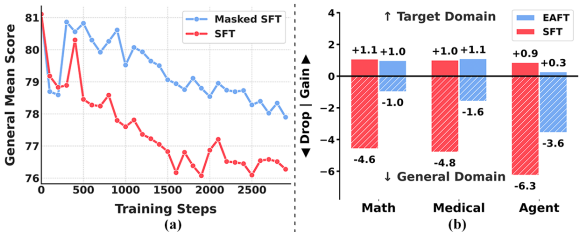


Figure 2: **(a)** Masking "Confident Conflict" tokens (the bottom 15% in both entropy and probability) effectively mitigates the general capability degradation observed in standard SFT. **(b)** Across Math, Medical, and Agent domains, EAFT matches SFT in target task improvements (upper bars) while significantly minimizing performance drops on general benchmarks (lower bars).

Building on this insight, we propose **Entropy-Adaptive Fine-Tuning (EAFT)**. Instead of using discrete thresholds, EAFT employs a soft gating mechanism that dynamically modulates the training loss based on token-level entropy.

Crucially, this approach differentiates EAFT from standard Cross-Entropy or probability-based re-weighting strategies (Wu et al., 2025; Lin et al., 2025; Sanyal et al., 2025; Shenfeld et al., 2025). These methods rely solely on prediction probability, and thus risk amplifying destructive gradients on "Confident Conflicts." In contrast, EAFT leverages entropy to distinguish rigidity from uncertainty. By down-weighting low-entropy tokens to suppress conflicting gradients, while concentrating supervision on high-entropy ones to facilitate adaptation, EAFT effectively balances domain proficiency with the preservation of general capabilities.

To verify the effectiveness and universality of our approach, we validate EAFT through extensive experiments on mathematical, extending to medical and agent domains. Our comprehensive evaluation covers diverse model families (Qwen, GLM) and scales ranging from 4B to 32B parameters. The results are presented in Fig. 2 and Tab. 1.

Quantitative results (Sec. 4.2) demonstrate that EAFT consistently outperforms standard SFT and existing mitigation strategies. It achieves a Pareto improvement: matching or exceeding baselines on target tasks while significantly mitigating catastrophic forgetting on general benchmarks.

Beyond performance, we provide an in-depth analysis of the method's intrinsic properties. We empirically verify that the entropy-adaptive mechanism successfully targets "Confident Conflicts" (Sec. 4.3), and further demonstrate that EAFT is both robust to hyperparameter variations (Sec. 5.1) and computationally efficient (Sec. 5.2).

In summary, our contributions are as follows:

- We **uncover** the distinct distributional gap between SFT and on-policy RL data. Through visualization and pilot experiments, we pinpoint "Confident Conflicts" (low-entropy, low-probability tokens) as the primary cause of catastrophic forgetting.
- We **propose** Entropy-Adaptive Fine-Tuning (EAFT), a novel objective that utilizes token-level entropy to modulate the training loss. This mechanism automatically down-weights destructive updates from conflicting data.
- We **validate** our approach through extensive experiments on Math, Agent, and Medical domains. The results establish EAFT as an effective and universal solution that successfully mitigates catastrophic forgetting across diverse model families and scales (4B–32B).

2 Related Work

Post-training Paradigms: SFT vs. RL. Post-training methods, primarily Supervised Fine-

Tuning (SFT) and Reinforcement Learning (RL), are widely used to align pre-trained LMs (Qwen et al., 2024; GLM et al., 2024; Yang et al., 2025). SFT optimizes the model to maximize the likelihood of ground-truth demonstrations (Off-policy). By contrast, RL optimizes the model based on its own generated responses guided by reward signals (On-policy). These signals typically originate from parameterized reward models or verifiable signals (Schulman et al., 2017; Ouyang et al., 2022; Shao et al., 2024; Zeng et al., 2025).

Emerging research highlights a fundamental dichotomy in their learning behaviors. While SFT is efficient, it is inherently prone to memorization, often fits specific training samples at the expense of generalization (Chu et al., 2025). RL demonstrates superior robustness: it can benefit from single training examples without severe overfitting (Wang et al., 2025), and it updates a smaller, more effective subspace of parameters compared to SFT (Mukherjee et al., 2025). A common thread connecting these results is that RL parameter updates are more local and targeted (Razin et al., 2023).

Our work investigates the root cause of SFT's instability. We argue that unlike on-policy methods which naturally operate within the model's distribution, standard SFT indiscriminately forces the fitting of "confident conflicts"—low-entropy samples that contradict the model's pre-trained knowledge.

Catastrophic forgetting. Catastrophic forgetting remains a foundational challenge in neural networks (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017). Initial efforts to mitigate forgetting focused on preventing parameters from drastically changing (Kirkpatrick et al., 2017; Li and Hoiem, 2017; Lopez-Paz and Ranzato, 2017).

In LLM post-training, this manifests as the "Alignment Tax" (Askell et al., 2021; Ouyang et al., 2022): fine-tuning for domain-specific capabilities (e.g., mathematical problem solving, tool utilization, or biomedical adaptation) often significantly degrades the model's general capabilities (Ouyang et al., 2022; Luo et al., 2023; Shi et al., 2025).

To overcome these limitations, recent works have explored dynamic training strategies that adjust optimization based on token-level metrics. TALR (Lin et al., 2025) dynamically scales learning rates based on token confidence to accelerate convergence. DFT (Wu et al., 2025) re-weights the SFT loss according to prediction probability. Others like RL's Razor (Shenfeld et al., 2025) employ KL divergence as a regularization term to constrain the model's drift from its base distribution.

However, existing dynamic methods predominantly rely on probability or KL divergence as proxies for difficulty or drift. We argue that probability alone is an insufficient statistic: a low-probability token can represent either epistemic uncertainty (valid knowledge to be learned) or a "Confident Conflict" (a destructive sample that contradicts the model's strong priors). By forcing the model to fit these conflicts based on probability, prior methods risk accelerating forgetting. Our work advances this by introducing Entropy as a gating signal.

3 Empirical Analysis & Methodology

In this section, we systematically investigate the causes of catastrophic forgetting in SFT and propose a targeted solution. We begin by defining the problem setup and key metrics in Sec. 3.1. We then present an empirical analysis identifying 'Confident Conflicts' as the primary source of destructive gradients in Sec. 3.2. Finally, based on these insights, we introduce our method, **Entropy-Adaptive Fine-Tuning (EAFT)**, in Sec. 3.3.

3.1 Preliminaries

SFT is the standard process of adapting a base model θ, denoted by its probability distribution P_θ, to a target dataset $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})_i\}_{i=1}^{N}$. For each sample, the response is a sequence of tokens $\boldsymbol{y} = (y_1, \ldots, y_T)$, where T denotes the sequence length. The adaptation is typically achieved by minimizing the Cross-Entropy (CE) loss, which maximizes the likelihood of the target sequences:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{t=1}^{T} \log P_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \qquad (1)$$

A key limitation of this objective is its uniform treatment of all tokens. It aggressively updates model parameters to fit every token y_t regardless of the model's prior knowledge or uncertainty.

To investigate the dynamics of how this uniform objective interacts with the model's internal state, we introduce two token-level metrics that serve as the foundation for our analysis and method:

1. **Probability.** $p_t = P_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})$, represents model's confidence in the ground-truth token.

2. **Predictive Entropy.** Let $P_t(v) \triangleq P_\theta(v|\boldsymbol{x}, \boldsymbol{y}_{<t})$ denote the distribution at step t. The entropy is defined as: $H_t = -\sum_{v\in\mathcal{V}} P_t(v) \log P_t(v)$ This

measures the model's predictive uncertainty over the vocabulary $\mathcal{V}$.

3.2 Analysis: The Origins of Forgetting

To understand why SFT leads to forgetting while on-policy RL does not, we compare the token-level statistics of standard SFT data against model-generated rollouts (the data source for on-policy RL). We visualize the distribution of probability p_t and entropy H_t for both datasets in Fig. 1.

Distributional Gap: Confident Conflicts. The visualization reveals a critical distributional shift. On-policy data falls into either high-probability (model is correct) or high-entropy (model is exploring). In sharp contrast, SFT data contains a significant cluster of tokens with both **Low Entropy ($H_t \downarrow$) and Low Probability ($p_t \downarrow$)**. We term these samples **'Confident Conflicts'**. They represent cases where the model holds a strong, stubborn prior belief (low entropy) that directly contradicts the ground-truth label (low probability).

Pilot Study: Masking Confident Conflicts. We hypothesize that these 'Confident Conflicts' are the primary drivers of forgetting. To verify this, we conducted a pilot experiment where we masked the loss for tokens falling within the **bottom 15% of both entropy and probability rankings**. As shown in Fig. 2, this simple intervention significantly mitigates the general capability degradation observed in standard SFT.

Notably, masking these specific tokens nearly eliminated catastrophic forgetting on our benchmarks. This finding confirms that the degradation of general capabilities stems primarily from forcing the model to accommodate these conflicting samples, rather than from the SFT process itself.

Theoretical Insight. We analyze the optimization dynamics to understand this damage. Consider the CE loss (Eq. 1). When the model is highly confident in a prediction that contradicts the target (low entropy, low probability), the CE loss induces a very large gradient. Because the model strongly favors another token, fitting the target requires substantial parameter updates, which can overwrite general representations in the base model. By contrast, when the model is uncertain (high entropy), the gradients are smaller and updates are gentler, helping preserve the model's original capabilities.

3.3 Entropy-Adaptive Fine-Tuning (EAFT)

While the pilot study validates our hypothesis, the hard masking strategy has two limitations: it discards training data, leads to ineffective learning on the target domain, and it relies on sensitive hyperparameters (τ, δ). To address this, we propose **Entropy-Adaptive Fine-Tuning (EAFT)**, a soft gating mechanism that dynamically adjusts the learning signal based on the model's uncertainty.

The EAFT Objective. We formulate the EAFT loss by scaling the standard supervision with the normalized entropy. This mechanism **prioritizes learning** from samples where the model is exploring, while effectively suppressing the gradients when the model is confident but conflicting. The objective is decomposed as:

$$\mathcal{L}_{\text{EAFT}}(\theta) = -\sum_{t=1}^{T} \underbrace{\tilde{H}_t}_{\substack{\text{Adaptive} \\ \text{Gating Signal}}} \cdot \underbrace{\log P_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})}_{\text{Standard Supervision}} \tag{2}$$

Here, the gating term $\tilde{H}_t$ is derived from the entropy of the **Top-K tokens**. This approximation greatly reduces computation compared with using the full vocabulary (analysis in Sec. 5.2). Normalized to the range $[0, 1]$ with $K = 20$, we calculate:

$$\tilde{H}_t = \frac{H_t^{\text{top-}K}}{\ln(K)} \approx \frac{H_t^{\text{top-20}}}{3.0} \tag{3}$$

where $H_t^{\text{top-}K}$ denotes the entropy calculated over the top-K probability distribution, and $\ln(K)$ serves as the normalization factor (the maximum entropy for K outcomes). This normalization creates a self-regulating mechanism:

- **Conflict Suppression ($\tilde{H}_t \to 0$):** When the model is stubborn (low entropy), the weight drops, effectively **masking** the destructive gradient from the conflicting label.
- **Knowledge Acquisition ($\tilde{H}_t \to 1$):** When the model is uncertain (high entropy) or exploring, the weight remains high, recovering the standard SFT objective to learn new patterns.

4 Experiments

In this section, we empirically validate the effectiveness of EAFT. Our experiments are designed to answer the following three key research questions:

- RQ1 (Performance): Can EAFT mitigate catastrophic forgetting without compromising performance on the target task?

Method	Math Domain			Math Avg.	General Domain			General Avg.
	AIME24	AIME25	GSM8K		MMLU	IFEval	CLUEWSC	
Qwen3-4B-Instruct	63.3	47.4	94.3	68.3	77.1	81.0	85.2	81.1
+ SFT	63.3	50.0	**94.8**	**69.4**	76.5	79.5	74.5	76.5 (-4.6)
+ SFT_{KL}	63.3	50.0	93.6	69.0	74.5	74.9	**89.4**	79.6 (-1.5)
+ FLOW	**66.7**	46.7	94.3	69.2	76.2	78.3	82.8	79.1 (-2.0)
+ DFT	56.7	40.0	93.9	63.5	75.9	77.0	81.4	78.1 (-3.0)
+ TALR	50.0	50.0	93.3	64.4	76.2	78.1	74.5	76.2 (-4.9)
+ EAFT	60.0	**53.3**	94.5	69.3	**76.6**	**80.1**	83.7	**80.1** (-1.0)
Qwen2.5-32B-Instruct	22.2	13.3	96.0	43.8	84.1	78.3	91.9	84.8
+ SFT	**53.3**	**50.0**	96.3	**66.5**	76.9	74.2	93.8	81.6 (-3.2)
+ SFT_{KL}	33.3	33.3	94.1	53.6	**81.4**	68.1	93.2	80.9 (-3.9)
+ FLOW	50.0	**50.0**	96.3	65.4	78.6	75.1	93.6	82.4 (-2.4)
+ DFT	33.3	36.7	95.9	55.3	77.8	70.0	**94.4**	80.7 (-4.1)
+ TALR	40.0	43.3	95.3	59.5	73.1	72.5	94.1	79.9 (-4.9)
+ EAFT	**53.3**	46.7	**96.5**	65.5	79.0	**78.4**	93.9	**83.7** (-1.1)
GLM4-9B-0414	6.7	6.7	90.1	34.5	70.2	74.4	85.1	76.6
+ SFT	**20.0**	10.0	90.3	40.1	57.3	69.8	84.8	70.6 (-6.0)
+ SFT_{KL}	13.3	6.7	90.1	36.7	60.0	66.4	85.3	70.5 (-6.1)
+ FLOW	16.7	**13.3**	91.1	**40.4**	57.5	71.5	85.2	71.4 (-5.2)
+ DFT	13.3	6.7	89.0	36.4	48.9	69.7	**86.0**	68.2 (-8.4)
+ TALR	15.6	**13.3**	91.2	40.0	57.4	71.3	84.5	71.5 (-5.1)
+ EAFT	13.3	**13.3**	**91.5**	39.4	**60.8**	**72.0**	85.3	**72.7** (-3.9)

Table 1: Main results on the target domain (Math) and general domain benchmarks. We evaluate performance on AIME24, AIME25 and GSM8K as the training target, alongside MMLU, IFEval, and CLUEWSC for general capabilities. The top two outcomes are **bolded** and underlined. All results are averaged over three independent runs. The "Avg." represents the average performance of the datasets in the corresponding domain.

- RQ2 (Mechanism): Does the entropy-adaptive gating mechanism work as intended in filtering "Confident Conflict" samples?
- RQ3 (Generalization): Is the efficacy of EAFT inherently domain-agnostic?

4.1 Experimental Settings

Datasets. We utilize prompts from Numina-Math (Li et al., 2024), BigMathVerified (Albalak et al., 2025), and Nemotron-CrossThink (Akter et al., 2025), synthesizing responses via Qwen3-235B-A22B-Instruct (Yang et al., 2025). We randomly select 19k correctly answered data pairs as our math training data. Details in App. E

Benchmarks. Our evaluation covers the target domain (Math: AIME 24/25 (AI-MO, 2025), GSM8K (Cobbe et al., 2021)) and general domain (MMLU (Hendrycks et al., 2020), IFEval (Zhou et al., 2023), CLUEWSC (Xu et al., 2020)). We report the average score ("Avg") across benchmarks, with full details available in App. F.

Models. To verify the scalability and generalizability of our method, we conduct experiments across a diverse model zoo spanning multiple families and parameter scales. Specifically, our evaluation includes models ranging from 4B to 32B parameters: Qwen3-4B-Instruct (Yang et al., 2025), GLM4-9B-0414 (GLM et al., 2024), and Qwen2.5-32B-Instruct (Qwen et al., 2024). Details in App. G

Baselines. We compare our method against standard SFT, a regularized variant with a KL divergence constraint (SFT_{KL}, see App. D for β sweep). Additionally, we include several advanced alignment baselines, including FLOW (Sanyal et al., 2025), DFT (Wu et al., 2025), and TALR (Lin et al., 2025). Details and implementation settings are provided in App. H and App. I

4.2 Main Results (Answering RQ1)

Tab. 1 presents the performance comparison across diverse model families and scales. The results provide a compelling answer to RQ1: **EAFT maintains competitive performance on target domain (Math) while significantly mitigating catastrophic forgetting on general capabilities.**

General Capabilities. The main advantage of EAFT lies in stability. Standard SFT causes a sharp drop in general benchmarks (e.g., a 10.7 point drop on CLUEWSC for Qwen3-4B). In contrast, EAFT effectively preserves the model's original knowledge, achieving the highest average score across general tasks among all baselines. This demonstrates that EAFT offers the most robust capability retention compared to other methods.

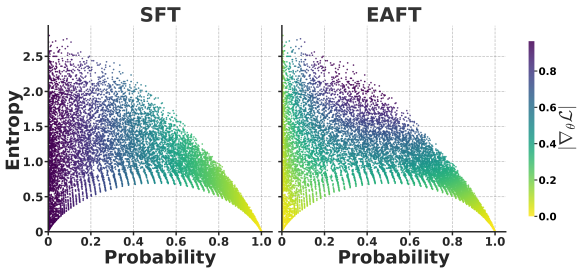


Figure 3: Gradient Magnitude Landscape. **Left:** SFT exerts strong optimization pressure (dark purple) on *Confident Conflicts* in the bottom-left. **Right:** EAFT effectively suppresses these gradients (light yellow), protecting the model's existing representations.

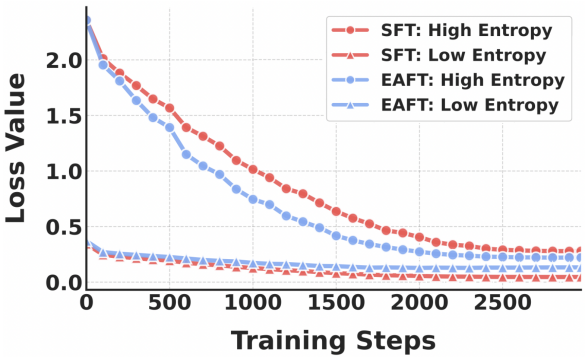


Figure 4: Training dynamics of token subgroups. EAFT matches SFT on high-entropy tokens while keeping losses stable on low-entropy conflicts, preventing over-optimization of conflicting priors. High and low entropy correspond to values ≥ 2.0 and ≤ 0.5, respectively.

Target Domain Performance. EAFT achieves highly competitive results on the target domain compared to other baselines. Specifically, the gap between EAFT and the best-performing method on math score is consistently less than 1 point. Notably, EAFT achieves the best performance on several sub-benchmarks. This demonstrates that EAFT effectively adapts to target domain, maintaining the same level of learning capability as standard SFT.

4.3 Mechanism Analysis (Answering RQ2)

To understand the inner workings of EAFT, we analyze the training dynamics from both spatial and temporal perspectives. The results provide a clear affirmative answer to RQ2: **EAFT effectively filters out "Confident Conflict" samples, preventing destructive gradient updates.**

Gradient Landscape. Fig. 3 visualizes the optimization strength for each token, where color intensity represents the gradient magnitude. The gradient distribution exhibits a distinct skew rather than uniformity. Due to the nature of Cross-Entropy loss, low-probability tokens are subjected to the strongest optimization pressure (indicated by the dark density in the left region). Crucially, in the "Confident Conflict" zone (bottom-left), the model is forcefully updated to fit labels that contradict its confident priors. These aggressive updates destabilize the model's established representation space. In contrast (right), EAFT dramatically alters this landscape. The "Confident Conflict" region becomes pale, indicating near-zero gradients. Here, the low entropy term directly suppresses the high gradients generated by the Cross-Entropy loss. This protects the model's established knowledge from being over-optimized by conflicting data.

Training Dynamics. Fig. 4 tracks the loss of cross-entropy of different token types throughout the training process, comparing EAFT (Blue) with Standard SFT (Red). We categorize tokens into high-entropy and low-entropy groups. **High-Entropy Tokens** (marked with dots •): EAFT exhibits a rapid loss reduction comparable to SFT. This confirms that our entropy-based gating effectively optimizes high-entropy tokens, ensuring the model adapts to the target domain without hindrance. **Low-Entropy Tokens** (marked with triangles ▲): SFT (Red) aggressively drives this loss toward zero, indicating that the model is being forced to memorize data that conflicts with its priors. In contrast, EAFT (Blue) maintains a stable loss throughout training. This demonstrates that the mechanism successfully prevents over-optimization on "Confident Conflict" tokens.

4.4 Universality (Answering RQ3)

To verify whether the efficacy of EAFT is domain-agnostic, we extend our evaluation to two distinct specialized domains: Biomedical (knowledge-intensive) and Agent Tool-Use (syntax-intensive). The results clearly answer RQ3: **EAFT is a domain-agnostic solution that consistently mitigates catastrophic forgetting across tasks.**

Medical Domain. We first examine the medical domain, a field demanding domain-specific knowledge application. For this experiment, we fine-tune the **Qwen3-4B-Thinking** (Yang et al., 2025) model using the **Huatuo-O1** (Chen et al., 2024a) dataset and conduct evaluations on the **MedMCQA** (Pal et al., 2022),**PubMedQA** (Jin et al., 2019)and **MedQA** (Jin et al., 2021) benchmarks (see App. F for more details). In Tab. 2, standard SFT triggers severe catastrophic forgetting while adapting

Method	Medical Domain			Medical Avg.	General Domain			General Avg.
	MedMCQA	MedQA	PubMedQA		MMLU	IFEval	CLUEWSC	
Qwen3-4B-Thinking	63.5	78.2	76.0	72.6	79.3	**85.0**	**94.1**	**86.1**
+ SFT	63.3	79.5	**78.0**	73.6	78.3	75.3	90.4	81.3 (-4.8)
+ EAFT	**63.9**	**80.0**	77.2	**73.7**	**80.1**	81.7	91.8	84.5 (-1.6)

Table 2: Results on target (Medical) and General domain benchmarks. We evaluate performance on MedMCQA, MedQA, and PubMedQA for the medical domain, alongside MMLU, IFEval, and CLUEWSC for general capabilities. The top two outcomes are **bolded** and underlined. All results are averaged over three independent runs.

Method	Agent Domain	General Domain			General Avg.
	BFCL v3	MMLU	IFEval	CLUEWSC	
Qwen3-4B-Instruct	60.5	**77.1**	**81.0**	**85.2**	**81.1**
+ SFT	**61.4**	74.5	77.8	72.2	74.8 (-6.3)
+ EAFT	60.8	76.1	78.6	77.7	77.5 (-3.6)

Table 3: Results on target (Agent Toolcall) and General domain benchmarks. We evaluate performance on BFCL for the agent toolcall domain, alongside MMLU, IFEval, and CLUEWSC for general capabilities. The top two outcomes are **bolded** and underlined. All results are averaged over three independent runs.

to the target domain, causing average performance on general benchmarks to drop significantly from 86.1 to 81.3. EAFT effectively mitigates this issue. It not only preserves general capabilities (maintaining an average of 84.5) but also marginally outperforms standard SFT on the target medical tasks (73.7 vs. 73.6). This result suggests that EAFT can inject specific knowledge without destructively overwriting the model's core representation.

Agent Tool-Use. We further evaluate EAFT on agentic tool-use tasks, which require strict adherence to syntactic constraints and formats. Specifically, we utilize the subset of **Nemotron-Agentic-Tool-Use-v1** (NVIDIA, 2025) to train the **Qwen3-4B-Instruct** (Yang et al., 2025) and assess its performance on the Berkeley Function Calling Leaderboard (**BFCL**) (Patil et al.) (see App. F). The results in Tab. 3 reveal a similar pattern regarding robustness. Although SFT fits the target distribution aggressively to achieve a slightly higher score (61.4), it does so at the cost of catastrophic forgetting, resulting in a sharp decline in general capabilities (81.1 $\rightarrow$ 74.8). In contrast, EAFT achieves a far superior balance: it remains highly competitive on the target task (60.8, within 1% of SFT) while maintaining general performance (77.5). These experiments confirm that EAFT is a domain-agnostic solution that consistently alleviates catastrophic forgetting across varying data distributions.

5 Analysis and Discussion

Having demonstrated the superior performance of EAFT, this section investigates the **intrinsic robustness** and **computational efficiency** of the proposed mechanism. Specifically, we conduct ablation studies to answer two fundamental questions:

- **Robustness:** Does performance rely on linear entropy scaling, or is the "entropy-aware" mechanism the primary driver? (Sec. 5.1)
- **Efficiency:** Can we approximate the token-level entropy efficiently without incurring significant computational overhead? (Sec. 5.2)

5.1 Robustness to Gating Function Variations

In Sec. 3, we adopted a linear gating function $\mathcal{L} = \tilde{H}_t \cdot \mathcal{L}_{\text{CE}}$. To verify that our gains stem from the intrinsic **mechanism of entropy awareness** rather than a specific hyperparameter choice, we generalize the gating signal to $f(\tilde{H}_t)$ and evaluate three distinct categories of variants:

- **Polynomial Scaling:** We test stricter suppression of confident samples using power functions $f(\tilde{H}_t) = (\tilde{H}_t)^p$ with $p \in \{2, 3\}$. We denote these variants as **EAFT**^{2} and **EAFT**^{3}.
- **Non-linear Gating (Sigmoid):** We employ a Sigmoid activation $f(\tilde{H}_t) = \sigma\big(\alpha(\tilde{H}_t - \beta)\big)$ where α controls the steepness and β determines the centering threshold. In our experiments, we set $\beta = 0.17$ (aligning with the bottom 15% entropy percentile used in the Masked SFT baseline) and $\alpha = 30$. This variant is denoted as **EAFT**$_{sig}$. See App. J for further comparisons.
- **Piecewise Hard Thresholding (Hard Mask):** We implement a binary filter that strictly discards the bottom 15% lowest entropy tokens:

$$f(\tilde{H}_t) = \mathbb{I}(\tilde{H}_t > \tau_{0.15}) \tag{4}$$

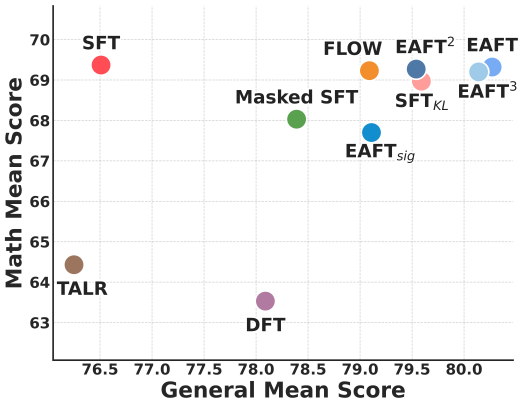


Figure 5: **Pareto trade-off analysis.** Unlike Masked SFT (drop in target score) or Standard SFT (severe forgetting), EAFT variants consistently occupy the optimal **top-right frontier**. This confirms that soft entropy-gating effectively preserves general capabilities without compromising target domain adaptation.

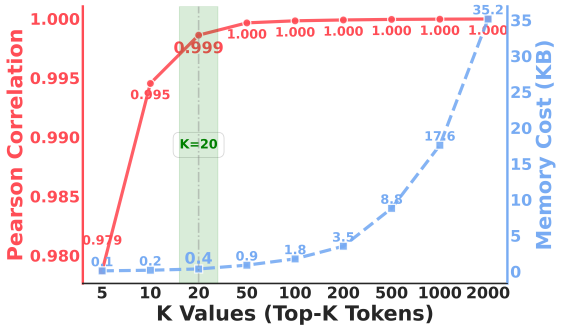


Figure 6: **Trade-off between Approximation Accuracy and Memory Cost.** The Red solid line (Left Axis) shows the Pearson correlation between Top-K and exact entropy, which saturates rapidly at 0.999. The Blue dashed line (Right Axis) tracks the additional memory overhead. The green shaded region highlights $K = 20$ as the optimal operating point, achieving near-perfect fidelity with negligible computational cost.

where $\mathbb{I}(\cdot)$ is the indicator function and $\tau_{0.15}$ denotes the threshold value corresponding to the 15^{th} percentile of the entropy distribution. This sets the loss of "confident" tokens to zero, retaining only uncertain samples for training. We refer to this baseline as **Masked SFT**.

Results and Discussion. Fig. 5 visualizes the trade-off between target domain performance (math average score) and general average score across all variants. We observe two pivotal insights:

Universality of Entropy Awareness. All entropy-aware variants (EAFT, EAFT^2, EAFT^3, EAFT$_{sig}$) consistently outperform SFT in general capabilities. This confirms that the reduction in forgetting stems from the core **mechanism of entropy monitoring** rather than the specific mathematical form. Simply down-weighting high-conflict tokens is sufficient to preserve general abilities.

The Necessity of Soft Gating. We observe a critical pitfall in the **Mask** baseline: while strictly discarding confident tokens prevents forgetting, it significantly harms target task performance (Math Score drops to 65.60 vs. EAFT's 69.27). This implies that "confident conflicts" carry essential adaptation signals that a hard cutoff destroys. In contrast, EAFT's soft gating reduces their impact without removing them, successfully occupying the **Pareto frontier** of learning versus retaining.

5.2 Efficiency of Top-K Approximation

Computing exact entropy over the entire vocabulary (often $|\mathcal{V}| > 100k$) creates unnecessary computational bottlenecks. In Sec. 3, we proposed a Top-K approximation strategy ($\tilde{H} \approx H^{\text{top-}K}$), premised on the property that the probability mass of LLMs is highly sparse and concentrated in the leading tokens (Fan et al., 2018; Radford et al., 2019; Holtzman et al., 2019).

Fig. 6 quantitatively validates this design by plotting approximation fidelity (Left Axis, Red) against memory cost (Right Axis, Blue). We observe a decisive trade-off: as K increases, the Pearson correlation with the exact entropy rises sharply and rapidly plateaus. Specifically, at $K = 20$ (green zone), the correlation reaches **0.999**, confirming that the long-tail distribution contributes negligibly to entropy. Conversely, the memory cost for the required operations (sort and log-sum-exp) remains virtually zero ($< \mathbf{0.4}$ KB) for $K \leq 20$. Based on these findings, **we adopt the Top-20 approximation to estimate entropy.** This setting provides an accurate estimate of the full-vocabulary entropy, while introducing no noticeable computational overhead compared to standard SFT.

6 Conclusion

In this work, we identify "Confident Conflicts" as the primary driver of catastrophic forgetting in SFT. We introduce Entropy-Adaptive Fine-Tuning (EAFT), a method that dynamically modulates training loss based on token-level entropy. By suppressing gradients from conflicting data, EAFT effectively prevents destructive updates while maintaining learning efficiency. Extensive experiments across diverse domains and model scales validate our approach. Ultimately, EAFT provides a simple

yet robust solution for balancing domain adaptation with the preservation of general capabilities.

7 Limitations

Scope of Applicability. It is important to clarify that EAFT is designed primarily for *domain adaptation* and *continual learning*, where the goal is to extend the model's capabilities without erasing existing knowledge. It is not a universal replacement for standard SFT, particularly in scenarios requiring **knowledge editing** or **counterfactual training** (e.g., teaching the model that "the sky is green" or correcting outdated facts). In such cases, the model's resistance to "Confident Conflicts" is undesirable, as the objective is precisely to override the prior belief. EAFT would interpret these necessary updates as conflicts and suppress them, hindering the intended learning.

Target Performance Trade-off. While EAFT successfully mitigates catastrophic forgetting, it does not aim to surpass standard SFT on the target domain metrics. As observed in our experiments, EAFT achieves a Pareto improvement—maintaining high general capabilities while closely approaching, but not necessarily exceeding, the peak specialization performance of SFT. For applications where maximizing target domain performance is the sole priority, regardless of the degradation in other areas, standard SFT may still be the preferred choice.

Dependence on Base Model Quality. EAFT operates on the assumption that the model's high-confidence priors represent valuable general knowledge worth preserving. However, if the base model exhibits **miscalibrated confidence** (e.g., being confidently wrong or hallucinating), EAFT inadvertently protect these erroneous behaviors. Future work could explore incorporating uncertainty calibration techniques to distinguish between true knowledge and confident hallucinations.

Acknowledgments

This work was supported by Zhongguancun Academy (C20250602 and C20250204), the National Natural Science Foundation of China (Grants U23B2052, 62225601, and 62476029), and the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

References

AI-MO. 2025. Ai-mo/aimo-validation-aime · datasets at hugging face. https://huggingface.co/datasets/AI-MO/aimo-validation-aime. Online; accessed 2025-12-28.

Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, et al. 2025. Nemotron-crossthink: Scaling self-learning beyond math reasoning. *arXiv preprint arXiv:2504.13941*.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, et al. 2025. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Google DeepMind. 2026. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/. Accessed: 2026-05-24.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Jiacheng Lin, Zhongruo Wang, Kun Qian, Tian Wang, Arvind Srinivasan, Hansi Zeng, Ruochen Jiao, Xie Zhou, Jiri Gesi, Dakuo Wang, et al. 2025. Sft doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in llms. *arXiv preprint arXiv:2509.20758*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv e-prints*, pages arXiv–2308.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. 2025. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*.

NVIDIA. 2025. Nemotron-agentic-v1 dataset. Hugging Face. Accessed: 2026-01-04.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 23 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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

Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua Susskind, and Etai Littwin. 2023. Vanishing gradients in reinforcement finetuning of language models. *arXiv preprint arXiv:2310.20703*.

Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. 2025. Upweighting easy samples in fine-tuning mitigates forgetting. *arXiv preprint arXiv:2502.02797*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. Rl's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42.

Haoxiang Sun, Lizhen Xu, Bing Zhao, Wotao Yin, Wei Wang, Boyu Yang, Rui Wang, and Hu Wei. 2026. Deepvision-103k: A visually diverse, broad-coverage, and verifiable mathematical dataset for multimodal reasoning. *arXiv preprint arXiv:2602.16742*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

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

Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. 2025. Implicit reward as the bridge: A unified view of sft and dpo connections. *arXiv preprint arXiv:2507.00018*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

xAI. 2024. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA. Accessed: 2025-04-26.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Yxn, Shushan Bai, Man Shu, and Xiangang Xi. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (aime) 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Bangkok, Thailand. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2026. Anchored supervised fine-tuning. *Preprint*, arXiv:2509.23753.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pages 48337–48383.

Appendix

A Qualitative Analysis: What are "Confident Conflicts"? 13

B Extensions to Multimodal Tasks 14

C Evolution of Target Domain Entropy 14

D Hyperparameter Sensitivity and Baseline Optimality 14

E Open-source Datasets 15
- E.1 Mathematical Datasets 15
- E.2 Medical Datasets 16
- E.3 Agent Datasets 16
- E.4 Multimodal Mathematical Datasets 16

F Benchmarks 16
- F.1 General Benchmarks 16
- F.2 Mathematical Benchmarks 16
- F.3 Medical Benchmarks 17
- F.4 Agent Benchmarks 17
- F.5 Multimodal General Benchmarks . 17
- F.6 Multimodal Math Benchmarks . . 18

G Models 18

H Baselines 19

I Implementation Details 19
- I.1 Training Setup 19
- I.2 Hyperparameter Settings 19
- I.3 Evaluation Protocol 19

J Extended Discussion 20
- J.1 Comparison with EAFT$_{sig}$ 20
- J.2 Comparison with RL and RFT . . 20
- J.3 Nature of Forgetting: Semantic Degradation vs. Style Overwriting 20

K LLM Usage 21

A Qualitative Analysis: What are "Confident Conflicts"?

To better understand the motivation behind our Entropy-Adaptive Fine-Tuning loss, we visualize the vocabulary distribution across different entropy and probability regimes in Figure 7. We categorize tokens into three distinct groups based on the model's predictive uncertainty:

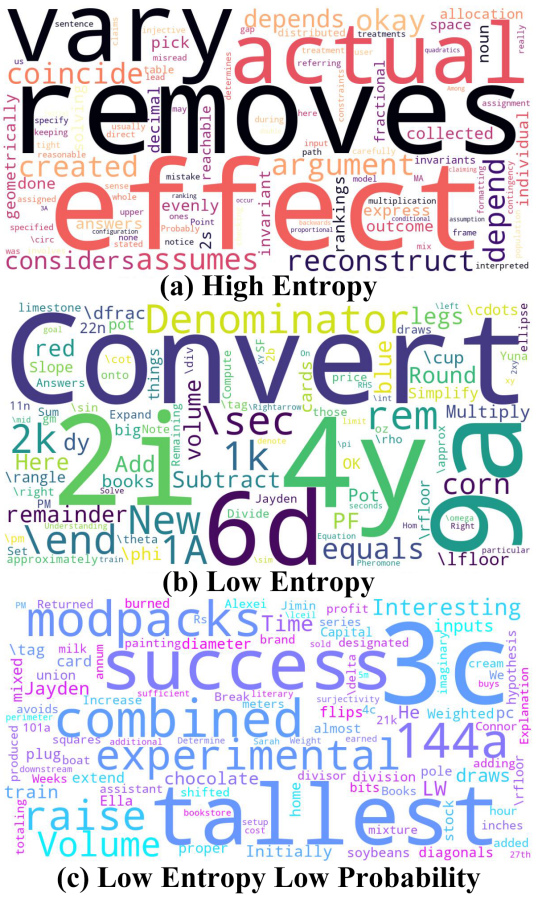


Figure 7: Word clouds visualizing tokens categorized by entropy and probability. **(a)** High entropy tokens often correspond to reasoning steps. **(b)** Low entropy tokens correspond to syntax and fixed patterns. **(c) Confident Conflicts tokens** often involve specific entities or data noise. In these visualizations, **word size** represents frequency, while **color** represents entropy.

- **(a) Branching Points (High Entropy):** As shown in the top panel, high-entropy tokens are predominantly abstract verbs (e.g., *vary, depends, reconstruct*), reasoning connectors, and general nouns. These tokens typically represent semantic **branching points** where the model faces multiple plausible continuation paths. They often correspond to complex reasoning steps or logical transitions, which are critical for the model to learn during SFT.

- **(b) Confident (Low Entropy):** The middle panel displays tokens where the model exhibits low entropy and high probability. This category is dominated by mathematical symbols (e.g., *Convert, 4y, 6d*), LaTeX syntax (e.g., `\frac`, `\sin`), and functional words. These tokens represent deterministic syntactic structures or rote

Method	MM Math Domain			Math Avg.	General Domain			General Avg.
	DynaMath	MathVision	MathVista		MMMU	MMStar	RealWorldQA	
Qwen3.5-2B(Thinking)	66.9	43.1	73.9	61.3	64.2	71.7	74.5	70.1
+ SFT	69.7	**45.8**	**76.1**	**63.9**	**60.0**	67.8	70.7	66.2 (-3.9)
+ EAFT	**70.1**	45.3	75.0	63.5	59.1	**68.4**	**74.4**	**67.3** (-2.8)
Qwen3-VL-4B-Thinking	72.9	55.6	77.0	68.5	70.8	73.2	73.2	72.4
+ SFT	**73.6**	**60.2**	**77.8**	**70.5**	65.3	71.9	72.4	69.9 (-2.5)
+ EAFT	73.2	59.1	77.7	70.0	**68.6**	**73.2**	**73.1**	**71.6** (-0.8)

Table 4: Results on Multimodal Math and General domain benchmarks. We evaluate performance on DynaMath, MathVision, and MathVista for the math domain, alongside MMMU, MMStar, and RealWorldQA for general multimodal capabilities. The top two outcomes are **bolded** and underlined.

memorization, which are generally easier for the model to master.

- **(c) Confident Conflicts (Low Entropy and Low probability):** The bottom panel illustrates tokens where the model is confident but potentially incorrect. This group consists largely of specific entities (e.g., *Jayden, modpacks*), rare domain terms, or noise. These often indicate a mismatch between the model distribution and the SFT data, or specific long-tail knowledge.

Insight: This qualitative distinction justifies our entropy-based adaptive strategy. The visualization highlights "Confident Conflicts" (Low Entropy, Low Probability) where the model's prior strongly mismatches the SFT data—a primary source of catastrophic forgetting. By differentiating these tokens from critical reasoning steps (High Entropy), our method mitigates the risk of **destroying the model's original distribution** while effectively adapting to new reasoning patterns.

B Extensions to Multimodal Tasks

We extend our experiments to continuous learning within the multimodal domain. Specifically, we evaluate EAFT on multimodal mathematics and general Visual Question Answering (VQA) benchmarks. For the experimental setup, we shuffled 16k data samples from (Sun et al., 2026). We evaluated our approach using the latest Qwen Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Qwen Team, 2026) across various parameter scales.

As shown in Table 4, we observe that both SFT and EAFT effectively improve performance on the target domain. However, standard SFT suffers from a severe performance drop in the general domain, whereas EAFT successfully maintains the original capabilities. Therefore, the core conclusions remain consistent with our text-based findings: EAFT maintains competitive performance on the target multimodal math domain while significantly mitigating catastrophic forgetting on general VQA capabilities. This further validates the effectiveness and universality of EAFT in the multimodal domain.

C Evolution of Target Domain Entropy

To further elucidate the convergence mechanism and understand how the model's familiarity with the target domain evolves, we monitor the target prediction entropy throughout the training process. Specifically, we track the supervised entropy metrics scaled by $1/\ln(20)$ over 3k training steps.

Figure 8 illustrates the trajectories of the maximum, mean, and minimum entropy. As training progresses, the mean entropy exhibits a steady and pronounced downward trend, starting from approximately 0.1 and gradually decaying to stabilize near zero. This continuous reduction in entropy strongly indicates that the model progressively resolves ambiguity and gains confidence in the target domain. Furthermore, the synchronization between this entropy decay and the loss convergence demonstrates that our mechanism effectively guides the model toward a highly certain and accurate solution on the target task.

D Hyperparameter Sensitivity and Baseline Optimality

To ensure a rigorous evaluation, we investigate whether the SFT_{KL} baseline can match our method when its core hyperparameter, the KL coefficient β, is optimally tuned. Following recent work (Zhu et al., 2026), we conduct an extensive hyperparameter sweep for SFT_{KL} by varying $\beta \in \{0.01, 0.1, 0.5, 1\}$, evaluating on both mathematical reasoning and general capabilities.

As shown in Table 5, while tuning β optimizes SFT_{KL}, our EAFT consistently achieves Pareto-

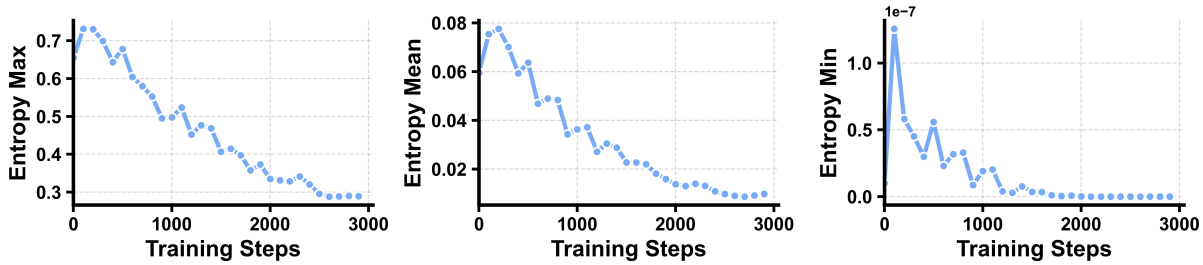


Figure 8: Trajectories of the maximum, mean, and minimum supervised prediction entropy (scaled by $1/\ln(20)$) throughout training. Alongside the steady decay of the mean entropy from $\sim$0.1 to near zero, the overall narrowing of the entropy bounds indicates that the model progressively resolves ambiguity and establishes uniform confidence across the target domain.

Method	Math Avg.	General Avg.
Qwen3-4B	68.3	81.1
SFT	**69.4**	76.5
SFT_{KL} $\beta = 0.01$	69.3	77.0
SFT_{KL} $\beta = 0.1$	69.2	77.8
SFT_{KL} $\beta = 0.5$	69.0	79.6
SFT_{KL} $\beta = 1$	67.8	76.0
EAFT	69.3	**80.1**

Table 5: Hyperparameter sweep of the β coefficient for SFT_{KL}, compared with standard baselines and our EAFT on Qwen3-4B.

optimal performance and surpasses the baseline frontier. This advantage stems from three key aspects. First, unlike the global static penalty of SFT_{KL} that rigidly trades off adaptation for preservation, EAFT employs a token-level dynamic gating mechanism to selectively inhibit destructive updates during high-confidence conflicts. Second, while SFT_{KL} is highly sensitive to β and requires expensive tuning for different tasks, EAFT delivers robust, out-of-the-box effectiveness across model scales from 4B to 32B parameters. Finally, EAFT eliminates the need for a frozen reference model, avoiding the substantial memory and compute overhead mandated by SFT_{KL}.

E Open-source Datasets

This section details the open-source datasets utilized for training and fine-tuning models, categorized into mathematical reasoning, medical knowledge, and agent capabilities.

E.1 Mathematical Datasets

The datasets listed below serve as prompt sources for our data distillation process. We utilized the **Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025)** model to generate responses to these prompts. For the final training set, we randomly sampled 19k instances where the model's outputs were verified as correct.

NuminaMath (Li et al., 2024) is a massive dataset dedicated to enhancing the mathematical reasoning capabilities of large language models. Representing the largest public collection in the field, it consists of approximately 860,000 pairs of competition-level math problems and detailed solutions. The dataset covers a broad spectrum of mathematical disciplines and difficulty levels.

BigMathVerified (Albalak et al., 2025) is a large-scale, high-quality dataset designed specifically for applying reinforcement learning to LMs in the domain of mathematics. It aggregates a vast collection of mathematical problems and solutions from diverse sources, prioritizing data quality and reasoning rigor. By providing a rich environment for training and fine-tuning, Big-Math aims to advance the ability of models to perform reliable multi-step mathematical reasoning and self-correction through reinforcement learning techniques.

Nemotron-CrossThink (Akter et al., 2025) is a large-scale dataset originally designed to extend self-learning mechanisms across various domains. While the full dataset covers a broad spectrum of reasoning tasks, **we exclusively select the mathematics subset** to align with our specific research focus. This subset contains high-quality synthetic problems with rigorous logical verification. By isolating these instances, we leverage the dataset's inherent self-correction patterns to specifically enhance the model's precision in math derivation.

E.2 Medical Datasets

Huatuo-o1 (Chen et al., 2024a) is a specialized supervised fine-tuning (SFT) dataset generated through knowledge distillation from the **DeepSeek-R1 (Guo et al., 2025)** model. Built upon the set of verifiable medical problems from **HuatuoGPT-o1 (Chen et al., 2024a)**, this dataset leverages DeepSeek-R1 to generate comprehensive, native reasoning chains. The primary objective is to transfer the advanced reasoning capabilities of the R1 model to target models, enabling them to initialize and internalize deep clinical thinking patterns and logical deductions essential for medical tasks.

E.3 Agent Datasets

Nemotron-Agentic-Tool-Use-v1 (NVIDIA, 2025) is a synthetic dataset designed to strengthen models' capabilities as interactive, tool-using agents. It focuses on multi-turn conversations where models decompose user goals and execute tool calls to complete tasks. For our experiments, we randomly sampled 20,000 trajectories from the non-thinking subset of this dataset to serve as the training set.

E.4 Multimodal Mathematical Datasets

DeepVision-103K (Sun et al., 2026) is a comprehensive, visually diverse multimodal mathematical dataset. It contains approximately 103K high-quality query-image-answer triplets covering a broad range of K12 mathematical topics, rich visual elements, and extensive knowledge points. By featuring verifiable, deterministic final answers, this dataset aims to enhance Large Multimodal Models' (LMMs) capabilities in visual perception, reflection, and mathematical reasoning.

F Benchmarks

In this section, we provide a detailed introduction to the benchmarks used in experiments.

F.1 General Benchmarks

- **MMLU (Hendrycks et al., 2020)** is a benchmark in evaluating the massive multitask language understanding and general knowledge capabilities of models. It consists of 57 distinct tasks. All of them are derived from various domains including STEM, the humanities, and the social sciences, ranging from elementary to professional levels. The problems in the MMLU dataset cover a wide variety of subjects such as elementary mathematics, US history, computer science, law, and medicine.

- **IFEval (Zhou et al., 2023)** is a benchmark designed to assess the instruction-following capabilities of large language models through objective and verifiable means. It consists of approximately 500 prompts containing various verifiable constraints, such as formatting requirements, word count limits, and keyword restrictions. Unlike benchmarks that rely on subjective human or model-based evaluation, IFEval employs a rule-based verification method to determine accuracy. This objective approach allows for a deterministic and reproducible evaluation of a model's ability to strictly adhere to complex user instructions.

- **CLUEWSC (Xu et al., 2020)** is a benchmark task designed to evaluate coreference resolution and common sense reasoning within the Chinese Language Understanding Evaluation (CLUE) framework. Adapted from the classic Winograd Schema Challenge, the dataset consists of difficult ambiguity resolution problems where a model must identify the correct antecedent of a pronoun in a sentence. The questions are constructed such that changing a single word in the sentence alters the correct reference. Therefore, solving CLUEWSC requires the model to utilize deep contextual understanding and general world knowledge, rather than relying on simple statistical associations or surface-level patterns.

F.2 Mathematical Benchmarks

- **AIME24 (AI-MO, 2025)** is a dataset in evaluating the mathematical reasoning ability of models. It consists of 30 challenging math problems. All of them are from the American Invitational Mathematics Examination. The problems in the AIME24 dataset cover a wide variety of mathematical fields such as algebraic equations and geometric puzzles. Due to the difficulty characteristics and the richness of question types, it has become a popular benchmark for evaluating the reasoning performance of models, and is widely used in multiple related research experiments.

- **AIME25 (Zhang and Math-AI, 2025)** consists of 30 challenging math problems. It is directly composed of the real questions from the American Invitational Mathematics Examination (AIME I & II) newly released in February

2025. AIME25's knowledge areas are extremely wide. It deeply covers core mathematical sections such as algebra, geometry, number theory, and combinatorial mathematics. This characteristic enables the AIME25 dataset to effectively distinguish the mathematical reasoning abilities of different models.

- **GSM8K (Cobbe et al., 2021)** is an elementary school math problem dataset released by OpenAI. These problems require 2 to 8 steps to solve, mainly through a series of basic calculations to obtain the final answer. This dataset is primarily used to test the logical and mathematical abilities of models and has been applied in multiple benchmark tests.

F.3 Medical Benchmarks

- **MedMCQA (Pal et al., 2022)** is a large-scale multi-subject dataset designed to assess the medical knowledge and reasoning capabilities of models. It comprises over 194,000 multiple-choice questions collected from prestigious Indian medical entrance examinations, such as AIIMS and NEET-PG. The dataset covers a vast spectrum of 21 medical subjects, ranging from basic biomedical sciences to advanced clinical disciplines like surgery and internal medicine. Given its high difficulty and broad professional coverage, MedMCQA serves as a crucial benchmark for evaluating how well large language models can handle complex healthcare scenarios.

- **PubMedQA (Jin et al., 2019)** is a biomedical question answering dataset designed to evaluate reasoning over scientific literature. It is derived from the titles and abstracts of research papers found in the PubMed database. The task requires models to answer research questions with "yes", "no", or "maybe" using the corresponding abstract as context. By focusing on evidence-based reasoning, PubMedQA effectively assesses a model's ability to interpret complex biomedical texts and draw accurate conclusions directly from scientific data.

- **MedQA (Jin et al., 2021)** is a large-scale dataset specifically designed to evaluate the clinical reasoning and professional medical knowledge of models. The dataset is derived from professional medical board examinations, with its English subset collected from the United States Medical Licensing Examination (USMLE). It consists of complex multiple-choice questions that simulate real-world clinical scenarios, covering diverse topics such as pathology, pharmacology, and patient management. Solving these problems requires deep domain knowledge and the ability to interpret patient case histories. Consequently, MedQA serves as a rigorous benchmark for determining whether models have achieved human-level competency in medicine.

F.4 Agent Benchmarks

- **BFCL v3 (Patil et al.)** is a comprehensive benchmark designed to evaluate the function-calling and tool-use capabilities of large language models. It comprises a diverse set of roughly 2K entries across multiple programming languages, including Python, Java, JavaScript, and REST APIs. The dataset assesses models on complex scenarios ranging from simple function calls to parallel, multiple, and nested calls, as well as multi-turn interactions. By employing an Abstract Syntax Tree (AST) evaluation method, BFCL v3 accurately measures a model's ability to generate syntactically correct and executable API calls, providing a more robust assessment than traditional string-matching metrics.

F.5 Multimodal General Benchmarks

- **MMMU (Yue et al., 2024)** is a massive multi-discipline multimodal understanding and reasoning benchmark designed to evaluate expert-level Artificial General Intelligence. It consists of thousands of challenging questions spanning over 30 distinct college-level subjects across science, engineering, arts, and medicine. Solving these problems requires models to deeply integrate domain-specific knowledge with complex visual perception and logical reasoning, making it a highly rigorous testbed for advanced vision-language models.

- **MMStar (Chen et al., 2024b)** is an elite multimodal benchmark carefully curated to assess the true vision-language capabilities of models. It specifically filters out questions that can be answered using text-only clues or linguistic priors, ensuring that purely visual perception is strictly necessary. The dataset covers a broad spectrum of visual tasks, forcing models to rely on deep, visually-dependent reasoning rather than surface-level text correlations.

- **RealWorldQA (xAI, 2024)** is a benchmark specifically designed to evaluate the spatial and physical understanding of multimodal representations in real-world scenarios. It consists of high-quality images captured from everyday environments paired with questions that demand accurate perception of object relations, spatial geometry, and physical common sense. This dataset is crucial for testing how well models can interact with and comprehend complex, unconstrained physical environments.

F.6 Multimodal Math Benchmarks

- **DynaMath (Zou et al., 2025)** is a dynamic mathematical reasoning benchmark designed to assess the robustness of multimodal models. Unlike static datasets, it generates multiple visual variants for mathematical problems to test if models can consistently apply mathematical logic under changing visual conditions. This requires models to not only understand the underlying mathematical concepts but also precisely map them to varied structural and visual representations.

- **MathVision (Wang et al., 2024)** is a comprehensive visual mathematical benchmark composed of problems from prestigious global mathematics competitions. The dataset features highly complex problems inherently bound to visual information, such as advanced geometry diagrams, function plots, and abstract visual logic puzzles. Because of its extreme difficulty and reliance on both deep mathematical deduction and precise visual interpretation, it serves as a top-tier evaluation suite for multimodal reasoning.

- **MathVista (Lu et al., 2024)** is an expansive benchmark that unified mathematical reasoning with rich visual contexts. It spans a wide array of task formats, including scientific diagrams, geographic maps, geometric figures, and mathematical tables. Evaluating on MathVista requires models to simultaneously perform accurate spatial perception, data extraction from images, and rigorous multi-step mathematical derivations.

G Models

Qwen3-4B-Instruct-2507 (Yang et al., 2025) is a lightweight large language model with 4 billion parameters released by Alibaba Cloud in July 2025. As the instruction-tuned variant of the Qwen3 series, it is engineered for efficient dialogue and general task execution, supporting a native context window of 256K tokens. Unlike its "Thinking" counterpart, this model prioritizes direct response generation without explicit chain-of-thought reasoning, optimizing it for low-latency performance and deployment on consumer-grade hardware.
Qwen3-4B-Thinking-2507 (Yang et al., 2025) is a reasoning lightweight model within the Qwen3 family, featuring 4 billion parameters. Unlike the standard instruction-tuned variant, this model is engineered to perform explicit chain-of-thought (CoT) reasoning, generating intermediate logical steps before producing a final answer. It supports a native context window of 256K tokens and is optimized for complex analytical tasks such as mathematical problem-solving and logical deduction.
Qwen2.5-32B-Instruct (Qwen et al., 2024) is a 32-billion parameter instruction-tuned large language model released by Alibaba Cloud in September 2024. Positioned as a mid-sized model within the Qwen2.5 series, it is designed to offer an optimal balance between computational efficiency and task performance. The model features a 64-layer Transformer architecture with Grouped Query Attention (GQA) and supports a context window of up to 128K tokens. It demonstrates significant improvements over its predecessors in instruction following, structured data understanding, and logical reasoning, making it highly suitable for deployment in resource-constrained environments that require high-quality generation.
GLM-9B-0414 (GLM et al., 2024) is a lightweight 9-billion parameter language model released by Zhipu AI (THUDM) in April 2025. As an iteration of the GLM-4 open-source series, this model version (0414) is specifically optimized for on-device deployment and agentic tasks. It features a native context window of 128K tokens and demonstrates state-of-the-art performance in tool use (function calling) and long-context understanding among models of similar size. The model is designed to provide a balance between inference latency and task complexity, making it a competitive alternative to larger 32B models for specific downstream applications.
Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025) is a flagship Mixture-of-Experts (MoE) model released by Alibaba Cloud. While it boasts a massive total parameter count of 235 billion, it utilizes a sparse architecture that activates only 22

billion parameters during inference. This design allows the model to achieve performance comparable to dense state-of-the-art models while maintaining the inference speed and computational efficiency of a much smaller 22B model.

Qwen3.5-2B (Qwen Team, 2026) is a next-generation lightweight multimodal language model featuring 2 billion parameters. Distinguished by its efficient hybrid architecture, it combines Gated Delta Networks with sparse Mixture-of-Experts to deliver high-throughput inference with minimal overhead. Built upon a unified vision-language foundation via early fusion training on multimodal tokens, it exhibits robust generalization in both text and visual tasks.

Qwen3-VL-4B-Thinking (Bai et al., 2025) is a reasoning-enhanced 4-billion parameter Vision-Language Model (VLM) from the latest Qwen3-VL series. It delivers comprehensive upgrades across visual perception, multimodal reasoning, and long-context processing. The "Thinking" edition specifically augments the model to excel in STEM and mathematical problem-solving by leveraging explicit causal analysis and extended logical reasoning chains.

H Baselines

- **SFT (Supervised Fine-Tuning):** The standard fine-tuning approach that maximizes the likelihood of target tokens using a uniform cross-entropy loss. It treats all tokens equally, making it prone to overfitting specific data patterns and "catastrophic forgetting" of general capabilities.
- **SFT$_{KL}$ (SFT with KL Regularization):** A prevalent robust baseline that adds a Kullback-Leibler (KL) divergence penalty to the loss function. It explicitly constrains the policy from deviating too far from the base model (reference model). While effective at preventing forgetting, it incurs significant memory overhead due to the need to maintain a frozen reference model.
- **FLOW (Sanyal et al., 2025):** A dynamic reweighting method that adjusts the importance of training samples based on their learning dynamics. FLOW monitors the loss trends to identify samples that are likely to cause forgetting, down-weighting them to maintain a smoother optimization trajectory compared to SFT.
- **DFT (Wu et al., 2025):** *Dynamic Fine-Tuning*. This method reinterprets SFT through a Reinforcement Learning lens, identifying that standard cross-entropy implicitly applies an unstable "inverse-probability weighting" ($1/\pi_\theta$) to gradients. DFT rectifies this by actively scaling the loss with the model's current prediction probability $\pi_\theta(y|x)$. This effectively dampens the gradients for low-probability target tokens (where the model is "surprised" or wrong), stabilizing optimization by preventing aggressive fitting of hard or noisy samples.
- **TALR (Lin et al., 2025):** *Token-Adaptive Loss Reweighting*. A granular approach that assigns varying weights to individual tokens based on their training difficulty (often measured by loss magnitude or gradient norms). TALR aims to focus the model's capacity on "hard" tokens while reducing the impact of easy or noisy tokens, though it typically lacks the uncertainty-awareness of our entropy-based method.

I Implementation Details

In this section, we provide a comprehensive overview of our training infrastructure, hyperparameter configurations, and evaluation protocols to ensure the reproducibility of our results.

I.1 Training Setup

All methods were implemented using **LLaMA-Factory** (Zheng et al., 2024), except for SFT$_{KL}$, which was trained using the **Verl** (Sheng et al., 2024) framework. All experiments were conducted on 8 NVIDIA A100 GPUs.

I.2 Hyperparameter Settings

To ensure a fair comparison, we adopted a unified set of hyperparameters across all baseline models and our approach, unless specified otherwise. The specific values for the shared hyperparameters are detailed in Tab. 6.

Regarding checkpoint selection, we prioritized the checkpoint that demonstrated the best mean performance averaged over all the evaluated benchmarks.

Method-Specific Hyperparameters. The KL-divergence coefficient β in SFT$_{KL}$ was set to 0.5.

I.3 Evaluation Protocol

For all evaluations, we conducted three independent runs and reported the average performance.

For mathematical reasoning benchmarks, we strictly followed the evaluation implementation of

Hyperparameter	Value
Learning Rate	1×10^{-5}
LR Scheduler	Cosine
Warm-up Steps	0.03 ratio
Optimizer	AdamW
Batch Size	64
Num of Epochs	10
Max Sequence Length	16384

Table 6: Hyperparameter settings used for all models.

Qwen2.5-Math (Yang et al., 2024). For medical domain tasks, we aligned our evaluation pipeline with the **MedEvalKit** framework (Xu et al., 2025).

J Extended Discussion

In this section, we elaborate on the design philosophy of the gating mechanism and position EAFT within the broader landscape of alignment techniques (e.g., RL (Shao et al., 2024) and RFT (Touvron et al., 2023)).

J.1 Comparison with EAFT$_{sig}$

While Section 5.1 demonstrates that Sigmoid-based gating (**EAFT**$_{sig}$) achieves competitive performance, we advocate for the linear formulation primarily due to its **hyperparameter robustness** and **ease of deployment**.

Sensitivity of Non-Linear Gating. The Sigmoid function introduces two sensitive hyperparameters: steepness (α) and the centering threshold (β).

- **High Sensitivity:** Performance fluctuates significantly with β. If β is too high, the method degenerates into standard SFT; if too low, it mimics a Hard Mask, discarding valuable training signals. This necessitates expensive grid searches for every new dataset or model scale.
- **Binary Bias:** High steepness (α) forces a near-binary decision boundary, ignoring uncertainty and treating moderately confident tokens the same as extremely confident ones.

The "Parameter-Free" Advantage. The linear formulation acts as a structural prior: *loss weight is directly proportional to uncertainty.* It effectively removes the need for hyperparameter tuning. This "out-of-the-box" robustness ensures EAFT generalizes across domains without requiring the delicate calibration needed for Sigmoid-based variants.

J.2 Comparison with RL and RFT

Existing alignment methods like On-policy RL (Schulman et al., 2017; Shao et al., 2024) or Rejection Sampling Fine-Tuning (RFT) (Touvron et al., 2023) mitigate the "alignment tax" but incur high computational or operational costs (Chen et al., 2025; Shenfeld et al., 2025). EAFT provides a distinct trade-off favoring efficiency.

Lightweight Efficiency (vs. RL). RL methods typically triple memory requirements by maintaining multiple models simultaneously (Policy, Reference, and Value/Reward models). **EAFT** retains the nearly identical memory footprint and computational graph of standard SFT. Crucially, it requires **no Reference Model** (using the model's own entropy as a proxy for "trustworthiness") and avoids the complex rollout generation phase, significantly accelerating training throughput.

Dynamic Adaptation (vs. RFT). Methods like RFT rely on *static* data curation—filtering or generating data based on the model's capability *before* training begins.

- **Static vs. Dynamic:** RFT assumes a fixed knowledge boundary. However, as the model learns, its uncertainty shifts.
- **On-the-fly Correction:** EAFT operates *dynamically*. If the model becomes overconfident about a hallucination mid-training, EAFT automatically down-weights the loss for that specific token. This allows the gating mechanism to evolve in real-time alongside the model parameters, offering an adaptive advantage that static data filtering cannot achieve.

J.3 Nature of Forgetting: Semantic Degradation vs. Style Overwriting

To clarify whether downstream forgetting stems from genuine capability degradation or superficial style overwriting, we conduct a quantitative error analysis on the MMLU, IFEval, and CLUEWSC benchmarks. Focusing on SFT models in Math and Agent domains, we isolate "newly degraded" instances—questions answered correctly by the base model but incorrectly after SFT. An advanced LLM (Gemini-3-pro (Google DeepMind, 2026)) is then leveraged to classify these failure modes into formatting mismatches (*Fmt_Err*) and critical semantic errors (*Sem_Err*).

As shown in Table 7, a striking 97% of these new failures are genuine semantic errors rather than superficial formatting shifts. This indicates that while "Confident Conflicts" involve formatting tokens, forcing updates on these rigid priors destabilizes the model's deeper representation space, triggering a ripple effect that leads to profound semantic degradation. By suppressing these destructive gradients, EAFT successfully preserves the model's core knowledge base.

Domain	Total	Fmt_Err	Sem_Err
Math	864	20	844 (98%)
Agent	946	30	916 (97%)
Overall	**1810**	**50**	**1760 (97%)**

Table 7: Quantitative error analysis of newly degraded instances after standard SFT.

K LLM Usage

We used large language models (LLMs) to improve the clarity and grammatical correctness of the manuscript. After using these tools, the authors carefully reviewed and edited all generated content, and take full responsibility for the final version of the paper.