

Confidence Drift During Task-Specific Language Model Fine-Tuning

Soham Tripathy

Viven AI

soham@viven.ai

Abstract

Task-specific fine-tuning can change not only what a language model predicts, but also how confident it is in those predictions. We study confidence and accuracy drifts throughout LoRA fine-tuning of seven decoder-only language models spanning multiple families and scales. Models are fine-tuned on AG News and evaluated at intermediate checkpoints on the source task and five untouched cross-task benchmarks. We find that cross-task accuracy generally remains stable or degrades, while confidence increases on both correct and incorrect predictions. Across many settings, confidence follows a non-monotonic trajectory: after an early transient and a relatively stable middle stage, it rises again late in training, coinciding with increasing train-validation NLL divergence. Our results show that accuracy-only evaluation can miss confidence drifts introduced by task-specific fine-tuning.

1 Introduction

Large Language Models (LLMs) have started to play a pivotal role in modern NLP applications and are being used in industry-grade applications for customer interaction and decision support (Chkribene et al., 2024). Yet there remain cases where a human still needs to remain in the loop. As the volume of decisions handled by LLMs continues to grow, manually reviewing each one becomes impractical. This makes calibration: how well an LLM’s confidence aligns with its actual accuracy, a crucial property for real-world applications (Geng et al., 2023). Ideally, an LLM should be confident precisely when it is right, and uncertain when it isn’t. High-confidence errors can be particularly problematic because confidence is often used to determine whether a prediction should be trusted or reviewed.

Calibration has received growing attention as a key indicator of the reliability of language models.

Prior work has examined the calibration properties of pretrained language models (Chen et al., 2023), while Durai (2025) studied how calibration evolves over the course of pretraining. However, LLMs are seldom used only on the task on which they were adapted. In practice, a task-specific fine-tuned model may subsequently be applied to inputs requiring different capabilities or belonging to entirely different tasks. This motivates studying calibration under cross-task shift: whether adaptation on a source task changes the model’s confidence on benchmarks that provide no fine-tuning signal. Unlike conventional domain shift, our setting changes the task itself, moving from news classification to question answering, common sense completion, physical reasoning, and reading comprehension.

In this work, we take a systematic look at how predictive performance (accuracy in our case) and confidence evolve over the course of task-specific fine-tuning. We track model checkpoints at intermediate stages of fine-tuning, evaluating each one not only on the source task but also on several untouched cross-task benchmarks on various metrics.

Our Key Findings

- ➔ *Cross-task accuracy stays roughly the same or deteriorates, but confidence generally increases across fine-tuning checkpoints.*
- ➔ *Confidence goes up on both correct and incorrect predictions; hence calibration can shift significantly even when accuracy barely moves.*
- ➔ *Confidence follows a non-monotonic trajectory across checkpoints: it rises early in fine-tuning, enters a relatively stable phase, and increases again during the late-stage overfitting regime, which we identify through the divergence between training and validation.*

2 Related Work

Modern neural networks are frequently miscalibrated despite high predictive accuracy, motivating post-hoc corrections such as temperature scaling (Guo et al., 2017). In NLP, pretrained Transformers are relatively well calibrated in-domain but degrade under distribution shift (Desai and Durrett, 2020), and strong task performance need not imply reliable confidence in larger generative models (Zhu et al., 2023). Fine-tuning itself alters calibration: it can induce severe miscalibration both in- and out-of-distribution (Kong et al., 2020), partly through the loss of pretrained features during downstream adaptation (He et al., 2023). Recent work documents tuning-induced miscalibration (Shum et al., 2024), overconfidence arising from the interaction between a model’s prior knowledge and its fine-tuning data (Wang et al., 2025), and calibration degradation from instruction tuning (Huang et al., 2025), including confidence gains without comparable accuracy gains across languages (Huang et al., 2026). From a temporal perspective, confidence can continue to rise on both correct and incorrect predictions rather than self-calibrating with further training (Chen et al., 2023), and pretraining exhibits distinct phases in which confidence and calibration move in opposite directions (Durai, 2025). Most recent work (Xiao et al., 2025) studies calibration degradation following preference alignment and proposes a calibration-aware fine-tuning methods to restore calibration. Their focus is primarily on explaining and mitigating miscalibration in RLHF/DPO-aligned models.

These studies either compare pretrained against final fine-tuned checkpoints or track pretraining. We instead track a dense sequence of intermediate checkpoints *during* downstream adaptation, and ask whether confidence drifts on benchmarks that receive no fine-tuning signal at all - treating fine-tuning as a continuous process rather than an endpoint transformation. Checkpoint-level analysis is practically useful because the final fine-tuned checkpoint is not necessarily the most reliable one. We find that confidence can continue to increase after task accuracy has saturated, including on incorrect cross-task predictions (which is detrimental). Tracking intermediate checkpoints can therefore inform early stopping and checkpoint selection,

and reveal reliability regressions that accuracy-only evaluation would miss.

3 Methodology

3.1 Problem Setup

Let $\mathcal{M}_{\theta_0}$ be a pretrained autoregressive language model. We fine-tune it on a source task $\mathcal{T}_s$ and save intermediate checkpoints $\{\theta_0, \dots, \theta_T\}$, where θ_0 is the base model and θ_T is the final fine-tuned model. Each checkpoint is evaluated on the source task and on a set of cross-task benchmarks $\{\mathcal{T}_1, \dots, \mathcal{T}_k\}$.

We adapt the model with LoRA (Hu et al., 2022), training a small subset of adapter parameters on the attention and MLP projections while the pretrained weights θ_0 remain frozen. We focus on LoRA because it is a widely used parameter-efficient adaptation method and keeps compute and training time low. The loss is applied at the answer position *only*; all prompt positions are masked out and hence the objective reduces to the classification cross-entropy:

$$\mathcal{L}(\theta) = -\log p_{\theta}(v(c_y) \mid x), \quad (1)$$

where $v(c_y)$ is the verbalizer token corresponding to the gold answer c_y . This ensures that no gradient flows from modeling the prompt tokens themselves.

All tasks are formulated as multiple-choice prediction. For an input x with candidates $\mathcal{C} = \{c_1, \dots, c_K\}$, we obtain a model score z_j for each candidate and renormalize across the set of candidates:

$$p_{\theta}(c_j \mid x) = \frac{\exp(z_j)}{\sum_{k=1}^K \exp(z_k)}. \quad (2)$$

The score z_j is the logarithmic probability of a *single* verbalizer token immediately after the Answer: cue, taken from a softmax over the full vocabulary. It is not a sum or mean over a multi-token continuation, and it is not length-normalized. For each model’s tokenizer, we verify that every verbalizer maps to a single distinct token. The predicted answer is the highest-probability candidate, and its probability is used as the model’s confidence in that prediction. Throughout this work, confidence refers to the probability obtained after renormalizing over the benchmark’s candidate answer set, rather than probability over the model’s full vocabulary. Consequently, absolute confidence values are not directly comparable across tasks with different numbers of answer choices. We therefore

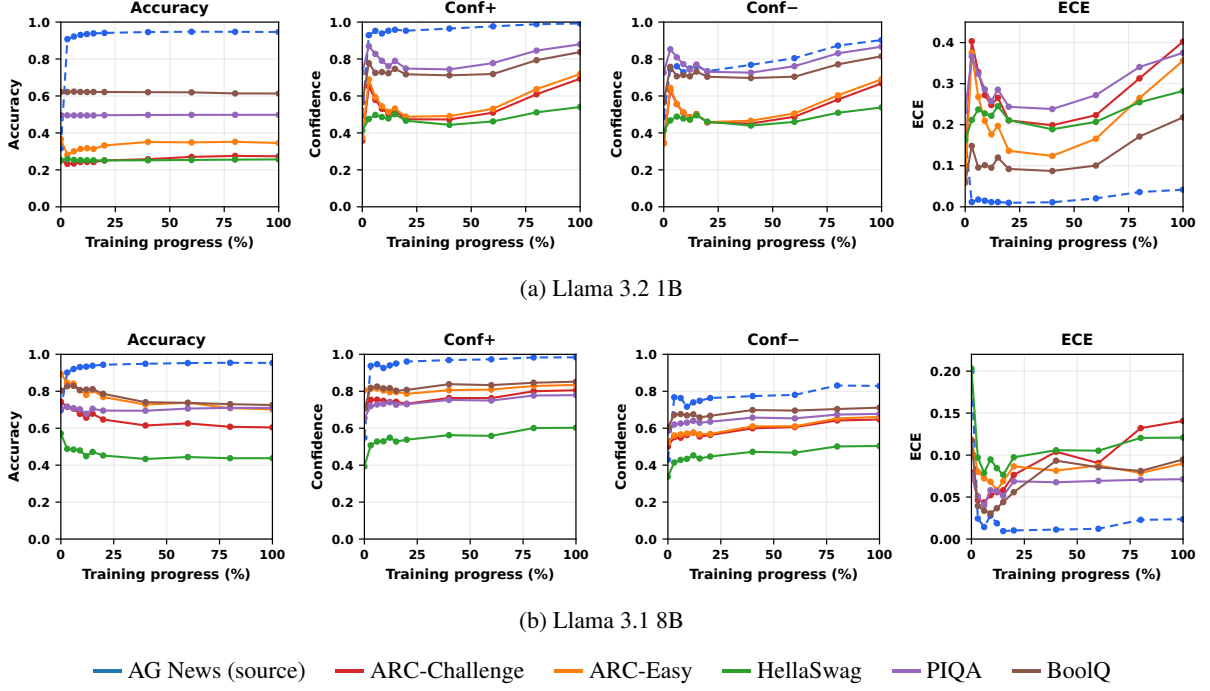


Figure 1: Trajectory metrics (mean across 3 seeds for each model) at checkpoints.

interpret confidence primarily through within-task changes across fine-tuning checkpoints of a single model, where the candidate set is fixed.

Thus, our confidence statistics (Conf^+ and Conf^-) and ECE are computed with respect to the prompted candidate set rather than the model’s unconstrained next-token distribution. We do not aggregate probability across alternative tokenizations or surface forms of the same answer (e.g., "A", "A", or "A."), so this form of tokenization ambiguity is not modeled by our evaluation. In open-ended generation, where multiple surface forms may represent the same semantic answer and substantial probability mass may lie outside a predefined candidate set, defining and measuring calibration becomes substantially more difficult.

3.2 Datasets

We fine-tune on AG News (Zhang et al., 2015), a four-way news topic classification dataset with 120,000 training and 7,600 test articles. AG News is a good source task for this study because it is large enough to expose a long fine-tuning trajectory and has clear objective answers to questions.

Each checkpoint is additionally evaluated on five benchmarks that are never seen during fine-tuning, chosen to span distinct reasoning abilities and different numbers of answer choices. These cover science question answering (ARC-Easy and ARC-

Challenge; Clark et al., 2018), commonsense completion (HellaSwag; Zellers et al., 2019), physical reasoning (PIQA; Bisk et al., 2020), and reading comprehension (BoolQ; Clark et al., 2019). Full dataset descriptions, example items, and choice counts are provided in Appendix A.

3.3 Evaluation Metrics

For example i , let p_i be the candidate-renormalized distribution of Eq. 2, $\hat{p}_i = \max_j p_{ij}$ its confidence, and $a_i \in \{0, 1\}$ whether the prediction is correct.

Confidence. We report mean confidence separately on correct and incorrect predictions; because fine-tuning may increase confidence on both correct and incorrect predictions.

$$\text{Conf}^+ = \frac{\sum_i a_i \hat{p}_i}{\sum_i a_i}, \quad \text{Conf}^- = \frac{\sum_i (1 - a_i) \hat{p}_i}{\sum_i (1 - a_i)}, \quad (3)$$

Negative log-likelihood. A proper scoring rule over the full predictive distribution rather than only its argmax,

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^N \log p_{i, y_i}. \quad (4)$$

Expected calibration error. Following Guo et al. (2017), examples are binned by confidence into

$M = 15$ equal-width bins $\{B_m\}$, and

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (5)$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the mean correctness and mean confidence in bin m . Lower is better.

4 Experimentation

4.1 Model Selection

We consider a range of decoder-only models spanning multiple families and scales in order to test whether our findings hold beyond a single architecture. The models include Llama-3.2-1B and Llama-3.1-8B, Qwen3-1.7B, Qwen3-4B and Qwen3-8B, SmolLM2-360M and SmolLM2-1.7B.

4.2 Experimental Setup

We LoRA-fine-tune each model on AG News and evaluate every checkpoint on the source-task test set and the five cross-task probes. LoRA uses $r = 16$, $\alpha = 32$ and dropout 0.05 on the attention and MLP projections. We optimise with AdamW at a peak learning rate of 2×10^{-4} , cosine decay, 5 warmup steps and an effective batch size of 256, for 5 epochs. Training uses a stratified 50% subsample of AG News train after holding out 2,000 validation examples; the 7,600-example test set is never used during fine-tuning. All experiments are run on NVIDIA L40S GPUs. Checkpoints are saved at $\{3, 6, 9, 12, 15, 20, 40, 60, 80, 100\}\%$ of total optimizer steps, plus the unadapted base model. At each state we plot accuracy, mean confidence (conditioned on correct vs. incorrect predictions), ECE against training progress, for both the source task and the cross-task suite. We additionally track train, validation, and test NLL on AG News to identify the onset of source-task overfitting. All datasets use fixed zero-shot prompt templates and fixed candidate verbalizers across checkpoints; no prompt parameters are changed during fine-tuning or evaluation. For AG News we use the word verbalizers "World", "Sports", "Business", and "Technology". Exact prompt templates and verbalizer tokens for every benchmark are given in Appendix C.

4.3 Experimental Results

4.3.1 Accuracy and Confidence

Figure 1 shows a clear separation between the evolution of cross-task accuracy and confidence dur-

Task	ΔAcc	ΔConf^+	ΔConf^-	ΔECE
<i>Llama-3.2-1B</i>				
AG News	+0.630	+0.501	+0.416	-0.130
ARC-C	+0.022	+0.336	+0.323	+0.302
ARC-E	-0.020	+0.344	+0.348	+0.292
HellaSwag	+0.006	+0.127	+0.125	+0.120
PIQA	+0.003	+0.150	+0.140	+0.142
BoolQ	-0.009	+0.272	+0.250	+0.159
<i>Llama-3.1-8B</i>				
AG News	+0.256	+0.436	+0.399	-0.176
ARC-C	-0.139	+0.086	+0.147	+0.061
ARC-E	-0.192	+0.028	+0.139	-0.028
HellaSwag	-0.134	+0.210	+0.166	-0.082
PIQA	+0.003	+0.124	+0.094	-0.005
BoolQ	-0.078	+0.146	+0.106	-0.023

Table 1: Mean change from the unadapted base model to the final checkpoint, averaged over three seeds. Cross-task accuracy stays nearly flat (1B) or declines (8B), while both Conf^+ and Conf^- increase on every task.

ing fine-tuning. Llama-3.2-1B and Llama-3.1-8B curves are means over three random seeds. Source-task accuracy on AG News improves rapidly during the early stages of adaptation and then largely saturates. In contrast, accuracy on most untouched tasks changes only modestly across subsequent checkpoints (it declines substantially in case of Llama 3.1 8B), although some task-specific degradation is observed (Appendix E). Importantly, Conf^+ and Conf^- follow similar trajectories across several tasks. Thus, the increase in confidence is not restricted to predictions that the model gets right: fine-tuning also makes the model more confident in many of its errors, despite relatively small corresponding changes in cross-task accuracy. Table 1 summarizes the Llama results; full trajectories for the remaining models are provided in Appendix E.

4.3.2 Confidence trajectory

Confidence exhibits a non-monotonic trajectory over the course of fine-tuning. After an initial transient increase, both Conf^+ and Conf^- remain relatively stable through the middle of training before increasing again toward the final checkpoints. This late-stage rise coincides with the onset of overfitting on the source task. As shown in Figure 3 (Appendix B), training NLL continues to decrease, while validation and test NLL reach a minimum and subsequently increase. Across all six seed runs, off-task confidence increases after the checkpoint minimizing source-task validation NLL, while training NLL continues to decrease. This consistent temporal association suggests that the late-stage

confidence increase accompanies source-task overfitting. We quantify that coincidence by the Pearson correlation between $\Delta_t = \text{NLL}_t^{\text{val}} - \text{NLL}_t^{\text{train}}$ and mean off-task $\text{Conf}^+/\text{Conf}^-$ (five probes). In the late window of finetuning (40–100%) this correlation is $r = +0.94/+0.95$ on Llama-3.2-1B and $r = +0.81/+0.80$ on Llama-3.1-8B. ECE exhibits more heterogeneous behavior across models and tasks, emphasizing that increasing confidence should not be equated with uniformly worsening calibration.

Robustness across seeds. To test whether the observed dynamics are driven by training stochasticity, we repeat fine-tuning of the Llama 1B and 8B models with three random seeds. Metrics are computed independently for each seed and then averaged. The late-stage confidence increase is highly consistent: from 60% to 100% of training, both Conf^+ and Conf^- increase in all six (three runs for 1B and 8B model) and on every cross-task benchmark, with the accuracy degrading slightly.

Full Finetuning We repeat the AG News experiments with full-parameter fine-tuning of Llama-3.2-1B and Llama-3.1-8B, using the same data splits, verbalizers, letter-based probes, and checkpoint percentages as in the LoRA experiments. Because of the higher computational cost, we use one seed per model and a learning rate of 2×10^{-5} . Appendix D shows the resulting trajectories. The 1B model still exhibits a clear late-stage increase in confidence, whereas the effect is less pronounced for the 8B model. For the 8B model, cross-task accuracy declines more under full fine-tuning than under LoRA, suggesting that full-parameter updates can induce stronger cross-task changes. Full fine-tuning was run on an NVIDIA A100 80 GB GPU.

Confidence on a fixed error set. To check that the $\text{Conf}^-/\text{Conf}^+$ rise is not an artefact of items flipping correctness, we freeze two sets: items incorrect at both base and final, and items correct at both (Figure 2). On the stably incorrect set, confidence rises on every probe (Llama-3.2-1B: ARC-Challenge 35% $\rightarrow$ 71%, ARC-Easy 34% $\rightarrow$ 76%, BoolQ 57% $\rightarrow$ 84%; Llama-3.1-8B: ARC-Challenge 53% $\rightarrow$ 69%, HellaSwag 34% $\rightarrow$ 48%, BoolQ 64% $\rightarrow$ 80%). The stably correct set rises as well, except 8B ARC-Easy, which is already high (85% and nearly flat). Confidence is therefore increasing on the same wrong

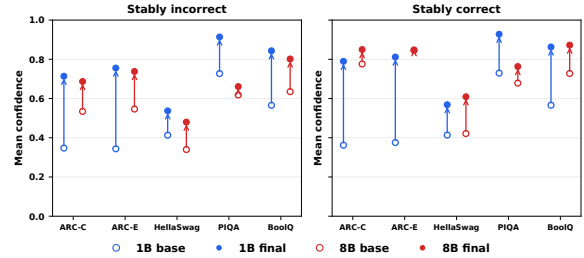


Figure 2: Mean confidence on items that stay incorrect (left) or stay correct (right) from the base model to the final checkpoint. Open markers are base, filled markers are final. Seed 1.

answers and the same right answers, not only because the error set changes.

5 Conclusion

We study how task-specific LoRA fine-tuning changes confidence beyond the source task by tracking intermediate checkpoints across model families and scales. Across our evaluated settings, cross-task accuracy changes comparatively little while confidence increases on both correct and incorrect predictions. Confidence further exhibits a late-stage increase that coincides with source-task overfitting, showing that continued adaptation can alter uncertainty without corresponding gains in cross-task competence. These results motivate evaluating confidence and calibration alongside accuracy when selecting fine-tuned checkpoints. The next steps should test whether these dynamics persist across source tasks and adaptation methods and investigate the causes of confidence growth in cross-tasks. These findings suggest that checkpoint selection should consider confidence and calibration alongside source-task accuracy, rather than defaulting to the final training checkpoint.

Limitations

First, we fine-tune on a single source task, AG News, so it remains unclear whether the observed confidence dynamics generalize to other adaptation tasks. Second, our evaluation is restricted to multiple-choice settings and uses candidate-normalized confidence; the results may therefore not extend directly to open-ended generation. Third, LoRA is our primary adaptation method; we include full-parameter fine-tuning only for Llama-3.2-1B and Llama-3.1-8B, each with one seed. Although the cross-task benchmarks are not used during fine-tuning, we cannot rule out exposure to

them during pretraining. Finally, ECE is sensitive to binning and does not indicate the direction of calibration error, so we interpret it alongside NLL and confidence trajectories. We will work on other source-tasks (other than AG News Dataset) in the near future to further strengthen our observations.

References

- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. [A close look into the calibration of pre-trained language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.
- Zina Chkirbene, Ridha Hamila, Ala Gouisse, and Unal Devrim. 2024. [Large language models \(llm\) in industry: A survey of applications, challenges, and trends](#). In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of EMNLP*, pages 295–302.
- Aneesh Durai. 2025. [Phases of uncertainty: Confidence-calibration dynamics in language model training](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainLP 2025)*, pages 11–16, Suzhou, China. Association for Computational Linguistics.
- Jiahui Geng et al. 2023. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.
- Guande He, Jianfei Chen, and Jun Zhu. 2023. Preserving pre-trained features helps calibrate fine-tuned language models. In *International Conference on Learning Representations*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jerry Huang, Peng Lu, and Qiuhaio Zeng. 2025. Calibrated language models and how to find them with label smoothing. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 25593–25612.
- Jerry Huang, Peng Lu, Qiuhaio Zeng, Yusuke Iwasawa, Yutaka Matsuo, Sarath Chandar, Edison Marrese-Taylor, and Irene Li. 2026. Investigating the multilingual calibration effects of language model instruction tuning. In *Proceedings of EACL*.
- Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of EMNLP*, pages 1326–1340.
- KaShun Shum, Minrui Xu, Jianshu Zhang, Zixin Chen, Shizhe Diao, Hanze Dong, Jipeng Zhang, and Muhammad Omer Raza. 2024. First: Teach a reliable large language model through efficient trustworthy distillation. In *Proceedings of EMNLP*, pages 12646–12659.
- Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, and Jianxin Li. 2025. Towards objective fine-tuning: How llms’ prior knowledge causes potential poor calibration? In *Proceedings of ACL*, pages 14830–14853.
- Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. [On the calibration of large language models and alignment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore. Association for Computational Linguistics.

A Datasets

Table 2 summarizes the splits and answer-choice counts for all datasets used in our experiments. Detailed descriptions and example items for each benchmark are provided below.

Dataset	Split	# Examples	# Choices
AG News	test	7,600	4
ARC-Easy	test	2,376	4
ARC-Challenge	test	1,172	4
HellaSwag	validation	10,042	4
PIQA	validation	1,838	2
BoolQ	validation	3,270	2

Table 2: Datasets used for evaluation, along with the split, number of examples, and number of answer choices per item.

A.1 AG News

AG News is a four-way topic classification dataset of news articles labeled as *World*, *Sports*, *Business*, or *Sci/Tech*. We use it as our in-task fine-tuning and evaluation benchmark, prompting the model to predict a single verbalizer token for the article topic.

A.2 HellaSwag

HellaSwag (Zellers et al., 2019) is a commonsense sentence-completion benchmark: given a short context, the model chooses the most plausible continuation among four endings. We use the validation split as a cross-task multiple-choice probe.

Context	Correct Ending
A man is sitting on a roof. he...	starts pulling up roofing on a roof.
A lady... grabs the [barbell]. the lady...	stands and lifts the weight over her head.

Table 3: Example items from HellaSwag.

A.3 ARC-Challenge

ARC-Challenge (Clark et al., 2018) consists of grade-school science questions that require multi-hop reasoning rather than surface retrieval. We evaluate on the test split as a cross-task science QA probe.

A.4 BoolQ

BoolQ (Clark et al., 2019) is a yes/no reading-comprehension task: given a Wikipedia passage

Question	Answer
If a planet rotates faster after a meteorite impact, what happens?	Planetary days will become shorter.
Why do engineers test building models under simulated earthquakes?	So that buildings will be made safer.

Table 4: Example items from ARC-Challenge.

and a naturally occurring question, the model answers *Yes* or *No*. We use the validation split as a cross-task QA probe.

Question	Answer
Does ethanol take more energy to make than it produces?	No
Is house tax and property tax the same?	Yes

Table 5: Example items from BoolQ.

A.5 PIQA

PIQA (Bisk et al., 2020) tests physical commonsense: given a goal, the model chooses the more sensible of two solutions. We use the validation split as a cross-task multiple-choice probe.

Goal	Preferred Solution
Ready a guinea pig cage	Use bedding of ripped paper strips (with water and food), not ripped jeans material.
Finish a dresser	Finish with woodgrain using a bobby pin, not replace drawer with a bobby pin.

Table 6: Example items from PIQA.

A.6 ARC-Easy

ARC-Easy (Clark et al., 2018) contains grade-school science questions that are typically answerable through direct retrieval or single-step reasoning, in contrast to ARC-Challenge. We evaluate on the test split as a complementary cross-task science QA probe.

B Negative Log Likelihood

Figure 3 shows source-task NLL on the AG News training, validation and test splits across fine-tuning checkpoints. Training NLL decreases throughout, whereas validation and test NLL reach a minimum in the middle of training and rise thereafter, tracking each other closely. We take the point at which

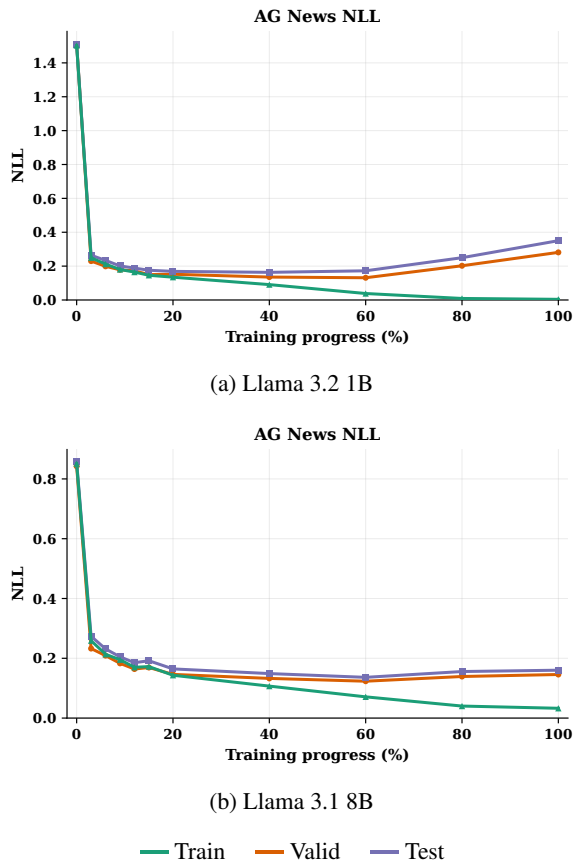


Figure 3: AG News NLL trajectory during fine-tuning.

the train–validation gap begins to widen as the onset of source-task overfitting, which is the regime in which the late-stage cross-task confidence increase of Section 4.3.2 occurs. Notably, source-task test NLL degrades while source-task accuracy is already saturated: the in-task counterpart of the accuracy–confidence separation we observe on the cross-task benchmarks.

C Prompt Templates and Verbalizers

All tasks are evaluated by next-token prediction after the Answer: cue. For candidate c_j , the score z_j is the full-vocabulary log-probability assigned to its single-token verbalizer $v(c_j)$. We then renormalize these scores over the K candidate verbalizers, as defined in Eq. 2.

AG News uses word verbalizers, whereas all off-task probes use letter verbalizers. For each model tokenizer, we verify that every verbalizer is a single distinct token and that the AG News verbalizer ids are disjoint from the probe letter ids. Therefore, fine-tuning on AG News never directly supervises the answer tokens used by the probes. Leading whitespace is part of each scored verbalizer (e.g., "

A" rather than "A").

AG News. The original *Sci/Tech* class is represented by Technology, because " Sci/Tech" is not a single token for all evaluated tokenizers. The displayed labels omit the leading spaces that are present in the scored verbalizers.

Classify the topic of the following news article.

Article: {article}

Choose exactly one answer:

World
Sports
Business
Technology

Answer:

ARC-Easy and ARC-Challenge. ARC examples contain either four or five choices. For a K -choice item, the prompt lists choices A through the K th letter, and scoring is restricted to the corresponding K letter verbalizers. Gold answers are resolved by their position in the item’s choice list, because some ARC examples use numeric rather than alphabetic answer keys.

Answer the following multiple-choice science question.

Question: {question}

A. {choice_1}
B. {choice_2}
C. {choice_3}
D. {choice_4}
[E. {choice_5}]

Choose exactly one answer: A, B, C, or D.

Answer:

For five-choice examples, the final instruction instead reads: Choose exactly one answer: A, B, C, D, or E.

HellaSwag.

Complete the following scenario with the most plausible continuation.

Scenario: {context}

A. {ending_1}
B. {ending_2}
C. {ending_3}
D. {ending_4}

Choose exactly one answer: A, B, C, or D.

Answer:

PIQA.

Choose the more sensible solution for the following goal.

Goal: {goal}

A. {solution_1}
 B. {solution_2}
 Choose exactly one answer: A, or B.
 Answer:

BoolQ. The prompt displays the semantic labels *Yes* and *No*, but the corresponding scored verbalizers are " A" and " B".

Answer the yes/no question using the passage.
 Passage: {passage}
 Question: {question}
 A. Yes
 B. No
 Choose exactly one answer: A, or B.
 Answer:

D Full Fine-Tuning

We repeat the AG News recipe with full-parameter updates of Llama-3.2-1B and Llama-3.1-8B (same splits, word verbalizers, letter probes, and %-checkpoints; one seed). Figure 4 uses the same four panels as Figure 1. Cross-task drift is therefore not a LoRA artifact only and updating every weight hurts untouched-task accuracy more.

E Additional Trajectory Results

Figures 5–9 show trajectory metrics (Accuracy, Conf^+ , Conf^- , and ECE) across fine-tuning checkpoints for the five additional models: Qwen3 1.7B, Qwen3 4B, Qwen3 8B, SmolLM2 1.7B, and SmolLM2 360M. Across all models, the same qualitative pattern observed for Llama holds: cross-task accuracy remains flat or deteriorates while confidence rises on both correct and incorrect predictions.

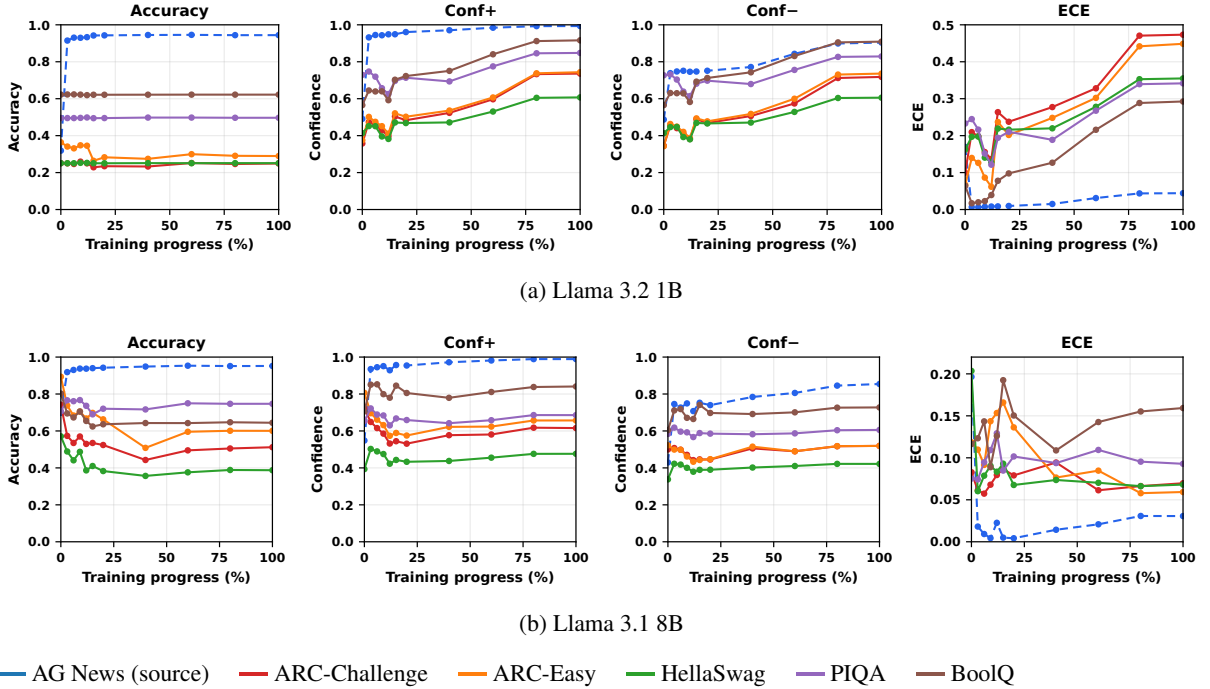


Figure 4: Trajectory metrics of full fine-tuning at checkpoints.

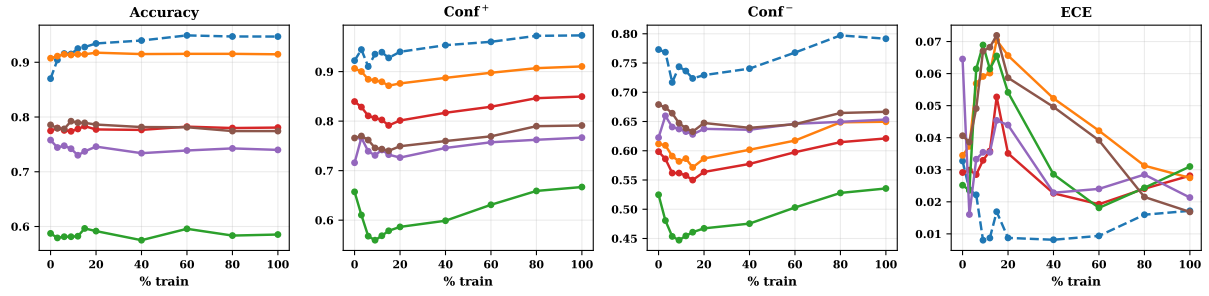


Figure 5: Trajectory metrics for Qwen3 1.7B.

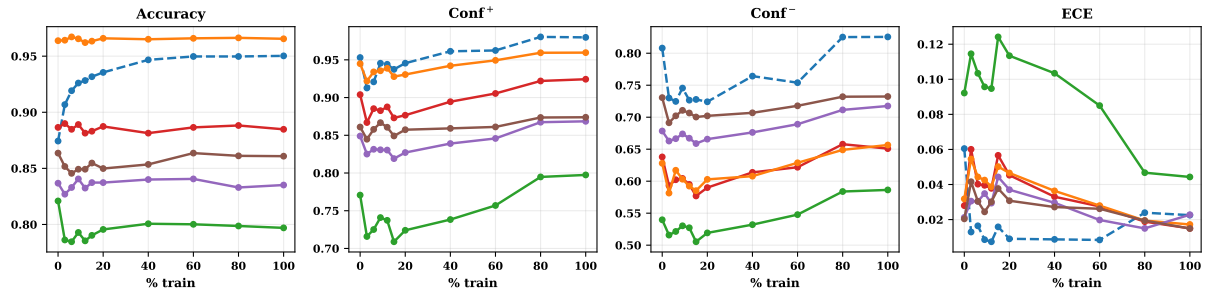


Figure 6: Trajectory metrics for Qwen3 4B.

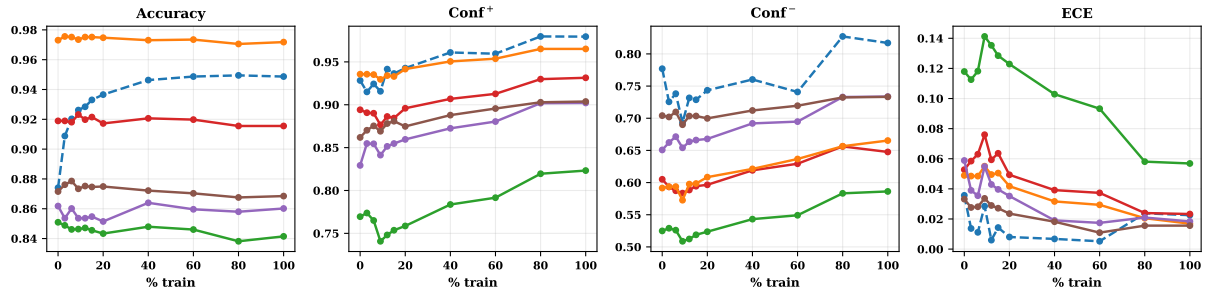


Figure 7: Trajectory metrics for Qwen3 8B.

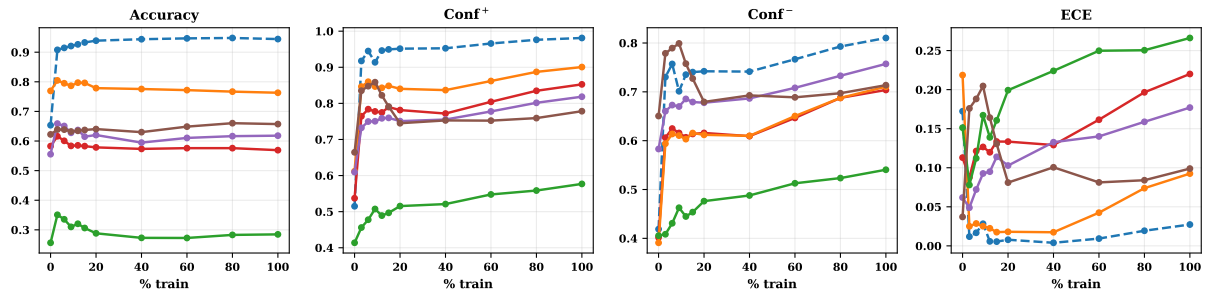


Figure 8: Trajectory metrics for SmolLM2 1.7B.

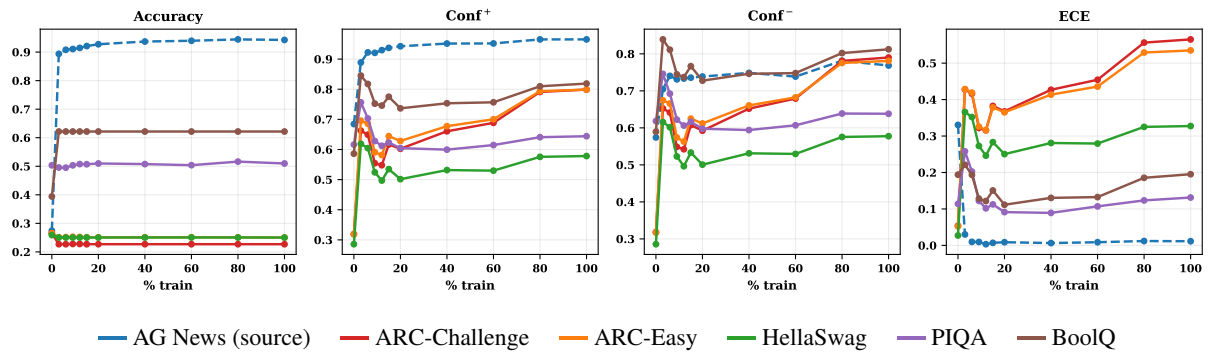


Figure 9: Trajectory metrics for SmolLM2 360M. Color legend (shared across all figures in this appendix): dashed line indicates the source task (AG News, dashed); solid lines indicate cross-task benchmarks.