# Phases of Uncertainty: Confidence–Calibration Dynamics in Language Model Training

**Aneesh Durai**
UC Berkeley
aneesh.durai@berkeley.edu

## Abstract

Autoregressive language models achieve strong performance across a wide range of natural language processing (NLP) tasks, yet their uncertainty estimates remain poorly understood, particularly during training. Prior work has primarily evaluated calibration and out-of-distribution (OOD) robustness at the final checkpoint, overlooking the dynamics that unfold earlier. We introduce a phase-based framework for tracking uncertainty metrics—including expected calibration error (ECE) and Kullback–Leibler (KL) divergence—across distinct stages of training. Using GPT-2 models trained across multiple random seeds, we find that uncertainty dynamics follow a consistent set of phases: models begin conservative and relatively well calibrated, but later phases introduce a paradoxical decoupling where confidence increases even as calibration worsens, especially under distribution shift. This paradox implies that the final checkpoint is not always the most reliable for deployment and motivates phase-aware strategies such as dynamic checkpoint selection or targeted calibration. Our findings highlight that uncertainty should be understood as a training-dependent property rather than a static one, opening new directions for scaling this framework to larger models, tasks, and distribution shift scenarios.

## 1 Introduction

Autoregressive language models have become central to a large portion of modern NLP, driving progress in tasks as varied as document summarization, dialogue, and code generation (Brown et al., 2020). Yet, the impressive in-distribution performance of these models hides a recurring issue: their behavior is far less predictable when the input departs from the training distribution (Hendrycks and Gimpel, 2017). In production settings, such out-of-distribution (OOD) cases are inevitable such as topic drift in conversational systems, domain mismatch in translation, or simply user queries that exploit corner cases in the model's learned representation.

Uncertainty estimation has become a way to address this problem. Approaches such as Bayesian approximations via dropout (Gal and Ghahramani, 2016) or calibration-based adjustments (Guo et al., 2017) offer ways to associate model predictions with confidence scores. However, most of this work evaluates a final trained output. What is less well understood, especially in language modeling, is how uncertainty evolves during training itself. Language models acquire the syntax, semantic, and task-specific reasoning in a staged manner, and their calibration profile is unlikely to be uniform across these stages (Desai and Durrett, 2020; Jiang et al., 2021).

Our key finding is that calibration does not improve monotonically with training: a mid-training phase emerges in which models grow more confident while becoming less calibrated.

In this work, we introduce a phase-based framework for tracking and analyzing the joint dynamics of calibration error and KL divergence between successive stages of training. By segmenting model training into distinct phases and evaluating these metrics both in-distribution and OOD, our approach offers a structured view of how and when models become more or less calibrated, and how their predictive distributions shift over time.

## 2 Related Work

### 2.1 Uncertainty Estimation in NLP

Quantifying predictive uncertainty has been a needed measure in modeling and modern neural networks. For classification tasks, baseline confidence scores such as the maximum softmax probability and predictive entropy are widely used to flag low-confidence predictions (Hendrycks and Gimpel, 2017). Bayesian-inspired techniques, including Monte Carlo dropout (Gal and Ghahramani,



*Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 11–16

2016) and deep ensembles (Lakshminarayanan et al., 2017), have adapted to NLP models to better capture the epistemic and aleatoric uncertainty. Recent work has explored these methods for both and structured prediction tasks like semantic parsing (Dong et al., 2017). However, most existing approaches report uncertainty only for the final converged models, and that overlooks how these measures are evolving during training.

### 2.2 Calibration of Language Models

Calibration measures the degree to which predicted probabilities align with empirical correctness (Guo et al., 2017). While overconfidence is a well-known issue in fields like computer vision, language models exhibit domain-specific calibration challenges (Desai and Durrett, 2020). Post-hoc techniques such as temperature scaling and histogram binning have been applied to NLP (Guo et al., 2017), but once again, their effectiveness is often evaluated only after full training.

Some other work has explored calibration in generative settings, (Kumar et al., 2019), yet there remains little understanding of how calibration quality changes mid-training, especially for large-scale autoregressive models.

### 2.3 OOD Robustness and Distribution Shifts

OOD detection aims to identify inputs that differ substantially from the training distribution. Density-based methods (Lee et al., 2018), and uncertainty-based rejection strategies (Hendrycks and Gimpel, 2017) have been explored in NLP, often under domain shift scenarios (Varshney et al., 2022). Despite this, the majority of studies evaluate robustness at convergence, providing little insight into the temporal dynamics of OOD behavior. The opportunity of the interplay between the training-phase uncertainty trends, calibration shifts, and OOD performance remains largely unexplored.

We address this gap by systematically tracking the uncertainty metrics, calibration scores, and KL divergence between training phases for autoregressive language models. By linking these evolving quantities to in-distribution and OOD generalization, we provide a temporal perspective on uncertainty and robustness, which offers a richer understanding than simple post-training evaluation alone.

## 3 Methodology

### 3.1 A Phase-Based View of Training Dynamics

Calibration in neural networks is typically assessed only at convergence, which obscures transient regimes where confidence and reliability can drift in opposite directions (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). Leveraging this observation, we take a temporal perspective and segment training into phases defined by persistent shifts in uncertainty and calibration traces.

### 3.2 Metrics

We track two uncertainty-related metrics at regular checkpoints.

**KL Divergence to Uniform (Confidence Proxy).** Let $p \in \Delta^{V-1}$ be the next-token predictive distribution over a vocabulary of size $V$, and let $u$ denote the uniform distribution ($u_i = 1/V$). Confidence is measured as

$$D_{\mathrm{KL}}(p \,\|\, u) \;=\; \sum_{i=1}^{V} p_i \log \frac{p_i}{1/V}. \qquad (1)$$

Higher values indicate sharper, more confident distributions; lower values indicate more diffuse predictions. This quantity is 0 when predictions are maximally uncertain (uniform) and increases as the distribution sharpens, making it a natural confidence proxy. It is closely related to predictive entropy, since $D_{\mathrm{KL}}(p \,\|\, u) = \log V - H(p)$. While other reference distributions could be considered, we adopt the uniform baseline because it provides a simple and interpretable notion of random guessing, against which sharper, more confident predictions can be measured.

**Expected Calibration Error (ECE).** Following Guo et al. (2017), tokens are binned by predicted confidence into $M$ equal-width bins $\{B_m\}_{m=1}^{M}$. Let $\mathrm{acc}(B_m)$ be the empirical accuracy and $\mathrm{conf}(B_m)$ the mean confidence in bin $m$. The ECE is

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{\sum_j |B_j|} \Big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\Big|. \qquad (2)$$

Lower values indicate better calibration.

### 3.3 Phase Detection

At each checkpoint we record KL-to-uniform and ECE over the ID validation set and an OOD corpus;

we smooth each per-seed KL trajectory with an exponentially weighted moving average and detect changepoints on KL. Let $t$ index checkpoints.

We then identify three regimes per seed:

1. **Phase I (Early Learning):** ends at the early local maximum of KL (searched in the first half of training) or a default tertile boundary if no clear maximum exists.
2. **Phase II (Confidence Surge):** begins after Phase I and ends at the subsequent local maximum of KL (or a default second-tertile boundary), enforcing a minimum phase length.
3. **Phase III (Stabilization):** the remaining steps to the final checkpoint.

Boundaries are constrained to respect minimum durations and ordered consistency (therefore I $<$ II $<$ III). We compute all metrics per phase and then report both per-seed summaries and seed-averaged statistics. This procedure captures non-monotonic behavior that endpoint-only evaluation can miss, such as periods where confidence rises while calibration degrades (Ovadia et al., 2019; Minderer et al., 2021).

# 4 Experiments and Results

## 4.1 Experimental Setup

We trained GPT-2 models for 3,000 optimization steps across five seeds. Training was conducted on the WikiText-2 (Merity et al., 2016) corpus for in-distribution (ID) evaluation, while out-of-distribution (OOD) generalization was assessed on the AG News (Zhang et al., 2016) dataset. At regular intervals, we computed both standard training metrics (loss) and uncertainty metrics for both ID and OOD test sets. This setup provides a comprehensive view of the interaction between confidence and calibration throughout training. Unless stated otherwise, significance is assessed with a two-sided paired t-test over checkpoints, aggregated across seeds.

## 4.2 Phase Detection Procedure

To identify interpretable regimes of uncertainty dynamics, we employed an automatic phase segmentation method based on changepoints in the KL trajectory. Consistently across all seeds, three phases emerged as shown in Table 1.

Table 1: Training phase boundaries identified across all seeds.

| Phase | Step Range |
|---|---|
| I (Early Transient) | 50–1500 |
| II (Confidence–Calibration Drift) | 1550–2900 |
| III (Convergence Plateau) | 2950–3000 |

Phase characteristics across seeds are summarized in Table 2. Figure 1 shows the dynamics for the average of the seeds. Phase I balances confidence and calibration, Phase II marks systematic divergence between them (the confidence–calibration paradox), and Phase III represents a plateau at degraded calibration levels.

## 4.3 In-Distribution Dynamics

During Phase I, models maintained relatively low calibration error (mean ECE $\approx 0.005$). As training progressed into Phase II, a paradoxical trend emerged: calibration degraded even as confidence increased. Specifically, mean ECE rose by $\sim$23.4% (from 0.0049 to 0.0058, $p = 2.05 \times 10^{-5}$), while KL divergence to uniform predictions increased by 0.5% (9.471 $\rightarrow$ 9.523). This indicates that the models became more confident but less calibrated. In Phase III, metrics stabilized (KL $\approx 9.529$ and ECE $\approx 0.0057$), but calibration did not return to the initial level.

Across all five experiments, this paradox held consistently: in every run, confidence increased while calibration worsened. Prior speculation that calibration might improve in later stages (e.g., Guo et al., 2017; Desai and Durrett, 2020) was not supported in our setting.

## 4.4 Out-of-Distribution Behavior

When evaluated on AG News, models exhibited the same paradox but with larger miscalibration. OOD ECE rose from $\sim$0.033 in Phase I to $\sim$0.040 in Phase II ($p = 2.31 \times 10^{-8}$), representing a $\sim$21% relative increase, alongside a concurrent increase in KL-to-uniform. As with ID, metrics stabilized in Phase III without recovery.

Notably, the paradox was amplified OOD: the models simultaneously became more confident and less calibrated under distribution shift, producing error rates far larger in magnitude than ID. This indicates that the confidence–calibration paradox is not only a training artifact but also a deployment concern for real-world distribution shifts.

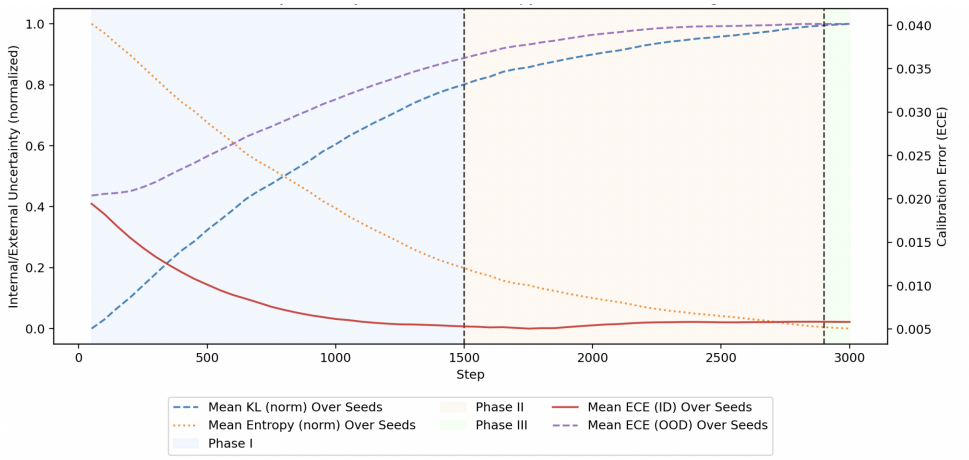


Figure 1: Phase dynamics of calibration and confidence. Confidence (KL, blue) rises steadily, while calibration error (ECE, red/purple) degrades. Phase II (yellow) highlights the paradoxical "danger zone" where all 5 seeds showed confident but unreliable predictions. Note the divergence between rising confidence and worsening calibration around step 1500.

Table 2: Phase characteristics averaged across seeds. ECE reported for in-distribution (ID) and out-of-distribution (OOD). KL-to-uniform is our primary confidence metric; entropy (H) is reported for reference only.

| Phase | KL | H | $\text{ECE}_{\text{ID}}$ | $\text{ECE}_{\text{OOD}}$ |
|---|---|---|---|---|
| I | $9.471 \pm 0.004$ | $1.354 \pm 0.004$ | 0.005 | $0.033 \pm 0.001$ |
| II | $9.523 \pm 0.003$ | $1.302 \pm 0.003$ | 0.006 | $0.040 \pm 0.001$ |
| III | $9.529 \pm 0.002$ | $1.296 \pm 0.002$ | 0.006 | $0.040 \pm 0.001$ |

## 5 Discussion and Future Work

We show three consistent training phases, document a mid-training confidence-calibration gap, and outline how to use these signals for safer checkpoint selection and calibration. Our results suggest that current practice may systematically deploy models from their least reliable phase. Monitoring only validation loss obscures the fact that Phase II coincides with worsening calibration. This paradox has several practical consequences: (1) calibration should be tracked jointly with loss during training, (2) deployments should avoid Phase II checkpoints (high confidence, poor calibration), and (3) interventions such as temperature scaling or selective regularization may be most beneficial when targeted specifically to this unstable phase. Without such precautions, models risk being deployed precisely when they are most deceptively unreliable.

Beyond these immediate implications, our phase-based framework highlights opportunities for future work. Scaling to larger architectures and reasoning-capable models will test the generality of the paradox. Expanding to broader OOD scenarios (e.g., multilingual or reasoning tasks) will help determine whether the observed dynamics extend beyond WikiText and AG News. Finally, phase-aware interventions could be designed to adaptively correct calibration drift in real time, reducing deployment risks for large-scale language models.

## 6 Conclusion

We introduced a phase-based framework for analyzing uncertainty and calibration dynamics throughout language model training. Across multiple seeds, we consistently observed a confidence–calibration paradox: models became less reliable precisely as their predictions grew more confident. This paradox was amplified under distribution shift, underscoring its practical importance for deployment safety.

By framing uncertainty as a training-dependent property rather than a static one, we provide a foundation for phase-aware monitoring, checkpointing, and intervention strategies. In practice, our results motivate monitoring calibration (ECE) jointly with validation loss, avoiding Phase II checkpoints when selecting release models, and applying simple post-hoc calibration such as temperature scaling at deployment.

## Limitations

First, we trained GPT-2 scale models for 3,000 iterations across five seeds, a modest but controlled scope. Second, our OOD evaluation was limited to a single dataset (AG News) and a restricted set of uncertainty metrics (ECE and KL). Third, our phase detection relies on inflection points in these metrics; whether analogous phase boundaries generalize to larger architectures or alternative metrics remains open.

Despite these constraints, the reproducibility of phase dynamics across seeds suggests that the phenomena are not small-scale artifacts but emergent properties of autoregressive training. Extending this analysis to larger models, broader OOD scenarios, and alternative calibration interventions represents a natural next step.

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *ArXiv*, abs/1706.04599.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3792–3803.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Preprint*, arXiv:1906.02530.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. *Preprint*, arXiv:2203.00211.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. *Preprint*, arXiv:1509.01626.

## Response to Reviewer Comments

### Reviewer RqHb

1. *Missing double-descent literature (Nakkiran et al.; OpenAI blog; Schaeffer et al.).* Our work studies calibration dynamics across training phases (ECE and KL-to-uniform), not necessarily accuracy/test-error curves. While double descent concerns error as a function of model size/data/epochs, our main result is a *confidence–calibration gap* during mid-training even as loss improves. We agree the phenomena are related but orthogonal; due to short-paper space we keep Related Work focused on calibration/uncertainty and OOD. We are happy to expand this connection in a longer version.

2. *Unclear how to utilize the research.* The camera-ready adds concrete guidance in *Discussion and Future Work* and *Conclusion*: (i) monitor calibration (ECE) jointly with validation loss; (ii) avoid Phase II checkpoints when choosing release models; (iii) apply simple post-hoc calibration (temperature scaling) and consider selective regularization targeted to Phase II.

3. *Missing dataset references.* We now cite WikiText-2 and AG News explicitly in *Experimental Setup*.

4. *Numbers not reflected in Figure 1.* We clarified phase boundaries (Table 1) and added a per-phase summary (Table 2). Figure 1 caption explicitly describes the divergence point (around step 1500) referenced in §4.3–4.4.

### Reviewer T3uL

1. *Entropy vs. KL to uniform are redundant.* We agree. The revision treats KL-to-uniform as the *primary* confidence proxy and includes entropy only as a reference column in Table 2. We explicitly note the identity $D_{\mathrm{KL}}(p\|u) = \log V - H(p)$ in *Metrics*.

### Reviewer 1agi

1. *Generality to newer/reasoning models.* We scope the empirical study to GPT-2 scale for 3,000 steps across five seeds on WikiText-2 (ID) and AG News (OOD). We state this limitation and the generalization agenda in *Limitations* and *Discussion and Future Work* (scaling to larger and reasoning-capable models).

2. *More seeds/tasks; statistical support.* We report five seeds and add a sentence in *Experimental Setup* specifying two-sided paired $t$-tests over checkpoints (aggregated across seeds) for reported comparisons. We agree that the next step would require multiple more runs.

3. *Intervening in the "danger zone."* We strengthened the deployment guidance in *Discussion*: phase-aware checkpoint selection, targeted regularization during Phase II, and post-hoc calibration (temperature scaling) at deployment.