Restoring Calibration for Aligned Large Language Models: A Calibration-Aware Fine-Tuning Approach

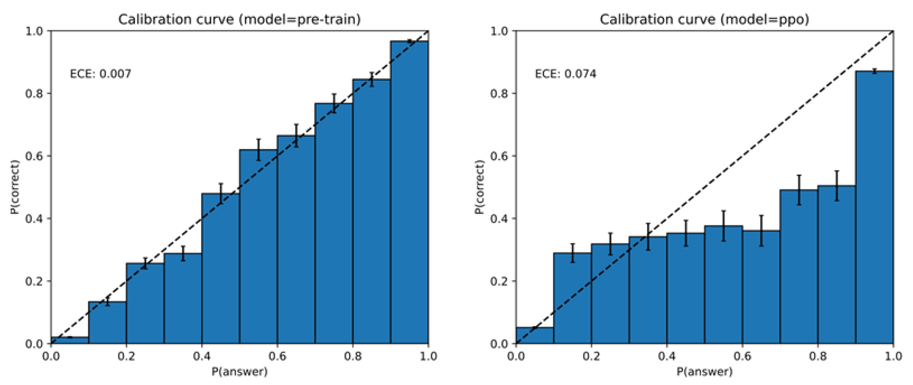
Jiancong Xiao*, Bojian Hou*, Zhanliang Wang*, Ruochen Jin, Qi Long^, Weijie J. Su^, Li Shen^

* Equal contribution: jcxiao@upenn.edu, bojianh@upenn.edu, aaronwzl@sas.upenn.edu

^ Correspondence to: qlong@upenn.edu, suw@upenn.edu, lishen@upenn.edu

Introduction

Preference Alignment leads to Poor Generalization (GPT-4 technical Report)



- **Calibration**: A model is well-calibrated when its confidence matches its accuracy
- **Calibration of LLMs:** evaluated on a multiple-choice question and its corresponding correct answer

Universal Issue across: Models, Alignment approaches

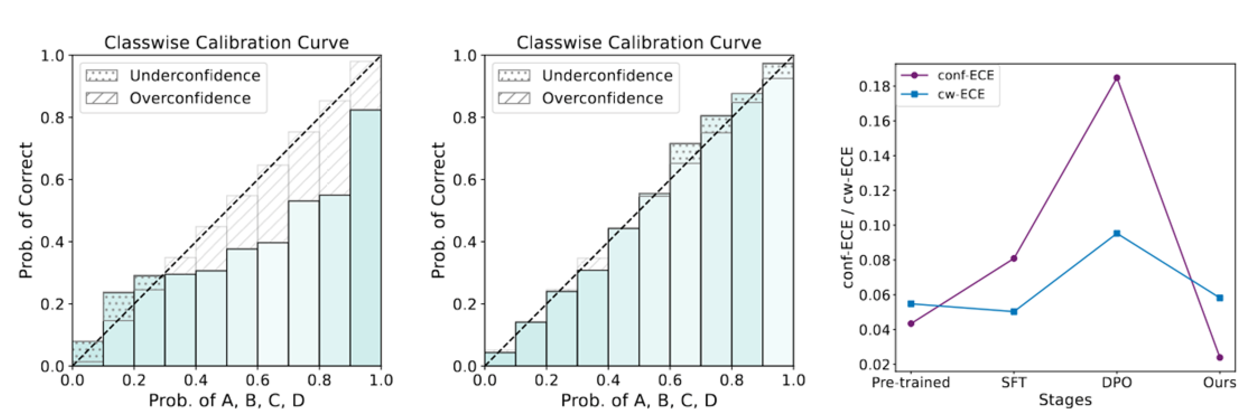
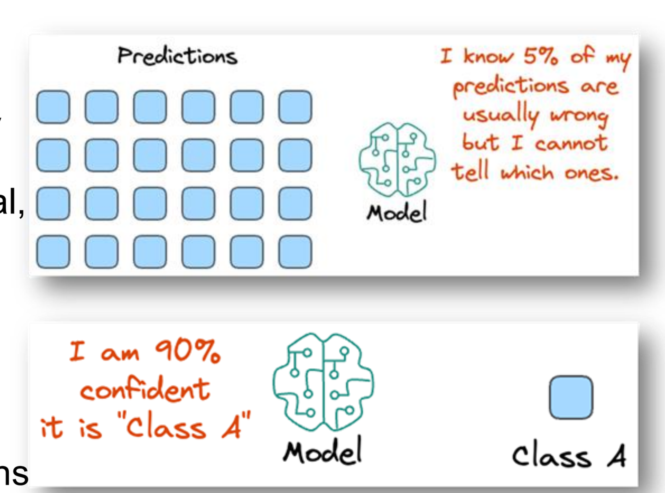


Figure 1. Calibration performance comparison between DPO and our approach on Llama3.1-8B-Tulu (a DPO-aligned version of Llama-3.1 (Touvron et al., 2023)). Left: Model calibration plots after DPO alignment, showing significant overconfidence. Middle: Calibration plots after applying our fine-tuning approach, demonstrating improved calibration. Right: The evolution of confidence ECE and classwise ECE across different stages (pre-trained, SFT, DPO, and our method) shows how our approach effectively restores calibration errors.

Why Good Calibration is Important?

- Blindly trusting ML model predictions can be fatal in high-stakes environments
- Despite high accuracy (e.g., 95%), models cannot identify which predictions are incorrect
- Assessing confidence for individual predictions is essential, not just accuracy rates
- In medical diagnostics and other high-risk scenarios, neglecting prediction confidence can lead to severe consequences
- Understanding and quantifying prediction uncertainty is crucial for responsible implementation
- Modern models tend to be overconfident in their predictions [1,2], which must be addressed in model design



[1] Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Hein, Matthias, Maksym Andriushchenko, and Julian Bitterwolf. "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Calibration Definitions

Definition 3.1 (Classwise Calibration). A probabilistic classifier $\hat{\boldsymbol{p}} : \mathcal{X} \to \Delta_k$ is classwise-calibrated, if for any class j and any predicted probability q_j for this class:

$$\mathbb{P}(y = j | \hat{p}_j(x) = q_j) = q_j.$$

Classwise-ECE (cw-ECE) is defined as:

$$\text{cw-ECE} = \mathbb{E}_{\hat{\boldsymbol{p}}(x)} \frac{1}{k} \sum_{j=1}^{k} \big|\mathbb{P}(y = j|\hat{p}_j(x)) - \hat{p}_j(x)\big|.$$

Definition 3.2 (Confidence Calibration). A probabilistic classifier $\hat{\boldsymbol{p}} : \mathcal{X} \to \Delta_k$ is confidence-calibrated, if for any $c \in [1/k, 1]$:

$$\mathbb{P}(y = \operatorname{argmax} \hat{\boldsymbol{p}}(x) | \max \hat{\boldsymbol{p}}(x) = c) = c.$$

Confidence-ECE (conf-ECE) is defined as:

$$\mathbb{E}_{\hat{\boldsymbol{p}}(x)} \frac{1}{k} \sum_{j=1}^{k} \big|\mathbb{P}(y = \arg\max \hat{\boldsymbol{p}}(x)| \max \hat{\boldsymbol{p}}(x)) - \max \hat{\boldsymbol{p}}(x)\big|.$$

Key Research Questions

1. Why does preference alignment affect calibration?
2. How can we restore calibration while maintaining the benefits of alignment?

Key Finding - Preference Collapse

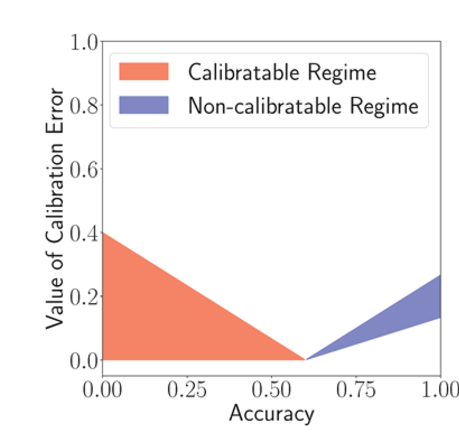
- **Preference Collapse Phenomenon**:
  - Definition: Aligned models excessively favor certain responses over others
  - Results: Preference ratio exceeding human preference proportions
  - π(y_w|x)/(π(y_w|x) + π(y_l |x)) > P(y_w≻y_l| x)
- **Multiple-Choice Generalization**:
  - Collapse appears with strong preference for one option (A/B/C/D)
  - Leads to high confidence regardless of correctness
- **Empirical Evidence**:
  - Observed across Llama3.1, Vicuna, Olmo2, and Mistral models which will be demonstrated in the following experimental results

Theoretical Framework - Probabilistic Generative Model

- **Generative View of Multiple-Choice QA**:
  - Data Generation: Test designer creates a probabilistic distribution over correct answers
  - This test designer can be regarded as a probabilistic generative model
- **Proposition**: Probabilistic generative models are inherently well-calibrated
  - Since these models generate the positions of correct answers according to their probability distributions, the observed accuracy always equals the model's confidence (i.e., its predicted probabilities).
- **Target Probabilistic Generative Model**:
  - Definition: Optimal solution maximizing accuracy under perfect calibration

Calibratable Regime v.s.Non Calibratable Regime

- **Calibration Regime**:
  ECE can reach zero without sacrificing accuracy (or broadly LLM Performance)
- **Non-Calibration Regime**:
  Fundamental trade-off between ECE and accuracy (or broadly LLM Performance)



(R)CFT: An EM-based Algorithm

- **SFT Loss:**
  Questioning Comprehension
  Accuracy
  Instruction Following

$$\mathcal{L}_{\text{SFT}_1} = -\log \pi(y_i|x_i).$$

- **SFT2 Loss:**
  Further Question understanding

$$\mathcal{L}_{\text{SFT}_2} = -\Big[\log \pi(y|x) + \sum_{t=2}^{T} \log \pi(x^t|x^{t-1}, \ldots, x^1)\Big]$$

- **ECE loss:**
  Calibration controlling

$$\mathcal{L}_{\text{ECE}} = \text{D}(\boldsymbol{p}(x), \text{conf}_\pi(x)),$$

Algorithm 1 (Regularized) Calibration-Aware FT

Require: Number of epochs L, Number of bins M;
Initialize model π_0 by the alligned LLMs;
for $l = 0$ *to* L **do**
  E-Step: *// Use max confidence to stratify samples*
  for $i = 1 : n$ **do**
    for $m = 1 : M$ **do**
      if $\max \text{conf}_{\pi_l}(x_i) \in (\frac{m-1}{M}, \frac{m}{M}]$; **then**
        $z_i = m$; *// z_i is defined as the latent variable*
      end
    end
  end
  M-Step: *// Calibrate model towards accuracy*
  for $m = 1 : M$ **do**
    $\mathcal{S}_m = \{(x_i, y_i) | z_i = m, i = 1, \ldots, n\}$;
    $q_m = \frac{1}{|\mathcal{S}_m|} \sum_{(x,y)\in\mathcal{S}_m} \mathbb{1}(\operatorname{argmax} \text{conf}_{\pi_l}(x) = y)$;
  end
  Update $\boldsymbol{p}(x_i)$ by Equation (6), $i = 1, \ldots, n$;
  $\pi_{l+1} = \frac{1}{n} \sum_{i=1}^{n} \min_\pi [\mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{ECE}}(\boldsymbol{p}(x_i), \pi_l(x_i))]$;
end

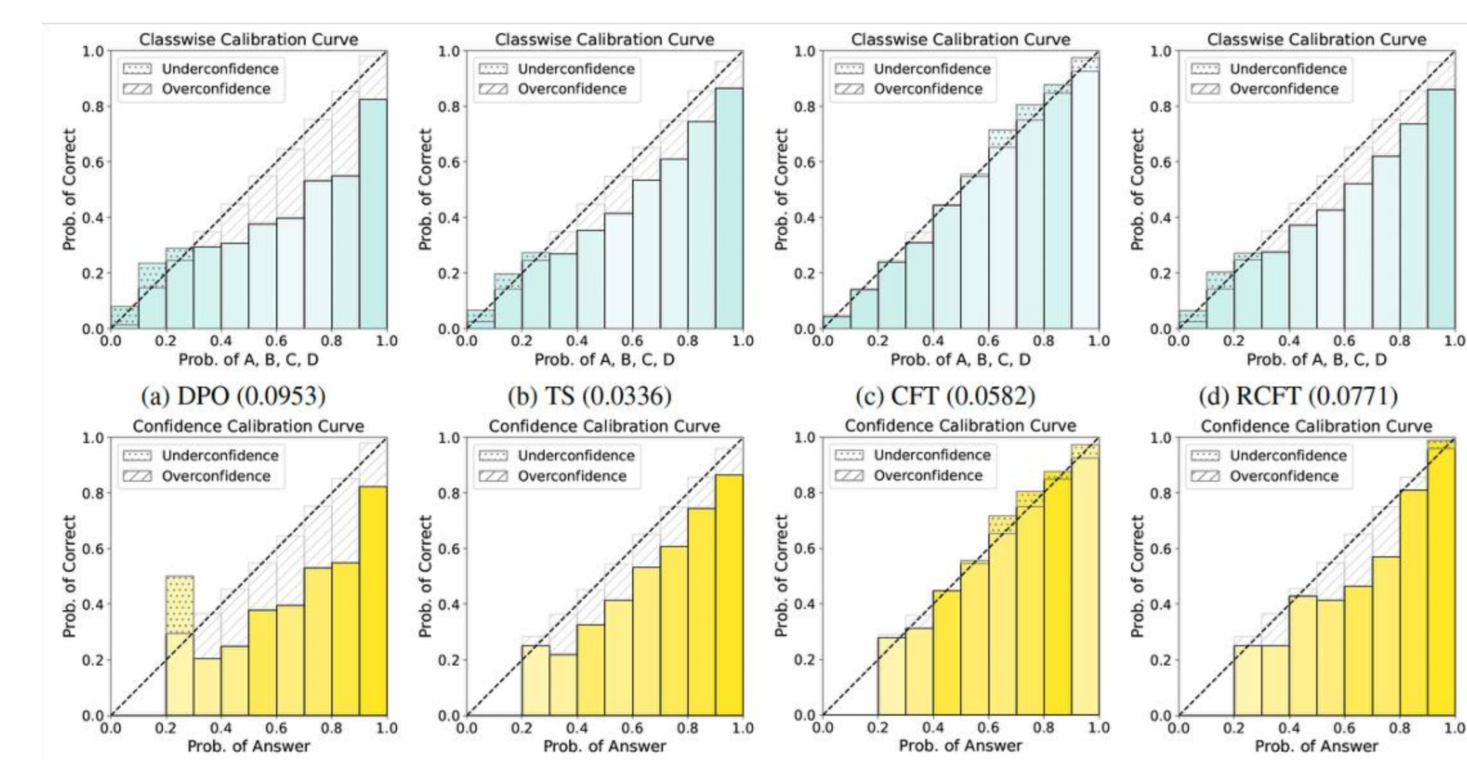
Experiments

Table 2. Performance comparison among DPO/RLHF, Temperature Scaling, Label Smoothing, CFT, and RCFT across four models (**Llama3.1-8B-Tulu**, **Vicuna-7B**, **Olmo2-7B**, and **Mistral-7B**) in in-domain and out-domain scenarios. Best results in each metric block are bold. Blue highlights indicate superior in-domain conf-ECE of our CFT while red highlights denote best in-domain accuracy of our RCFT. "↓"/"↑" means the smaller/larger the better. "-" means the results of Temp. Scale. are the same as the original DPO/RLHF version.

Model	Method	conf-ECE ↓ In-Domain	conf-ECE ↓ Out-Domain	cw-ECE ↓ In-Domain	cw-ECE ↓ Out-Domain	Accuracy ↑ In-Domain	Accuracy ↑ Out-Domain
Llama3.1-8B-Tulu	DPO	0.1953	0.1212	0.0953	0.0650	0.6228	0.7810
	Temp. Scale.	0.1126	**0.0679**	**0.0336**	0.0514	-	-
	Label Smooth.	0.1898	0.1009	0.0692	0.0639	0.6372	0.7116
	CFT(Ours)	**0.0239**	0.0688	0.0582	**0.0375**	0.6410	**0.8000**
	RCFT(Ours)	0.0897	0.0810	0.0771	0.0526	**0.8341**	0.7991
Vicuna-7B	RLHF	0.1422	0.0852	0.0979	0.0560	0.4344	0.5233
	Temp. Scale.	0.0598	**0.0224**	0.0488	**0.0484**	-	-
	Label Smooth.	0.1221	0.0823	0.0517	0.0544	0.4517	0.5767
	CFT(Ours)	**0.0379**	0.0331	0.0583	0.0491	0.4481	0.6172
	RCFT(Ours)	0.0474	0.0672	**0.0459**	0.0530	**0.6015**	**0.6035**
Olmo2-7B	DPO	0.1555	0.1325	0.0873	0.1331	0.6210	0.6635
	Temp. Scale.	0.0665	0.1160	**0.0355**	0.1196	-	-
	Label Smooth.	0.1010	0.0499	0.0791	0.1298	0.6808	0.6431
	CFT(Ours)	**0.0544**	**0.0225**	0.0804	**0.0637**	0.6606	0.7085
	RCFT(Ours)	0.0989	0.0781	0.0806	0.0707	**0.8510**	**0.7099**
Mistral-7B	DPO	0.2010	0.1318	0.0909	0.1103	0.6331	0.7567
	Temp. Scale.	0.0802	0.0991	**0.0399**	0.0909	-	-
	Label Smooth.	0.1874	0.1121	0.0900	0.0990	0.6479	0.6997
	CFT(Ours)	**0.0651**	**0.0424**	0.0712	**0.0614**	0.6514	**0.7863**
	RCFT(Ours)	0.0979	0.0731	0.0877	0.0739	**0.8297**	0.7768

Table 3. Win rate comparisons among DPO/RLHF (DPO used in Table 3), CFT and RCFT across four models (Llama3.1-8B-Tulu, Vicuna-7B, Olmo2-7B and Mistral-7B) on three datasets (AlpacaEval, Arena-Hard and Ultrafeedback). The best performance for each dataset is in bold. The competitive performance indicates that our methods can preserve the alignment performance.

Model	AlpacaEval (vs DPO) CFT vs DPO	AlpacaEval (vs DPO) RCFT vs DPO	AlpacaEval DPO	AlpacaEval CFT	AlpacaEval RCFT	Arena-Hard DPO	Arena-Hard CFT	Arena-Hard RCFT	Ultrafeedback DPO	Ultrafeedback CFT	Ultrafeedback RCFT
Llama-3.1-8B-Tulu	51.68 vs 48.32	46.83 vs 53.16	21.4	**22.6**	19.6	44.6	**45.0**	43.6	0.7295	**0.7460**	0.7118
Vicuna-7B	46.46 vs 53.54	50.43 vs 49.57	2.60	2.60	**3.60**	1.00	1.00	1.00	0.2271	**0.2279**	0.2257
Olmo2-7B	62.48 vs 37.52	46.12 vs 53.88	**24.2**	22.9	23.1	19.4	19.2	**20.2**	0.7493	**0.7588**	0.7517
Mistral-7B	46.96 vs 53.04	49.81 vs 50.19	26.0	**26.8**	25.2	**18.9**	18.3	18.0	0.7066	0.7124	**0.7221**



(a) DPO (0.0953) (b) TS (0.0336) (c) CFT (0.0582) (d) RCFT (0.0771)

(e) DPO (0.1953) (f) TS (0.1126) (g) CFT (0.0239) (h) RCFT (0.0897)

Figure 5. Calibration Plots of (a, e) DPO, (b, f) Temperature Scaling (TS), (c, g) our CFT, (d, h) our RCFT on Llama-3.1-8B-Tulu. (a-d) are the classwise calibration curve and (e-h) are the confidence calibration curve. Each panel plots the model's predicted probabilities (i.e., confidence) on the x-axis against the observed accuracy (fraction correct) on the y-axis, binned into ten groups. The diagonal line in each panel represents *perfect calibration*. The depth of the color indicates the sample density in that column. DPO has the worst calibration performance. Other three methods improve the calibration performance where our CFT has the lowest con-ECE (shown in the parenthesis). The figures of conf-ECE (e-h) omit the first two bins because the model selects an answer with the largest predicted probability which is always larger than 0.25 in the four options prediction task (so no samples exist below that threshold).

Take-away Messages

- **Preference-aligned Model lies in the Calibratable Regime**
  CFT can restore calibration without sacrificing LLM performance
- **Overly fine-tune models,, they shift into the non-calibratable regime**
  RCFT navigate the trade-off between ECE and LLM performance

Github: https://github.com/PennShenLab/RestoreLLMCalibration

Acknowledgements: This work was supported in part by NIH grants U01CA274576 and P30AG073105, ARPA-H Award D24AC00253, NSF grant DMS-2310679, a Meta Faculty Research Award, Wharton AI for Business, and a PSOM AI2D Seeding Project.