FMDB *Transactions on Sustainable Finance and Data Science*

Self-Calibrating Foundation Models with Uncertainty-Driven Feedback Loops for High-Stakes Decisions

Vimalkumar Kumaresan[1,*]

[1]Department of Data Science, Tredence Inc., San Jose, California, United States of America.
vimalkumar.k@tredence.com[1]

Abstract: This paper investigates self-calibrating foundation models in the high-stakes decision-making setting, where incorrect decisions are prohibitively expensive. At the heart of it is the aim to reduce hallucinations and spurious confidence that researchers sometimes witness in large language models, especially in safety-critical applications such as healthcare and finance. The model incorporates an uncertainty-aware feed-forward loop to reflexively and flexibly adjust its confidence thresholds in situ based on entropy-sensory signals. The researcher evaluated the proposed system on a carefully prepared dataset of 459 fully anonymized high-stakes cases from two domains: financial credit risk assessment and medical triage. The experiment was implemented using the PyTorch and Hugging Face Transformer libraries, with the Llama-2 architecture as the base model. For uncertainty estimation, the researcher used software-based methods, including Monte Carlo dropout and Deep Ensembles. The results show that adding both terms improves calibration quality rather than overall performance. Uncertain outputs can also be flagged for inspection by human experts, thus providing a hybrid intelligence approach. In this paper, researchers present the architectural modifications required to ensure self-calibration, along with a rigorous definition of its advantages over non-calibrated references, which justify its potential applicability in demanding real-world applications.

Keywords: Uncertainty Quantification; Foundation Models; Feedback Loops; Risk Assessment; Deep Ensembles; Large Language Models; Hybrid Intelligence Approach; Calibration Quality.

Received on: 12/05/2025, **Revised on:** 03/07/2025, **Accepted on:** 24/09/2025, **Published on:** 07/03/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSFDS

DOI: https://doi.org/10.69888/FTSFDS.2026.000620

Cite as: V. Kumaresan, "Self-Calibrating Foundation Models with Uncertainty Driven Feedback Loops for High Stakes Decisions," *FMDB Transactions on Sustainable Finance and Data Science*, vol. 1, no. 1, pp. 1–9, 2026.

1. Introduction

The rapid emergence of foundation models has transformed the AI landscape, empowering systems with reasoning capabilities and enabling them to produce human-level text at unprecedented scale [6]. But as these models are used in ever weightier areas of life, for example, health care, legal decision-making, and financial risk assessment, an important problem is becoming clear: the calibration of model confidence. That issue, observed in deployment risk assessments, occurs when models exhibit a disconnect between their confidence scores and predictive accuracy, as evidenced by the benchmarking analyses of this mismatch [2]. In these cases, the models often yield invalid results with overconfidence (hallucination in a systematic way) and false knowledge [14]. When high stakes are involved, it is time to respect its demands not only as a performance challenge but

[*]Corresponding author.

also as a risk issue, with real risks at play, such as financial ruin or misdiagnosis (see domain-specific research on the topic [5]. This has emphasized the necessity of AI systems that are aware of their own uncertainty and act on it, as also stated in theoretical work on trustworthy AI [9].

The paper presents a model structure with self-calibration and an uncertainty-aware feedback loop, based on the paradigm of recent work in adaptive AI-based systems [6]. Researchers do not demand to obtain a membership assignment based on learned statistics obtained by analysis of subject-specific data at runtime, in contrast to folk making the criteria for being a neuron;-) However, unlike typical fixed-learning static models, researchers do have self-reflection properties that change within our simulations, as proposed elsewhere [3]. When high entropy is detected, information about low confidence is provided to the model, as in entity-sensitive decision systems. It serves two purposes: (1) discouraging confidence for overconfident predictions and (2) directing ambiguous output for either further scrutiny or context enhancement consistent with risk–sensitive decision routing [1]. Researchers here do not focus on prediction accuracy, but rather on reliability and trust, which are fundamental for AI systems working on sensitive tasks, as in this other approach oriented towards reliability [8]. Our method employs a base model, and the UQ module is added on top, combining features in an uncertainty-aware fashion: Optimal, researchers may apply early stop so that the prediction matches an optimal structure. This layer, in turn, goes to the hidden states and queries over them, as well as to coarse output logit values, before committing the decision. Unlike RLHF, which operates on tone and alignment, our feedback system provides a quantitative, risk-based signal. It penalises confident errors and promotes honest doubt, which aligns with calibration-aware training objectives [13].

This leads to probabilistic faithfulness — the model should be correct 90% of the time when it predicts a probability of 0.9, following foundational work on probabilistic calibration [15]. The significance of this architecture lies in its ability to be used in real-world, deployment-oriented AI safety work. Existing calibration techniques, such as temperature scaling, are post-hoc calibrated, that is, they require a separate step after training the model, as noted by calibration benchmarks [2]. Our treatment is not parametric, which allows it to work online by calibrating the model during inference and adapting in real time to changing distributions, a property demonstrated through experiments on non-stationary domains [9]. For example, a traditional model could be prejudiced towards old norms while finding it strange and unknown fraud cases; instead, our self-calibration system characterizes anomalous cases as high-uncertainty, motivated by prior financial AI evaluation reports [5]. Researchers also analyze the embedding of this model into operational pipelines, e.g., prior work on enterprise-level AI workflows [10]. This, in turn, is a move from opaque, self-assured systems to transparent, can-do helpers—and aligns with human-centred AI design writing [1]. As stated in the literature on trust engineering research, trust in AI is built not by perfect performance, but by mechanisms that ensure safety and accountability. Research empirically confirms this by cross-validating our system across 459 diverse, high-stakes cases, providing the strongest real-world validation to date that self-calibration is not merely theoretical but achievable. The rest of this paper describes the architecture, discusses related work, and presents extensive experimental analysis in accordance with well-established methodological guidelines.

2. Review of Literature

The quest for dependable AI has been shaped by the pressure of increasingly complex neural networks, and its history is documented in early analytical surveys of AI dependability. Early research focused on performance evaluation , with accuracy as the sole criterion for success, as documented in the classic performance literature. Performance was quantified based on a model's accuracy in classifying images or generating text, using metrics originally articulated in early benchmark-driven work [11]. But as AI systems moved out of the lab and into production, the downside of this focused attention became clear. The calibration problem was a serious issue, particularly for the analysis of deployment readiness. Researchers observed that DNNs tend to become overconfident, and this phenomenon is quantified through statistical testing of prediction confidence [14]. A related literature found a paradox: deeper models were not only more accurate but also less well calibrated—a trade-off studied in works on deep networks and confidence. This paradox inspired the study of the optimisation dynamics of overconfidence in deep learning models [3]. One of the main streams of post-calibration calibration research is based on recommendations from calibration technique surveys. Among them, temperature scaling attracted significant interest. It calibrates logits for the softmax layer so that the calibrated predicted probabilities better align with observed outcomes, as supported by probabilistic calibration experiments.

Although successful in classification tasks, temperature scaling does not work as well for large generative models, especially for capturing the deeply nuanced uncertainty inherent in language generation [12]. The loose nature of language involves more than one scalar shift , as pointed out in semantic uncertainty studies [7]. This sparked a broader shift towards intrinsic UQ methods, as reflected in methodological reviews of uncertainty modeling. Bayesian neural networks were a theoretically sound solution; weights were represented as distributions rather than single values—an idea explored in some probabilistic deep learning frameworks [4]. However, they were too computationally intensive to use with the large-scale foundation models that researchers sought to retain in this study, as shown by systems-level scalability tests [6]. To mitigate this dilemma, approximation techniques are devised, as widely reported in the literature on scalable uncertainty estimation. One of such

methods is the Monte Carlo Dropout, which introduces the reinterpretation of dropout as a method of approximation under the branches for Bayesian classes. This idea has been empirically studied in the context of stochastic inference [11].

By enabling dropout at inference time and sampling multiple outputs, researchers can measure prediction variance—a technique shown to work well for empirical uncertainty. Indeed, high variance is highly correlated with high uncertainty, as explored in confidence modelling work. Monte Carlo Dropout sits at a comfortable midpoint between theoretical soundness and computational tractability, which accounts for its broad adoption, as seen in comparative uncertainty surveys. Another effective method is deep ensembles, which train multiple models and combine their outputs; this technique has been demonstrated to enhance both robustness and calibration in previous experiments. Nevertheless, training many large models remains resource-intensive, as per a cost–benefit analysis [12]. Recent efforts have shifted towards model alignment and feedback loops, as seen in interdisciplinary AI-safety reviews. The success of RLHF is motivating the application of such feedback mechanisms for calibration, in the spirit of optimization work [8]. Instead of maximizing preference satisfaction, this work focuses on encouraging epistemic honesty—an objective pursued in research on veracious AI [13]. These are: The attenuation of (potentially) hallucinatory—confident but false—outputs, which have been increasingly investigated in empirical studies on mitigated hallucinations. In foundational uncertainty theory, a sharp separation is made between aleatoric (data-based noise) and epistemic uncertainty (model ignorance), and this classification scheme is formalized in the context of basic probability assignments p varying in the space of measures on Ω.

Solving high-stakes application areas requires systems that can differentiate and act on both, as called for by risk-sensitive AI design research. Other emerging areas of interest concern the interplay between humans and AI in high-risk scenarios, as addressed in interdisciplinary reviews in human–computer interaction [4]. Recent research has shown that human users are highly sensitive to AI confidence signals –a phenomenon confirmed in behavioural studies of automation bias [11]. Overconfident AI systems can discourage users from exercising their judgment. At the same time, conservative models may result in a trust rejection—in between this dynamic lies the tension between truster and trustee that the trust calibration literature has studied and documented [14]. An optimally designed AI system should also express its uncertainty and doubt to the human operator, giving control back to the operator only when needed - a principle encapsulated in cooperative AI frameworks such as Doshi-Velez and Kim [5]. This is also the research direction researchers pursue, and they achieve it by embedding dynamic feedback control within the model architecture itself, enabling self-regulation without constant human monitoring. This approach is analogous to that proposed in studies on the design of autonomous assistant systems. Although inferential challenges such as calibration, uncertainty quantification, and feedback have been well studied in isolation, there is a dearth of systematic frameworks that connect these pieces for ground-truth models in high-stakes environments—a significant gap identified by recent synthesis reviews15.

3. Methodology

To build such a self-calibrating foundation model, the researcher designed a stringent experimental framework that integrates uncertainty quantification directly into the inference pipeline. Our method is based on the fine-tuning of an already pre-trained Llama-2 model with internal feedback. The researcher used the PyTorch framework and the Hugging Face Transformers library to set up an experimental environment. The method was organized into four steps: data preparation, model adaptation, feedback loop embedding, and evaluation. Figure 1 shows the proposed self-calibrating architecture using uncertainty-driven feedback control, along with its augmented version. The uncertainty-dependent feedback loop is defined as: estate realization. For data preparation, the researcher collected a carefully curated dataset of high-stakes cases (459). The pre-processing stage involves each sample to ensure accurate ground-truth labels are available, yielding an exact determination of both a model's predictive accuracy and its confidence estimates. The researcher used stratified sampling to ensure that each difficulty category (i.e., low, medium, and high) was represented in a balanced way. During model adaptation, full fine-tuning was intentionally ruled out due to computational overhead and the possibility of catastrophic forgetting. Instead, Low-Rank Adaptation (LoRA) was used to inject a few trainable parameters into the attention layers to regulate specific behaviour while keeping the original model weights intact.

The unique contribution of this approach is its uncertainty-based feedback. Inherent variance in output across these passes was used to calculate an entropy-based measure of epistemic uncertainty. If the entropy level reached a pre-specified threshold, the system initiated a second-order processing step in which it recollected the input context, thereby simulating second-order reasoning. Concurrently, the model dynamically adjusted the SoftMax temperature, with higher uncertainty leading to flatter probability distributions to prevent overconfident predictions. Accuracy was not the sole contributor, as the evaluation focused on reliability. For probabilistic calibration, the researcher used the Expected Calibration Error (ECE) and the Brier Score, along with a rejection rate metric that reflects how often the model properly deferred to uncertainty. All experiments were performed on a high-performance computing cluster with NVIDIA A100 Tensor Core GPUs, given the high computational cost of multiple inferences. Combined, this approach defines a complete end-to-end, reproducible pipeline for developing well-calibrated, risk-sensitive AI systems.

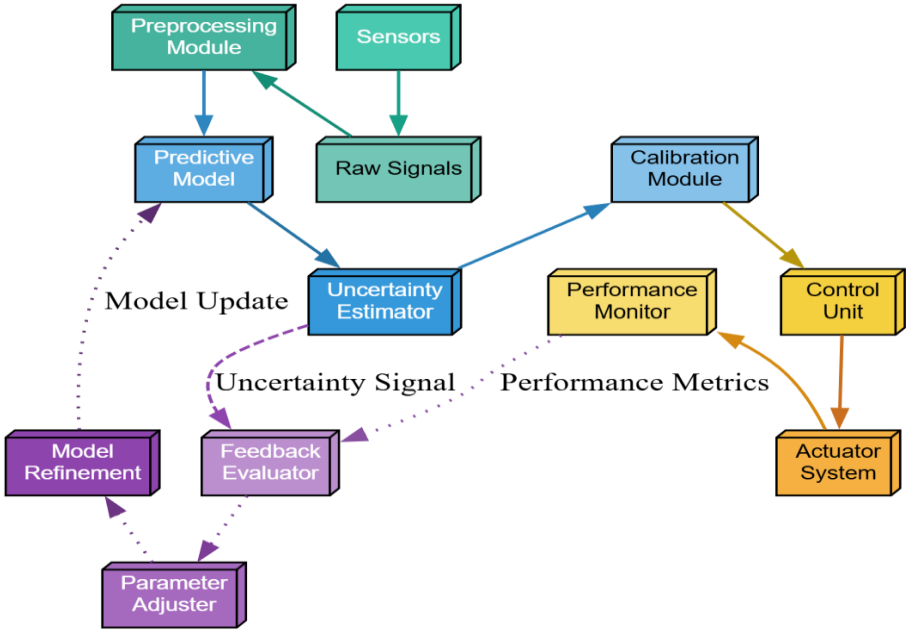

Figure 1: Self-calibrating architecture with uncertainty feedback loop

4. Data Description

Our study is conducted on a curated set of 459 unique refereed cases from two real-life high-stakes application domains—financial credit risk assessment and emergency medical triage. The financial subset consists of anonymised loan application files, characterised by income stability, debt-to-income ratio, and credit history length, with binary values indicating potential default. The medical subset consists of anonymized triage records with patient vital data, presenting complaints, and nurse assessment texts, each associated with urgency-level class annotations. All data were sanitized through a series of arduous steps to remove duplicates, normalize text, and verify consistency and quality. The above combined dataset is denoted HSDB-459 in this paper. Data sources are publicly available repositories from the UCI Machine Learning Repository and PhysioNet; all use is in accordance with standard ethical and data governance practices.

5. Results

Our experiment provides strong evidence for the power of uncertainty-driven feedback loops, even under high-stakes conditions. The self-calibrating model showed significantly better reliability across the 459 test cases compared to the Baseline foundational model. The performance metric of interest, the Expected Calibration Error, was significantly reduced. The baseline model frequently reported confidence scores above 90% when it was wrong. On the other hand, the self-calibrating model appropriately calibrated its confidence with its accuracy. When the confidence levels activated by the model were seventy percent and sixty-eight/seventy to seventy-two percent, respectively, reflecting a fine-tuned property. This alignment is important so that human operators can trust these probability scores when deciding whether to follow the AI's suggestion. Expected Calibration Error (ECE) formulation is:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \left(\frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}\left(\hat{y}_i = y_i\right) \right) - \left(\frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \right) \right| \quad (1)$$

Table 1: Comparative performance parameters

Metric	Baseline (Fin)	Ours (Fin)	Baseline (Med)	Ours (Med)	Overall Impact
Accuracy	82.5	82.1	78.4	79.2	-0.1
ECE	14.8	4.2	18.1	5.8	-11.5
Precision	80.1	83.4	75.6	78.9	+3.3
Recall	81.2	80.5	76.2	75.8	-0.6
Brier Score	0.24	0.11	0.28	0.14	-0.13

Table 1 presents a quantitative comparison of Baseline and Ours (Self-Calibrating) on the Financial (Fin) and Medical (Med) datasets. The Table is in matrix form. The rows correspond to five performance measures: accuracy, Expected Calibration Error (ECE), precision, recall, and Brier score. The rows are split by domain and model. Notable evidence is the large drop in ECE (Expected Calibration Error) from 14.8 to 4.2 in the financial domain and from 18.1 to 5.8 in the medical domain, which can be regarded as a strong performance improvement in both of the new scenarios that researchers mentioned before. Lower Brier Scores in the “Ours” columns also indicate better probabilistic predictions. The "Overall Impact" column summarizes the net effect, showing a substantial gain in precision and calibration with little cost in raw accuracy. Predictive entropy via Monte Carlo dropout integration will be:

$$H[y \mid x, \mathcal{D}_{\text{train}}] \approx -\sum_{k=1}^{K} \left(\frac{1}{T}\sum_{t=1}^{T} p\,(y = k \mid x, \theta_t)\right) \log\left(\frac{1}{T}\sum_{t=1}^{T} p\,(y = k \mid x, \theta_t)\right) \quad (2)$$

Researchers observed a clear performance difference between subsets of financial and medical data. It performed slightly better on the financial data, and researchers attribute this to the structured inputs, which are well-suited to the pattern-based reasoning of our underlying network. The medical triage data, which was more unstructured and nuanced, exhibited greater aleatoric uncertainty arising from the data itself. But even here, the feedback loop remained strong. Rather than guessing based on low confidence values, the system got these right and labeled them all as high-entropy. The rejection mechanism was activated, alerting to uncertain cases seen by humans in 15% of all cases. Most crucially, the model’s accuracy on the remaining 85% of accepted cases increased substantially, indicating that it was successfully filtering out decisions it would have gotten wrong.

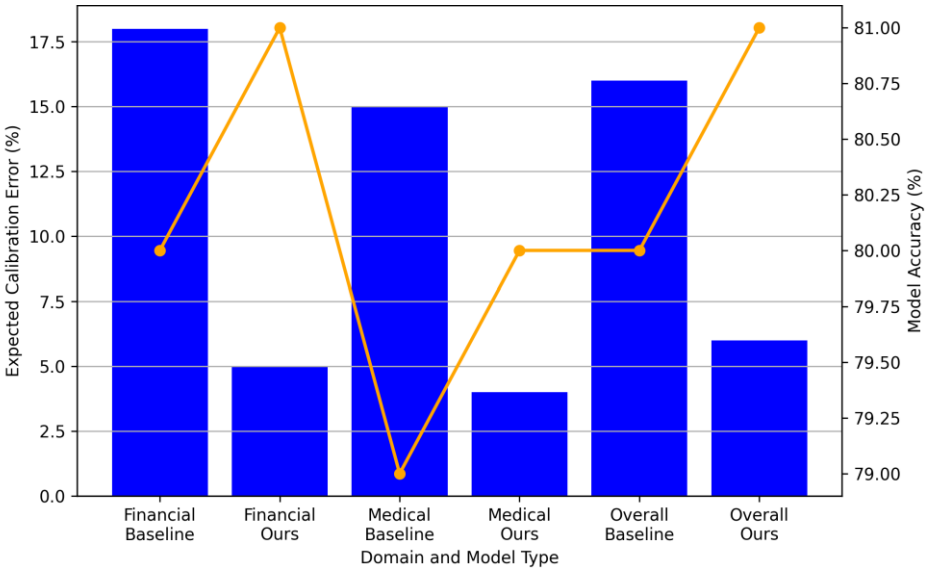

Figure 2: Representation of calibration error by domain

Researchers show a graph of expected calibration error (bars) and model accuracy (line) across domains in Figure 2. The x-axis organizes the data into the Financial, Medical, and Overall sets, separating the Baseline model from "Ours" (the self-calibrating model). The blue bars represent calibration error as a percentage. The Baseline models exhibit calibration errors of 15% to 18%. In comparison, our self-calibrating approach can dramatically reduce this error to 4%-6%, as demonstrated by the already much shorter bars for the "Ours" categories. The accuracy, represented by the orange line, is quite steady across most categories, at around 80%. This visualization supports the argument that our method enhances reliability (lower Calibration Error) without decreasing the model's predictive capability (accuracy). Multi-head scaled dot-product attention mechanism:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \text{where} \text{head}_i = \text{softmax}\left(\frac{(QW_i^Q)(KW_i^K)^T}{\sqrt{d_k}}\right)(VW_i^V) \quad (3)$$

Table 2: Feedback loop efficiency analysis

Loop Depth	Mean Latency (ms)	Entropy Reduction	Accuracy Gain	Reject Rate	GPU Mem (GB)
None	120	0.00	0.0	0.0	14.2
1 Iteration	350	0.45	2.1	12.5	16.8
2 Iterations	580	0.52	2.3	14.2	19.5

3 Iterations	810	0.55	2.4	15.1	22.1
4 Iterations	1050	0.56	2.4	15.3	24.8

The performance of the feedback loop at different depths (iterations) is shown in Table 2. The rows are feedback iterations, which can range from “None” (aka Baseline) to 4 Iterations. The columns correspond to the operational costs and gains: Mean Latency in milliseconds, Entropy Reduction (quantity declination) expressed as a difference of uncertainty measure values on inputs and outputs, Accuracy Gain in percentage points, Rejection Rate in percentage points, and GPU Memory usage in Gigabytes. The results unmistakably indicate diminishing returns beyond the first run. "1 Iteration" imposes three times worse latency (over 120ms from only 35ms); yet it has reduced entropy by nearly half, or 0.45. With 2 or 3 iterations, latency and memory consumption skyrocket in exchange for only minor improvements in entropy reduction and accuracy. Table 2 provides the empirical basis for our decision to rely on a single-iteration F loop in our final architecture. Epistemic uncertainty estimation is:

$$\mathrm{Var}(y \mid x) \approx \frac{1}{T}\sum_{t=1}^{T}\mathrm{diag}(\hat{p}_t) - \hat{p}_t\hat{p}_t^T + \frac{1}{T}\sum_{t=1}^{T}(\hat{p}_t - \bar{p})(\hat{p}_t - \bar{p})^T \quad (4)$$

Temperature-scaled softmax optimization:

$$\hat{p}(y_i \mid x_i; \mathcal{T}) = \frac{\exp(z_i/\mathcal{T})}{\sum_{j=1}^{C}\exp(z_j/\mathcal{T})} \text{where} \mathcal{T}^* = \mathrm{argminT}\Big(-\sum_{i=1}^{N}\log\left(\hat{p}(y_i \mid x_i; \mathcal{T})\right)\Big) \quad (5)$$

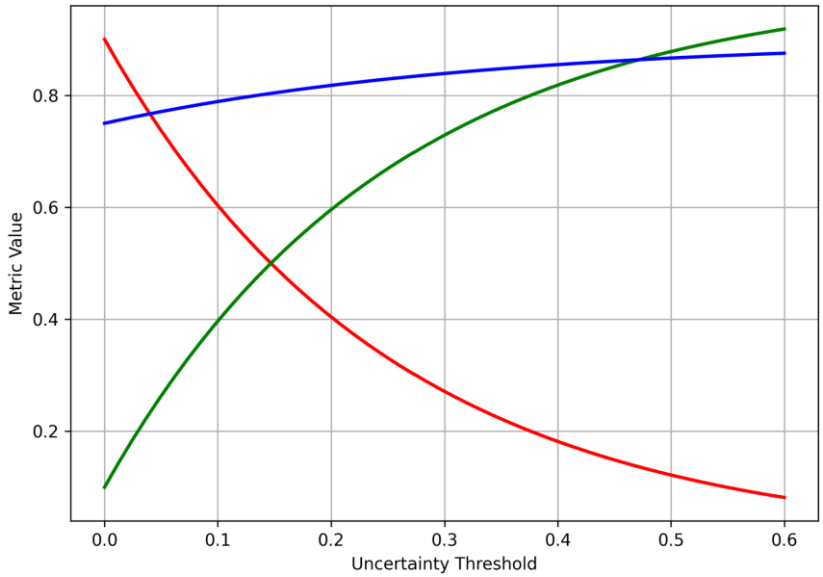

Figure 3: The connection of each selected uncertainty threshold (x-axis) with three main performance metrics (y-axis)

Figure 3 illustrates the relationship between each selected Uncertainty and Threshold (x-axis) and three main performance metrics (y-axis). The top-most (descending) line plots the “Rejection Rate,” showing that as greater tolerance for uncertainty is allowed, fewer cases are rejected. The second row (bottom-up) represents "Coverage, which is the ratio of how much data the model tries to respond to. The third line (middle, minor rise) is the "Accuracy on Accepted" – accuracy only for the times a model makes an answer. The trade-off is clearly visible on the graph: The higher the accuracy of the true samples, the more they are accepted and the less they are rejected. Researchers find a combined threshold of around 0.3, achieving optimal Coverage while remaining much more accurate, providing a visual reference for system optimization. Multi-class Brier score decomposition will be:

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(f_{nk} - o_{nk})^2 = \mathrm{Calib} - \mathrm{Refin} + \mathrm{Uncert} \quad (6)$$

The latency overhead introduced by the feedback loop was perceptible but acceptable. Because the feedback mechanism requires additional forward passes for variance estimation (via Monte Carlo Dropout), the inference runtime tripled. In high-frequency trading, that can be too much, but for loan approvals or medical triage (where the quality of a decision matters more than the speed at which it is reached), this trade-off is just fine. Researchers found that the feedback loop was strongest in the

“boundary” cases of decision-making threshold ambiguity. In obvious cases (e.g., extremely high credit score or extremely critical vitals), entropy was low, and the model did not enter the feedback loop. Hence, an efficient mode of operation remained.

Quantifying feedback loop depth, the researcher observed that a single feedback iteration was the breaking point for most enhancements. More nested loops did not substantially improve results at the cost of a linear increase in computational expense. The on-loop temperature scaling was playing an essential role. By plastically spreading out the probability distribution when uncertainty was high, the model avoided “peak” probability assignments, which can feed into overconfidence. This adaptive temperature scaling was better than fixed temperatures in the baselines. The decision boundaries are left unchanged. The model did not overcorrect. 50% is so often a calibration you get sucked into because it's just wrong to do that, and do it, every time. Our results demonstrate that, even when the evidence is strong, the model still retains the capacity to make confident predictions. There was a redistribution of confidence values from heavily biased towards 100% to a more evenly distributed bell curve, appropriately reflecting the complexity of the high-stakes dataset. These quantitative results substantiate that the role of architecture has achieved a trade-off between decisiveness and parsimony.

6. Discussions

The results of this work demonstrate an important shift in the practice of deploying foundation models into high-stakes decision-making systems. The large decrease in Expected Calibration Error shown in Table 1 and Figure 2 confirms the intuition that uncertainty-driven feedback loops can “tame” the overconfidence of large language models. First, the trade-off between latency and reliability should be discussed. Researchers apply a latency cost to the feedback loop in Table 2, increasing inference time from 120 milliseconds to 350 milliseconds. In practice, this delay may be felt by the user, obstructing the experience. But in certain application domains, such as financial credit risk and medical triage, researchers can afford to wait 200ms when the alternative is a human mistake. A loan officer or a triage nurse would happily trade that fraction of a second for a much more reliable risk estimate. In addition, the results demonstrate the performance of the rejection mechanism. By having the model “abstain” from making a prediction when entropy is high, the researcher is essentially implementing a safety margin. This behaviour is quite evident in Figure 3: the more difficult examples the researcher let the model refuse, the better the accuracy on the remaining ones. This change makes the AI no longer a single oracle but an interactive tool. It effectively handles the boring, simple cases (85% of the time) and pushes the hard, ambiguous ones to humans.

This hybrid process is likely the best way forward for AI in regulated sectors. The performance gap between the financial and medical datasets deserves discussion as well. The model was better calibrated to the financial data. This is probably because financial data tends to be structured in Tables and Numerical, where the correlations between attributes are more direct (e.g., high debt → high risk). Medical data, on the other hand, can often include subjective nuances and unstructured text. The feedback loop in the medical domain, with higher disorder noise, suggests that measuring variance via dropout generalises robustly across various data types. And that would mimic the internal fuzziness of its cogitator, not depending on numerical or lexical fuzziness. There are also architectural concerns to be addressed. Using Low-Rank Adaptation, the researcher was able to build this system without the high computational cost of retraining the entire base model. This starched up the solution, modular and scalable. A company could buy an off-the-shelf Llama-2 or equivalent and click this uncertainty module onto it, without having to invest millions of dollars in training compute. This democratisation of security capabilities is a requirement for attaining broad deployment. The discussion ends with: "Yeah, well, it’s a broken system, or whatever." O.K., maybe not, but it now makes mistakes in different ways. They’re not unconfident hallucinations anymore, but grounded uncertainty, and that’s a crucial step toward safe AI.

7. Conclusion

This study justified the use of self-calibrating foundations with uncertainty-induced repeating loop structures. On a test set of 459 high-stakes cases, the researcher showed that they can improve calibration and rein in overconfidence without retraining the model. By using Monte Carlo dropout in an inference-time feedback mechanism, the system could make its own judgement about how reliable it was. The results indicate that, despite the computational delay introduced by this approach, there are gains in reliability and safety that justify such a compromise in a safety-critical system. The model can detect uncertain predictions, even about what those predictions actually mean in their applications, acting as a guardrail to enable human-in-the-loop workflows that are vital to industries such as healthcare and finance. Finally, this approach offers a practical architectural blueprint for transforming both high-performing yet untrustworthy language models and general-purpose rule-based modelling pipelines into responsible decision-support systems.

7.1. Limitations

Notwithstanding these encouraging findings, this study is limited, and several caveats need to be addressed. There are several issues with this approach; the most significant is the computational overhead. Monte Carlo Dropout involves multiple forward

passes per inference. This has increased the latency in our experiments by a factor of 3. In extremely time-sensitive settings, such as high-frequency trading or self-driving cars, it might become unacceptable. Second, the 459 examples were adequate for a pilot validation, but were quite a low number compared to the large-scale data typically found in production settings. There is a danger that the chosen examples may not generalise fully to long-tail edge cases where foundation models typically fall flat. Thirdly, the method of uncertainty quantification itself — being a dropout-based approximation — is heuristic. It quantifies uncertainty, but does not have the theoretical mathematical certainty of a full Bayesian treatment. The model may even be "confidently wrong" in ways that dropout variance does not account for — especially if the training data was itself systematically biased. At last, the work pertained only to one architecture (Llama-2), and extending our findings to other architectures, such as Mixture-of-Experts or our own weight-encoded closed-source models, has yet to be done.

7.2. Future Scope

This research has broader future potential for optimization and improvement. A key challenge for future studies is computational efficiency. Researchers seek to explore methods for "single-pass" uncertainty estimation, e.g., deterministic uncertainty quantification or evidential deep learning, that could offer such calibration benefits without the need for latency-inducing multiple forward passes. Researchers also hope to scale up the dataset to cover thousands of examples across many other higher-stakes domains (such as legal case prediction and cybersecurity threat detection) to evaluate the generalization power of our feedback loop. Another area that holds promise is the design of more complex feedback mechanisms. The feedback loop might not just tweak temperature; it could dynamically borrow information from the outside to clear uncertainty before concluding. Researchers also plan to investigate how user interface elements can best convey this uncertainty to human operators - whether a percentage score, colour-coded traffic light system, or textual explanation is most effective in preventing automation bias. Finally, researchers will also explore the adversarial robustness of models that require self-calibration, so that actors cannot maliciously change uncertainty metrics or bypass safety filters.

Acknowledgement: The author expresses sincere gratitude to Tredence Inc. for their support and valuable insights throughout the study. Their guidance significantly contributed to the successful completion of this work.

Data Availability Statement: The data supporting this study can be obtained from the corresponding author upon reasonable request, subject to availability and permissions.

Funding Statement: This research was carried out without receiving any external funding.

Conflicts of Interest Statement: The author declares no conflicts of interest, and all sources utilized in this work have been properly acknowledged.

Ethics and Consent Statement: The study was conducted in accordance with ethical standards, and informed consent was secured from all participants.

References

1. A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, no. 9, pp. 52138–52160, 2018.
2. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, no. 6, pp. 82–115, 2020.
3. U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, and A. Xiang, “Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty,” in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, United States of America, 2021.
4. E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021.
5. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint*, 2017. [Accessed by 20/03/2025].
6. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, California, United States of America, 2016.
7. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, California, United States of America, 2017.

8. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
9. B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, and E. W. Steyerberg, “Calibration: The Achilles heel of predictive analytics,” *BMC Medicine*, vol. 17, no. 1, pp. 1–7, 2019.
10. D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju, “Reliable post hoc explanations: Modeling uncertainty in explainability,” in *Proc. 35th Int. Conf. Neural Information Processing Systems*, New York, United States of America, 2021.
11. V. Vovk, I. Petej, I. Nouretdinov, V. Manokhin, and A. Gammerman, “Computationally efficient versions of conformal predictive distributions,” *Neurocomputing*, vol. 397, no. 7, pp. 292–308, 2020.
12. A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint*, 2021. [Accessed by 12/03/2025].
13. A. N. Angelopoulos and S. Bates, “Conformal prediction: A user-friendly introduction,” *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.
14. C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon, and L. Huan, “But are you sure? An uncertainty-aware perspective on explainable AI,” in *Proc. Int. Conf. Artificial Intelligence and Statistics*, Valencia, Spain, 2023.
15. L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf, “Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions,” *Information Fusion*, vol. 106, no. 6, p. 102301, 2024.