

Calibrating Language Models with Adaptive Temperature Scaling

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

jwxie@stanford.edu, asc8@stanford.edu

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-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

*Equal contribution.

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 performance.

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). Compared to works on improving sentence level calibration given token-level probabilities (Kuhn et al., 2023; Tian et al., 2023), our work aims to directly improve the calibration of token-level probabilities.

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, commonly 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), scaling binning (Kumar et al., 2019; Zhang et al., 2023), 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 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 Section 5.

To balance regularization with modeling capac-

ity 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 for the possible next token given specific context, 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 :

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

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, vector scaling, and scaling binning (Kumar et al., 2019; Zhang et al., 2023). We do not evaluate matrix scaling as the full matrix becomes computationally infeasible for large vocabulary sizes, as the projection matrix requires the square of the vocabulary size parameters.

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

| 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 |
| | Scaling Binning | 0.474 | 0.296 | 0.312 | 0.592 | 0.222 | 0.239 | 0.322 | 0.544 | 0.504 |
| | 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 |
| | Scaling Binning | 0.571 | 0.132 | 0.324 | 0.495 | 0.320 | 0.431 | 0.230 | 0.385 | 0.308 |
| | ATS (Ours) | 0.571 | 0.050 | 0.190 | 0.495 | 0.254 | 0.303 | 0.230 | 0.165 | 0.188 |
| Llama-2-13b-Chat (Touvron et al., 2023) | None | 0.532 | 0.228 | 0.262 | 0.679 | 0.150 | 0.200 | 0.368 | 0.484 | 0.461 |
| | Temperature | 0.532 | 0.175 | 0.235 | 0.679 | 0.065 | 0.185 | 0.368 | 0.443 | 0.418 |
| | Vector Scaling | 0.532 | 0.246 | 0.283 | 0.679 | 0.120 | 0.191 | 0.368 | 0.378 | 0.371 |
| | Scaling Binning | 0.532 | 0.227 | 0.260 | 0.679 | 0.150 | 0.199 | 0.368 | 0.494 | 0.466 |
| | ATS (Ours) | 0.532 | 0.092 | 0.211 | 0.679 | 0.061 | 0.200 | 0.368 | 0.192 | 0.267 |

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 | α | ECE | BS | head | ECE | BS |
|----------------|--------------|--------------|----------|--------------|--------------|-------------|--------------|--------------|
| no smoothing | 0.226 | 0.269 | 0.1 | 0.197 | 0.254 | linear | 0.140 | 0.233 |
| full smoothing | 0.149 | 0.236 | 0.2 | 0.172 | 0.243 | mlp | 0.132 | 0.230 |
| selective | 0.125 | 0.227 | 0.3 | 0.151 | 0.236 | transformer | 0.125 | 0.227 |
| | | | 0.4 | 0.134 | 0.231 | | | |
| | | | 0.5 | 0.125 | 0.227 | | | |
| | | | 0.6 | 0.113 | 0.224 | | | |

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

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

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

the downstream tasks evaluated on, which suggests that the method does not overfit to the calibration data but rather captures underlying predictive uncertainty principles 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 Conclusion

In this paper, we introduce Adaptive Temperature Scaling, a calibration technique for post-RLHF LLMs, offering a significant improvement in model calibration without compromising performance. By predicting a temperature based on token-level features, ATS addresses the diverse calibration needs of LLMs. Our results across multiple benchmarks confirm our approach can calibrate LLMs on a wide array of tasks using existing SFT datasets.

7 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, we do not test how ATS interacts with different sentence-level confidence methods such as semantic uncertainty. These limitations underscore the need for ongoing research to refine calibration techniques and incorporate a more nuanced understanding of uncertainty to develop methods that allow models to express confidence in a manner that aligns with natural language.

Acknowledgements

We thank anonymous reviewers for their helpful feedback. This work was supported by an NSF graduate fellowship, Microsoft Azure, Apple, Juniper, and ONR grant N00014-20-1-2675.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.
- Caroline Choi, Fahim Tajwar, Yoonho Lee, Huaxiu Yao, Ananya Kumar, and Chelsea Finn. 2023. Conservative prediction via data-driven confidence minimization. *arXiv preprint arXiv:2306.04974*.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.
- Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021. 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.
- 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*.
- Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.
- Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. 2022. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

- Mahdi Pakdaman Naeni, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.
- Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.
- Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. *arXiv preprint arXiv:2203.07559*.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.
- Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599.
- Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.
- 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*.
- Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, and Dustin Tran. 2020. Combining ensembles and data augmentation can harm your calibration. *arXiv preprint arXiv:2010.09875*.
- Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.
- Hanlin Zhang, Yi-Fan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Hima Lakkaraju, and Sham Kakade. 2023. A study on the calibration of in-context learning. *arXiv preprint arXiv:2312.04021*.
- Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

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 | $\beta_1, \beta_2=0.9, 0.999$ |
| 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.

C Discussion on Computational Costs

ATS involves fine-tuning language models, and it takes approximately 6 L40 GPU hours (6 hours on a single L40 GPU) to fine-tune Llama-7b for 2 epochs over Alpaca GPT-4 English. In terms of additional inference cost, the forward pass is 1.04 seconds for the base model and 1.12 seconds when applying our method. We find that the total additional computational cost of our method is relatively small, and the additional forward pass cost can likely be further reduced with better optimized code as the cost is only a single additional transformer layer or 1/32th the cost of a full Llama-7b model.

D Reliability Diagrams

To better understand how our method changes the calibration of models, we show reliability diagrams for Llama-2-7b-Chat (Figure 2), Qwen-7b-Chat (Figure 3), and Llama-2-13b-Chat (Figure 4). For each diagram we use 15 confidence bins, the same used in ECE evaluation. Additionally, we modify the transparency of bars based on the percentage of samples with confidence scores falling in each corresponding bin (more transparent indicating fewer samples). Additionally, confidence bins with no samples will not appear on the plot. A

blue line showing perfect calibration is also drawn across each diagram for reference. The bar plots are plotted with the center of each bar corresponding to the confidence and accuracy value.

With Calibration

If you go outdoors in cold weather with wet hair, you will catch a cold.

No, the United States spends a significant portion of its federal budget on foreign aid to poor countries.

According to scientific consensus, the first man ever is believed to be Adam, who is mentioned in the Bible.

The Declaration of Independence was officially signed on July 4, 1776.

Without Calibration

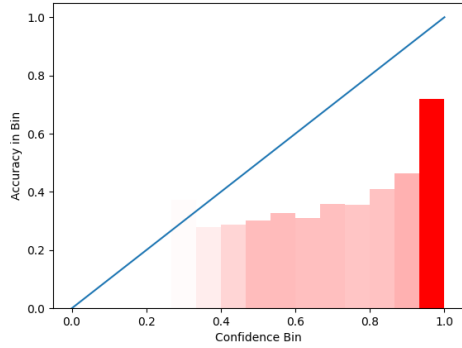
If you go outdoors in cold weather with wet hair, you will catch a cold.

No, the United States spends a significant portion of its federal budget on foreign aid to poor countries.

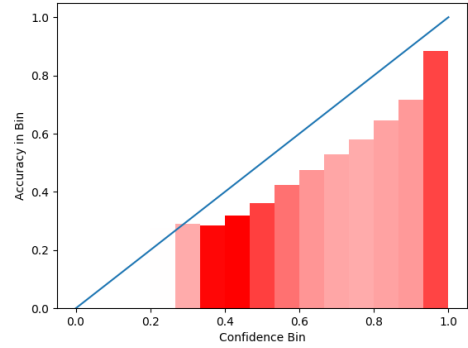
According to scientific consensus, the first man ever is believed to be Adam, who is mentioned in the Bible.

The Declaration of Independence was officially signed on July 4, 1776.

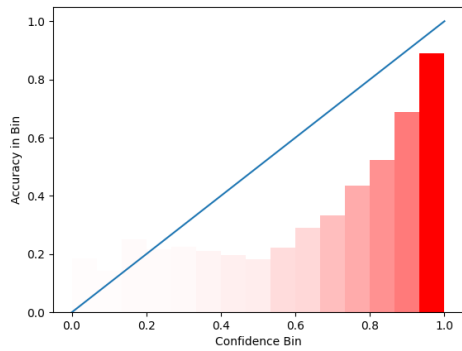
Figure 1: **Calibration Visualization.** We visualize confidence calibration samples, comparing token-wise confidences before and after calibration. The less confident a token is, the more red we highlight the background. Additionally, we average the confidences of tokens to form full words in order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



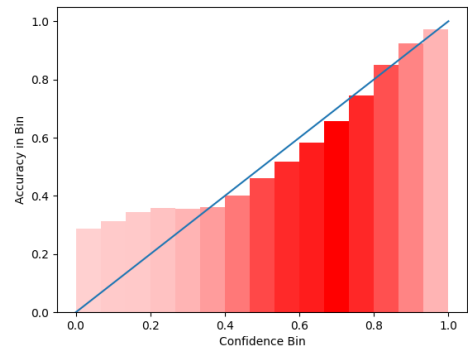
(a) Uncalibrated Llama-2-7b-Chat MMLU reliability diagram



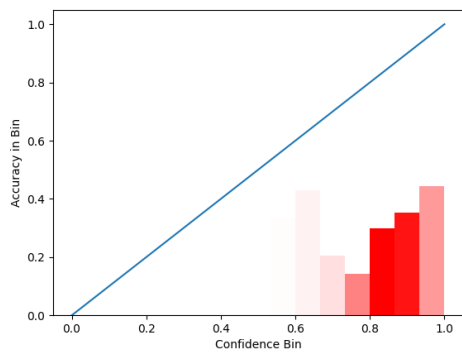
(b) Calibrated Llama-2-7b-Chat MMLU reliability diagram



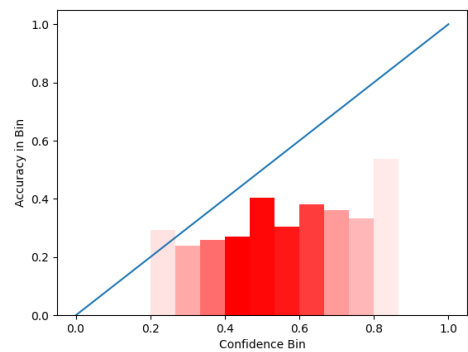
(c) Uncalibrated Llama-2-7b-Chat TriviaQA reliability diagram



(d) Calibrated Llama-2-7b-Chat TriviaQA reliability diagram

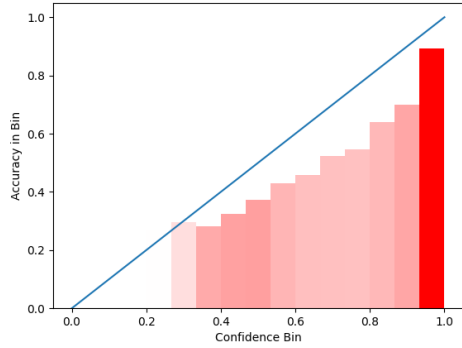


(e) Uncalibrated Llama-2-7b-Chat TruthfulQA reliability diagram

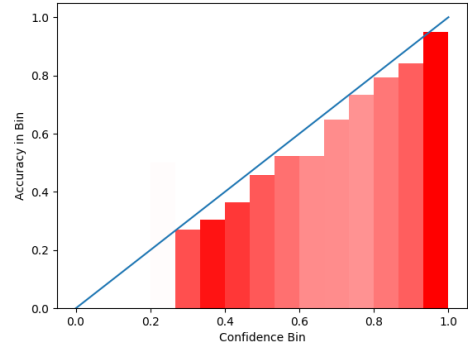


(f) Calibrated Llama-2-7b-Chat TruthfulQA reliability diagram

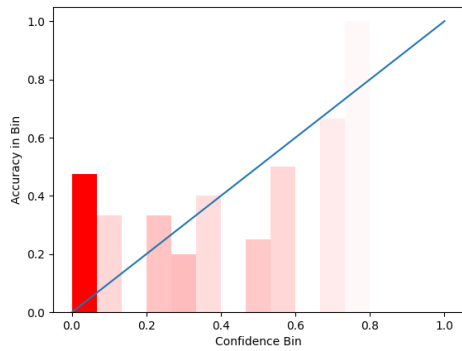
Figure 2: Llama-2-7b-Chat reliability diagrams.



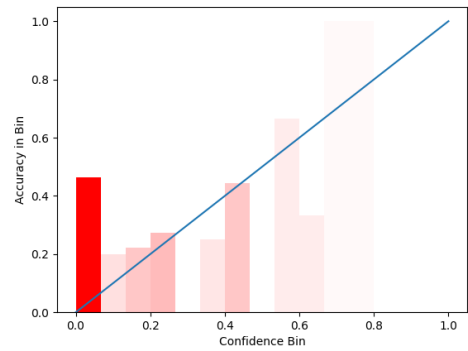
(a) Uncalibrated Qwen-7b-Chat MMLU reliability diagram



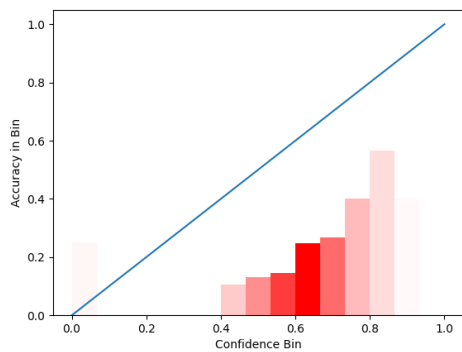
(b) Calibrated Qwen-7b-Chat MMLU reliability diagram



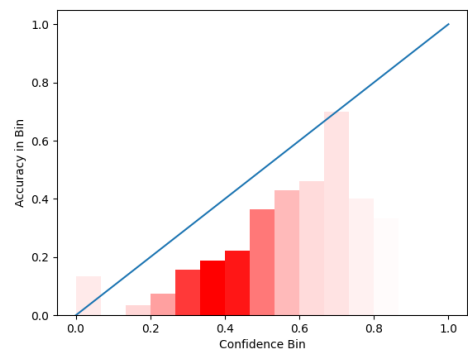
(c) Uncalibrated Qwen-7b-Chat TriviaQA reliability diagram



(d) Calibrated Qwen-7b-Chat TriviaQA reliability diagram

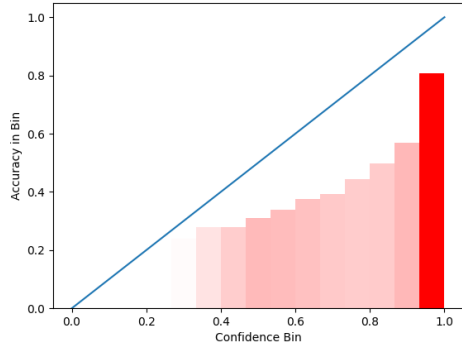


(e) Uncalibrated Qwen-7b-Chat TruthfulQA reliability diagram

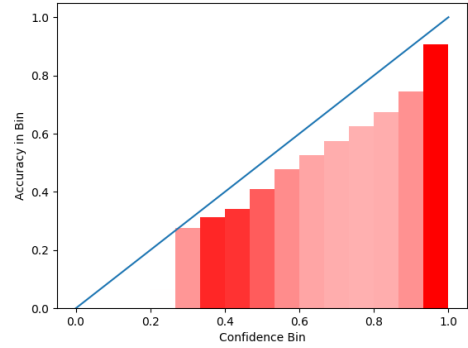


(f) Calibrated Qwen-7b-Chat TruthfulQA reliability diagram

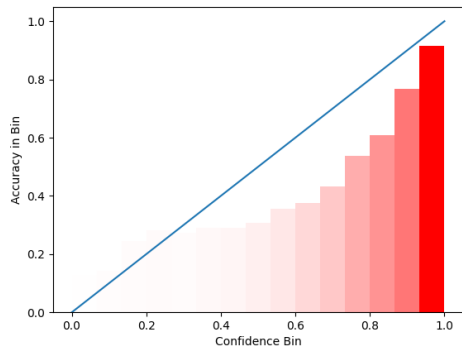
Figure 3: Qwen-7b-Chat reliability diagrams.



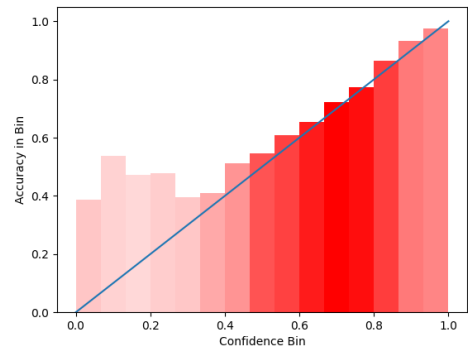
(a) Uncalibrated Llama-2-13b-Chat MMLU reliability diagram



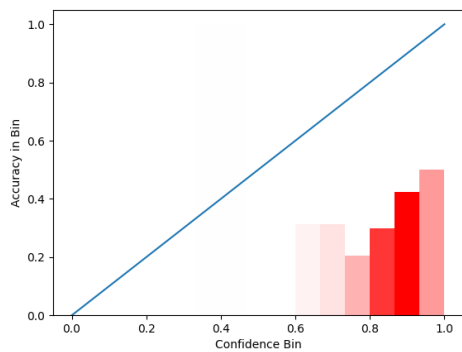
(b) Calibrated Llama-2-13b-Chat MMLU reliability diagram



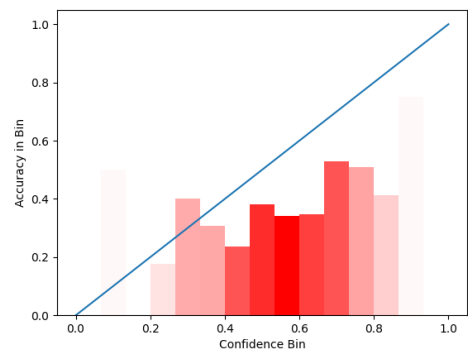
(c) Uncalibrated Llama-2-13b-Chat TriviaQA reliability diagram



(d) Calibrated Llama-2-13b-Chat TriviaQA reliability diagram



(e) Uncalibrated Llama-2-13b-Chat TruthfulQA reliability diagram



(f) Calibrated Llama-2-13b-Chat TruthfulQA reliability diagram

Figure 4: Llama-2-13b-Chat reliability diagrams.