CALIBRATING LANGUAGE MODELS WITH ADAPTIVE TEMPERATURE SCALING

Johnathan Xie[*] Stanford University
Annie S. Chen[*] Stanford University
Yoonho Lee Stanford University
Eric Mitchell Stanford University
Chelsea Finn Stanford University

ABSTRACT

The effectiveness of large language models (LLMs) is not only measured by their ability to generate accurate outputs but also by their calibration—how well their confidence scores reflect the probability of their outputs being correct. While unsupervised pre-training has been shown to yield LLMs with well-calibrated conditional probabilities, recent studies have shown that after fine-tuning with reinforcement learning from human feedback (RLHF), the calibration of these models degrades significantly. In this work, we introduce Adaptive Temperature Scaling (ATS), a post-hoc calibration method that predicts a temperature scaling parameter for each token prediction. The predicted temperature values adapt based on token-level features and are fit over a standard supervised fine-tuning (SFT) dataset. The adaptive nature of ATS addresses the varying degrees of calibration shift that can occur after RLHF fine-tuning. ATS improves calibration by over 10-50% across three downstream natural language evaluation benchmarks compared to prior calibration methods and does not impede performance improvements from RLHF.

1 INTRODUCTION

Large language models (LLMs) have become a cornerstone of modern artificial intelligence, offering impressive capabilities in natural language processing tasks. However, the reliability of LLMs is intertwined with their ability to generate confidence scores that accurately reflect the likelihood of their outputs being correct. This calibration, aligning a model's confidence with its accuracy, is essential, especially when LLMs are deployed in real-world scenarios where decisions based on incorrect outputs can have significant consequences.

While unsupervised pre-training methods have shown success in producing well-calibrated LLMs, a challenge arises when these models undergo fine-tuning through reinforcement learning from human feedback (RLHF). While RLHF fine-tuning is effective in enhancing model performance on specific tasks and aligning outputs with human preferences, recent studies indicate a notable degradation in the calibration of LLMs post-RLHF fine-tuning (Achiam et al., 2023; Tian et al., 2023; Kadavath et al., 2022). This degradation compromises the model's ability to provide reliable confidence scores, an issue that becomes critical when these models are applied to tasks requiring high levels of trust and accuracy. An important question arises: how can we maintain the performance gains achieved through RLHF fine-tuning while ensuring that the model's confidence scores remain reliable?

To address this challenge, our work introduces Adaptive Temperature Scaling (ATS), a post-hoc calibration technique that predicts a temperature scaling parameter for each token prediction based on a language model's hidden features. Basic temperature scaling is a widely-used calibration method that applies a single temperature parameter across all outputs of a model. This technique, while effective in some contexts, assumes uniform calibration needs across all inputs, which is often not the case for complex models like LLMs. ATS, in contrast, predicts a unique temperature scaling parameter for each set of token predictions. This input-specific approach allows ATS to refine the calibration process, addressing the varying degrees of calibration shift that can occur after RLHF fine-

[*]Equal contribution. Correspondence to jwxie@stanford.edu and asc8@stanford.edu.

tuning. For instance, certain inputs or topics might be more susceptible to miscalibration post-RLHF, and ATS can adaptively adjust the scaling for these instances more aggressively than for others where the model's confidence remains relatively well-aligned with its accuracy. Importantly, our approach reduces the need for task-specific calibration, which may be difficult to achieve in many cases, given the wide variety of downstream tasks that LLMs may be used for.

We conduct experiments on MMLU, TriviaQA, and TruthfulQA to evaluate the effectiveness of ATS in improving the calibration of LLMs following RLHF fine-tuning. Our findings demonstrate that ATS improves the calibration of post-RLHF LLMs by 10-50% on average, while having no effect on model outputs.

2 Related Work

Recent literature has extensively discussed the challenges of maintaining calibration in LLMs, particularly highlighting the degradation in calibration post-RLHF (Lin et al., 2022; Park and Caragea, 2022; Kadavath et al., 2022; Xiao et al., 2022; Kuhn et al., 2023). The concept of verbalized confidence has been explored as a way to counteract this degradation (Xiong et al., 2023; Tian et al., 2023), and dialogue models have been shown to express uncertainty in a well-calibrated manner (Mielke et al., 2022; Zhou et al., 2023). Unlike these approaches, we focus on recalibrating the conditional probabilities of LLMs post-RLHF.

The calibration of neural networks has been a topic of significant interest, with foundational concepts such as proper scoring rules (Gneiting et al., 2007) laying the groundwork. Model mismatch and distribution shift often degrade calibration, common quantified with common metrics including Expected Calibration Error (ECE) (Naeini et al., 2015) and Brier score (Brier, 1950). Modern neural networks have been found to exhibit overconfidence (Guo et al., 2017; Thulasidasan et al., 2019; Wen et al., 2020), especially in the context of image classification (Geirhos et al., 2018; Taori et al., 2020; Wen et al., 2020; Hendrycks et al., 2021).

Various methods have been proposed for calibrating neural networks, including temperature scaling (Guo et al., 2017), Platt scaling (Platt et al., 1999; Niculescu-Mizil and Caruana, 2005), label smoothing (Müller et al., 2019), and more sophisticated approaches (Hendrycks et al., 2018; Katz-Samuels et al., 2022; Choi et al., 2023; Jiang et al., 2023). While these methods offer strategies for improving model calibration, our approach uniquely adapts the temperature scaling parameter for each token prediction based on its hidden features, tailoring the method to the problem of language modeling.

3 Background and Problem Setting

We consider access to a conversation SFT dataset of $\mathcal{D} = \{(x, y)\}$ with vocabulary V where $x \in V^{l_x}$, denotes the instruction, each with sequence length l_x, and $y \in V^{l_y}$ is the corresponding response with sequence length l_y. We wish to calibrate language model $\pi(y|x)$. While we do not make any assumptions about the training process of π, we find our calibration method is most useful for language models following an RLHF process where token-level calibration is often significantly degraded compared to base language models which are generally well calibrated Achiam et al. (2023).

For a given sample (x, y), we generate a set of unnormalized logits $\hat{z} = \pi(x) \in \mathbb{R}^{l_x + l_y \times |V|}$ where each $\hat{z}_i$ defines the unnormalized logits for the $i + 1$-th token and $|V|$ is the vocabulary size. Prior methods Guo et al. (2017); Platt et al. (1999) propose various scaling methods for calibrating models by transforming logits. In matrix scaling, a calibration head is used to produce calibrated logits $\hat{q} = W\hat{z} + b$ where W, b are learnable parameters. In the case of language modeling where $|V|$ is large, learning a full transform matrix becomes computationally infeasible, so we compare to vector scaling, where W is constrained to a diagonal matrix. Temperature scaling is the case when W is constrained further to a scalar matrix and b to the zero-vector. To learn these parameters, these methods minimize the cross-entropy over the SFT dataset calculated over response tokens.

4 Adaptive Temperature Scaling

Architecture. Temperature scaling, while effective in classification settings, struggles to adapt logits well in language modeling as the confidence scores that are most important (such as those that contain actual answers or facts) account for only a small portion of natural language sequences. Therefore, optimizing a single temperature parameter often results in post-RLHF language models still being overconfident post scaling. Additionally, language model miscalibration largely varies

Model	Calibration	MMLU			TriviaQA			TruthfulQA		
		Acc	ECE	BS	Acc	ECE	BS	Acc	ECE	BS
Llama-2-7b-Chat (Touvron et al., 2023)	None	0.474	0.298	0.313	0.592	0.221	0.239	0.322	0.507	0.480
	Temperature	0.474	0.270	0.295	0.592	0.187	0.224	0.322	0.492	0.463
	Vector Scaling	0.474	0.324	0.333	0.592	0.211	0.234	0.322	0.499	0.471
	ATS (Ours)	0.474	**0.125**	**0.227**	0.592	**0.069**	**0.217**	0.322	**0.197**	**0.264**
Qwen-7b-Chat (Bai et al., 2023)	None	0.571	0.141	0.215	0.495	0.272	0.311	0.230	0.372	0.304
	Temperature	0.571	0.093	0.215	0.495	0.269	0.308	0.230	0.313	0.262
	Vector Scaling	0.571	0.144	0.218	0.495	**0.252**	0.308	0.230	0.369	0.302
	ATS (Ours)	0.571	**0.050**	**0.190**	0.495	0.254	**0.303**	0.230	**0.165**	**0.188**

Table 1: **Model Calibration Comparison**. We find that ATS yields significant improvements over other calibration methods for both LLama-2-7b-Chat and Qwen-7b-Chat.

loss	ECE	BS
no smoothing	0.226	0.269
full smoothing	0.149	0.236
selective	**0.125**	**0.227**

Table 2: **Smoothing type**. Selective smoothing outperforms cross-entropy (no smoothing) and full smoothing (standard label smoothing).

α	ECE	BS
0.1	0.197	0.254
0.2	0.172	0.243
0.3	0.151	0.236
0.4	0.134	0.231
0.5	0.125	0.227
0.6	**0.113**	**0.224**

Table 3: **Loss weighting**. A high smooth loss weight is necessary to correct for language model overconfidence.

head	ECE	BS
linear	0.140	0.233
mlp	0.132	0.230
transformer	**0.125**	**0.227**

Table 4: **Head architecture**. We use that using a Transformer head in the same configuration as LLaMa-2-7b-Chat performs best.

based on the type of token being predicted following RLHF. Matrix and vector scaling can in theory perform adaptive confidence prediction by using logits as features; however, they are prone to overfitting, as we find in Sec. 5.

To balance regularization with modeling capacity in our calibration head, we instead propose to use a head architecture that predicts a singular temperature for every token prediction. For an input pair (x, y), we first produce input-dependent features $\hat{h} \in \mathbb{R}^{l_x+l_y,h}$ using the language model π. We then learn a calibration head to produce a temperature vector $c_\theta(\hat{h}) = \tau \in \mathbb{R}^{l_x+l_y}$. We exponentiate τ to ensure positive values then transform logits to yield calibrated logits $\hat{q} = \hat{z} \circ e^\tau$. In practice, we find that directly using the logits $\hat{z}$ as features can be inefficient (with a large vocabulary size) and also less effective compared to hidden states. Therefore, we use the last hidden state of the language model π as the features for predicting τ. With this architecture formulation we retain the ability to predict confidences adaptively depending on the context, while also never changing the "ranking" of predictions as each set of token logits are scaled by only a single value.

Loss function. To improve the process of calibration, we take inspiration from selective classification works (Choi et al., 2023) and use a loss function which adapts targets depending on the correctness of the original language model. For a logit, label pair $\hat{q} \in \mathbb{R}^v$, $y \in V$, and weighting hyperparameter $\alpha \in [0, 1]$ we optimize the following loss function ℓ:

$$\ell(\hat{q}, y) = \begin{cases} -(1-\alpha)\log\left(\sigma_{SM}(\hat{q})_y\right) & \arg\max \hat{q} = y \\ -\frac{\alpha}{|V|}\sum_{i=1}^{|V|}\log(\sigma_{SM}(\hat{q}))_i & \arg\max \hat{q} \neq y \end{cases} \tag{1}$$

This loss function uses a uniform distribution as the target when the model is incorrect and a standard one-hot cross-entropy when the model is correct.

5 Experiments

In this section, we aim to evaluate our proposed method on multiple benchmarks to demonstrate its effectiveness in improving calibration of LLMs fine-tuned with RLHF. We compare our method to no calibration as well as existing temperature scaling methods. Additionally, we ablate the main components of our method including the loss function, loss weighting, and head architecture.

Evaluation Setting. We evaluate using two 7B parameter post-RLHF models LLama-2-Chat-7b Touvron et al. (2023) and Qwen-Chat-7b. As the calibration dataset, we use the Alpaca GPT-

4 Peng et al. (2023) instruction tuning dataset, which contains a diverse set of instructions with high quality answers. We then evaluate model calibration on three downstream tasks.

We perform multiple choice evaluation on the MMLU Hendrycks et al. (2020) by aggregating statistics across the entire dataset. Specifically we concatenate the confidences and correctness labels from all subjects, then calculate the calibration metrics. We also evaluate on two free response datasets, TriviaQA Joshi et al. (2017) and TruthfulQA Lin et al. (2021).

Metrics. In multiple choice inference, we have a set of tokens ids O which represent the valid options for a multiple choice answer, so the confidence scores are $p = \sigma_{SM}(\hat{q}_{l_x, j \in O})$ where σ_{SM} denotes the softmax function. To calculate confidences over a long sequence of response tokens for an input x, we sample a generation $\hat{y}$ of length $l_{\hat{y}}$ from the original language model then concatenate to the instruction to form $\hat{z}$ and $\hat{q}$ following calibration. Then, we calculate an average over transition probabilities on the response tokens. We use the Expected Calibration Error (ECE) Guo et al. (2017) and Brier score Brier (1950) to evaluate calibration. We also report accuracy but each method does not significantly affect accuracy.

Baselines. We compare our method to the post-RLHF model without calibration, temperature scaling, and vector scaling. We do not evaluate matrix scaling as the full matrix becomes computationally infeasible for large vocabulary sizes.

5.1 Results

We report the results of our method compared to the baselines in Table 1. Overall, we find that our method improves calibration by 10-50% across the three benchmarks in terms of ECE and Brier Score compared to the next best method for both LLama-2-7b-Chat and Qwen-7b-Chat. More specifically, for Llama-7b-Chat, applying ATS achieved the lowest ECE and BS across all downstream benchmarks, showing how adjusting the temperature scaling parameter for each token prediction can significantly improve calibration. Qwen-7b-Chat also saw a significant improvement in calibration, although in the case of TriviaQA, ATS actually makes Qwen-7b-Chat slightly underconfident compared to vector scaling. Importantly, the calibration dataset used for training ATS, Alpaca GPT-4, is unrelated to the downstream tasks evaluated on, which suggests that the method does not merely overfit to the calibration data but rather captures underlying predictive uncertainty principles that are applicable across various tasks.

5.2 Ablation Studies

To analyze our method, we ablate the main components: loss objective, loss weight, and head architecture, measuring calibration metrics on MMLU.

Loss objective. We compare different loss objectives, standard cross-entropy, cross-entropy with label smoothing, and selective smoothing (ours) in Table 2. For label smoothing we performed a sweep and found a smoothing value of 0.3 to be optimal. We find that selective smoothing outperforms both the typical cross-entropy loss and label smoothing. One possible explanation for cross-entropy and standard label smoothing being less effective is that learning adaptive temperature values with a cross-entropy loss can actually cause the model to increase confidence when the model is incorrect. In comparison, by using a uniform distribution target for incorrect predictions, this will never happen.

Loss weight. We perform a sweep of smooth loss weight in Table 3. While increasing the loss weight to 0.6 (compared to 0.5) benefits MMLU calibration, in practice we found this higher loss weight began to perform worse for TriviaQA, and we did not sweep higher values as the model begins to become underconfident.

Head architecture. In Table 4, we ablate the choice of head architecture. We find that a causal transformer layer identical to those used in the LLama-2-7b-chat model performs best. Given that the inference cost of a single additional layer is relatively negligible, using a full transformer layer is generally best for calibration performance as it can aggregate hidden state values from prior tokens for the specific task of predicting calibration.

6 LIMITATIONS

While ATS offers a significant improvement in model calibration without compromising post-RLHF performance by adapting the temperature scaling parameter based on token-level features of each input, limitations remain. In particular, ATS does not address semantic uncertainty or verbalized confidence, and per-token log-probabilities may be unavailable in some RLHF-LMs. These limitations underscore the need for ongoing research to refine calibration techniques and incorporate a more nuanced understanding of semantic uncertainty to develop methods that allow models to express confidence in a manner that aligns with natural language.

REFERENCES

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [page 1, 2]

J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. [page 3]

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950. [page 2, 4]

C. Choi, F. Tajwar, Y. Lee, H. Yao, A. Kumar, and C. Finn. Conservative prediction via data-driven confidence minimization. *arXiv preprint arXiv:2306.04974*, 2023. [page 2, 3]

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. [page 2]

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007. [page 2]

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. [page 2, 4]

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. [page 2]

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. [page 4]

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. [page 2]

M. Jiang, Y. Ruan, S. Huang, S. Liao, S. Pitis, R. B. Grosse, and J. Ba. Calibrating language models via augmented prompt ensembles. 2023. [page 2]

M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. [page 4]

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. [page 1, 2]

J. Katz-Samuels, J. B. Nakhleh, R. Nowak, and Y. Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022. [page 2]

L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023. [page 2]

S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021. [page 4]

S. Lin, J. Hilton, and O. Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022. [page 2]

S. J. Mielke, A. Szlam, E. Dinan, and Y.-L. Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872, 2022. [page 2]

R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. [page 2]

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. [page 2]

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. [page 2]

S. Y. Park and C. Caragea. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. *arXiv preprint arXiv:2203.07559*, 2022. [page 2]

B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. [page 4]

J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. [page 2]

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599, 2020. [page 2]

S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. [page 2]

K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023. [page 1, 2]

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [page 3]

Y. Wen, G. Jerfel, R. Muller, M. W. Dusenberry, J. Snoek, B. Lakshminarayanan, and D. Tran. Combining ensembles and data augmentation can harm your calibration. *arXiv preprint arXiv:2010.09875*, 2020. [page 2]

Y. Xiao, P. P. Liang, U. Bhatt, W. Neiswanger, R. Salakhutdinov, and L.-P. Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*, 2022. [page 2]

M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023. [page 2]

K. Zhou, D. Jurafsky, and T. Hashimoto. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*, 2023. [page 2]

A Appendix

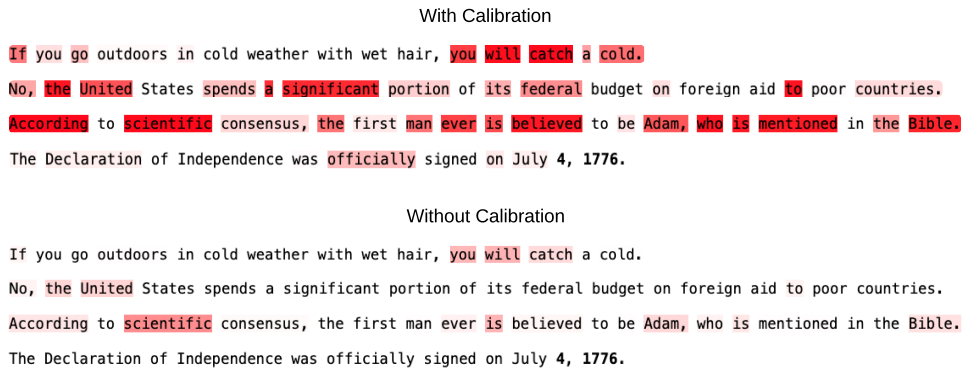

Figure 1: **Calibration Visualization.** We visualize confidence calibration samples, comparing token-wise confidences before and after calibration. The less confident a token is, the more red we highlight the background. Additionally, we average the confidences of tokens to form full words in order to create a more interpretable visualization.

A Confidence Visualizations

In Figure 1, we compare confidence calibration on TruthfulQA dataset samples. We compare the Llama-2-7b-chat model without any calibration to after calibration with our method. Our method is able to cause the language model to become significantly less confident on tokens containing inaccuracies.

B Hyperparameters

config	value
optimizer	AdamW
optimizer betas	β_1, β_2=0.9, 0.999
weight decay	0.0
learning rate	$5e-5$
learning rate schedule	cosine decay
epochs	2
batch size	8

Table 5: Calibration training hyperparameters.

In Table 5 we list the main hyperparameters used for training calibration methods over Alpaca GPT-4.