Why Are LLMs Confidently Wrong? Correcting Overconfident Errors via Causal Head Intervention

Jing Ren[1,#] **Bowen Li**[1,#] **Ziqi Xu**[1,*] **Xuechao Yang**[1] **Feng Xia**[1]

[1]RMIT University

{S4065838, S3890442}@student.rmit.edu.au
{ziqi.xu, xuechao.yang, feng.xia}@rmit.edu.au

[#]Equal contribution [*]Corresponding author

https://aisuko.github.io/confix

Abstract

Modern large language models (LLMs) often exhibit overconfident errors, assigning high confidence to incorrect predictions and thereby undermining reliability in real-world use. Existing calibration methods typically rely on post-hoc adjustment or retraining, operating at the output level without addressing the underlying causes of miscalibration. We propose **CoCHI** (**C**lass-**o**riented **C**ausal **H**ead **I**ntervention), a framework that improves calibration by identifying and intervening on internal components responsible for distinct confidence behaviours. By analysing predictions through correctness-confidence patterns, CoCHI enables targeted intervention on model internals, reducing overconfident errors while preserving correct predictions. Our approach provides a mechanistic perspective on calibration, linking reliability to identifiable structures within the model rather than treating it as an output-level issue. Experiments across multiple LLMs show consistent improvements in calibration metrics, including Expected Calibration Error, Negative Log-Likelihood, and Brier Score, while maintaining accuracy on five multiple-choice question answering benchmarks. These results suggest that effective calibration requires not only adjusting outputs, but also addressing the underlying causal mechanisms within the model.

1 Introduction

Modern neural models have achieved remarkable performance across a wide range of tasks. However, they frequently produce overconfident errors, assigning high confidence to incorrect predictions [1, 2, 3, 4] and thereby undermining reliability in high-stakes applications. This issue is further amplified in large language models (LLMs) [5, 6, 7], which have become a dominant paradigm in modern AI and are widely deployed in domains such as healthcare, education, and scientific reasoning, where incorrect yet confident outputs can lead to significant downstream risks [8, 9, 10, 11, 12].

To address this problem, a large body of work has focused on *calibration*, aiming to align predicted confidence with empirical accuracy [13, 14, 15]. Existing approaches, such as temperature scaling and its variants, primarily operate at the output level by adjusting logits or confidence scores, while more recent methods introduce learned calibrators or inference-time interventions [16]. Despite their effectiveness in some settings, these approaches share a fundamental limitation: they treat miscalibration as a surface-level phenomenon and do not address the *internal mechanisms* that give rise to overconfident errors. As a result, they often fail to generalise and provide limited insight into why models are confidently wrong.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we take a different perspective and ask: *what internal components of a model are responsible for overconfident errors, and can we intervene on them directly?* Recent advances in mechanistic interpretability [17, 18, 19, 20] have made significant progress in moving beyond correlational post-hoc explanations toward causally grounded, intervention-based understanding of how LLMs implement computations internally. In particular, causal intervention methods such as causal head gating [17], causal mediation analysis [18, 21], and causal tracing [19] allow us to identify internal components that are causally responsible for specific model behaviours by systematically perturbing internal activations and observing the resulting change in the model's output distribution. However, existing interpretability approaches largely focus on post-hoc analysis and explanation, rather than leveraging these insights to actively improve model reliability. Consequently, the connection between internal computational circuits and miscalibration remains largely unexplored.

To bridge this gap, we propose **CoCHI** (**C**lass-**o**riented **C**ausal **H**ead **I**ntervention), a framework that improves calibration by directly targeting the internal causes of overconfident errors in large language models. Our key insight is that miscalibration is not merely an output-level issue, but arises from specific internal components that systematically promote incorrect yet highly confident predictions or introduce noise into otherwise correct ones. To capture this, we decompose model predictions into four calibration classes based on correctness and confidence, and perform class-conditioned causal analysis to identify attention heads associated with distinct behaviours. Based on this analysis, CoCHI constructs a structured intervention that selectively suppresses heads responsible for overconfident errors while removing non-contributing components in correct predictions. This results in a minimal, targeted modification to the model's internal computation that improves calibration without altering model parameters, providing a mechanism-aware alternative to conventional post-hoc approaches.

Contributions. Our main contributions are as follows:

- We propose a structural causal mask for calibration, operating directly on model internals rather than relying on post-hoc output adjustment.
- We partition model predictions into four fine-grained classes based on correctness and confidence, enabling precise identification of attention heads that causally promote overconfident errors or fail to support correct predictions.
- We introduce CoCHI (Class-oriented Causal Head Intervention), an inference-time framework that performs class-aware causal analysis and constructs a structured mask to selectively suppress problematic attention heads.
- Extensive experiments across five multiple-choice question answering benchmarks show that CoCHI consistently improves calibration metrics while preserving accuracy, outperforming traditional post-hoc calibration methods.

2 Related Work

Uncertainty Estimation and Overconfident Errors. A large body of work studies uncertainty estimation in neural models, including Bayesian approaches [22, 23, 24, 25, 26], deep ensembles [27], and post-hoc techniques such as temperature scaling [16], which rescale logits using a learned temperature parameter. In the context of LLMs, uncertainty estimation and calibration have received increasing attention [28]. Recent efforts extend post-hoc calibration to LLMs, including adaptive temperature scaling [29] and learned calibrators such as Thermometer [30], aiming to better align predicted confidence with empirical accuracy. Other approaches adjust hidden representations, for example by training a linear layer on top of the final hidden states for confidence prediction [31], or exploit prompt-based strategies and self-consistency to estimate uncertainty [32, 33, 34, 35, 15]. Despite their effectiveness in some settings, these methods primarily operate at the output or representation level, but recent work such as Inference-Time Intervention [36] shows that undesirable behaviours (e.g., hallucination) are governed by internal mechanisms. These methods primarily enhance truthful representations, while overlooking the role of mechanisms that specifically drive overconfident errors, where incorrect predictions are assigned high confidence scores. Recent work has shown that attention heads encode rich signals about model uncertainty, enabling unsupervised uncertainty estimation without additional training [37]. However, these approaches primarily focus on extracting uncertainty signals from internal representations, without modifying the model's behaviour.

This limitation motivates our approach, which targets the structural causal components responsible for overconfident errors and performs structured intervention to improve calibration.

Circuit Discovery in Mechanistic Interpretability Circuit discovery aims to identify the minimal subgraph of a language model responsible for a specific behaviour. The seminal work of [38] manually identified the Indirect Object Identification (IOI) circuit in GPT-2 Small, comprising 28 attention heads across seven functional classes. To scale this process, Conmy et al. [39] proposed Automatic Circuit Discovery (ACDC), an iterative pruning method based on activation patching, which successfully rediscovered known circuits such as Greater-Than and recovered most components of the IOI circuit. Subsequent approaches, including Efficient Attribution Patching (EAP), further improved the efficiency and scalability of automated circuit discovery [40]. Beyond circuit discovery, mechanistic interpretability has increasingly incorporated tools from causal inference [41]. These approaches leverage counterfactual interventions and structured perturbations to uncover how structural heads contribute to model outputs [42, 43, 44, 45]. In particular, causal head gating identifies functional roles of attention heads by assigning graded scores for facilitation, interference, or irrelevance based on their causal effect on task performance, going beyond purely correlational analyses [17]. Recent mechanistic studies have shown that verbalised confidence in LLMs is governed by dedicated internal circuits that are largely decoupled from factual correctness [46]. These findings suggest that overconfidence arises not from miscalibration at the output layer, but from specific internal mechanisms that consistently generate confident responses regardless of their accuracy [47]. Building on this insight, we identify and suppress attention heads that contribute to high-confidence errors, leading to improved calibration and reliability.

3 Preliminaries

We briefly introduce the key components for identifying the causal roles of attention heads in transformer models via differentiable intervention.

Transformer with Head-wise Gating. Consider a transformer model with L layers and H attention heads per layer. For each attention head (ℓ, h), we introduce a continuous gating variable $G_{\ell,h} \in [0, 1]$ that scales the output of the head:

$$Z_{\ell,h} = G_{\ell,h} \cdot \big(A_{\ell,h} V_{\ell,h}\big), \tag{1}$$

where $A_{\ell,h}$ denotes the attention weights (after softmax) and $V_{\ell,h}$ denotes the value projections.

The gated head outputs are concatenated and projected as in standard multi-head attention. By varying $G_{\ell,h}$, we perform soft interventions on individual heads without modifying model parameters.

Gate Optimisation. The gating variables are optimised while keeping model parameters frozen. Given a dataset $\mathcal{D}$, we minimise the negative log-likelihood objective:

$$\mathcal{L}(G) = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \log P(y \mid x; G), \tag{2}$$

where $P(y \mid x; G)$ denotes the model output under gating configuration G.

Optionally, a regularisation term can be introduced to encourage sparsity or density of the gating values. This formulation allows us to assess the causal contribution of each head by observing how scaling its output affects model behaviour.

Dual Optimisation and Role Identification. To disambiguate causal importance from incidental effects, we optimise two sets of gates:

- G^+: optimised under a retention bias, encouraging gates towards 1,
- G^-: optimised under a sparsity bias, encouraging gates towards 0.

The divergence between G^+ and G^- reveals the causal role of each head: heads that remain active under sparsity are necessary for the task, while those suppressed even under retention bias are likely harmful.

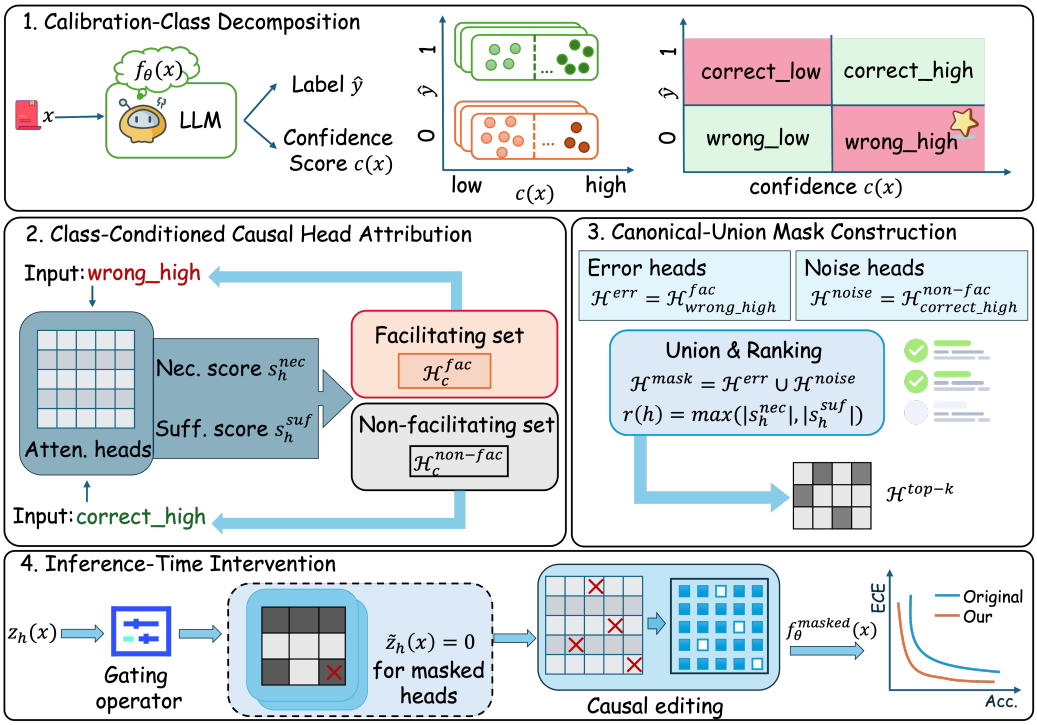


Figure 1: The framework of CoCHI.

Causal Head Taxonomy. Based on the optimised gates, each attention head is assigned a causal role. **Facilitating heads** contribute positively to performance and remain active under sparsity (G^-), indicating necessity. **Interfering heads** degrade performance and are suppressed even under retention bias (G^+), indicating harmful contributions. **Irrelevant heads** have negligible impact, with gating values varying under different regularisation regimes. This taxonomy provides a causal, rather than correlational, characterisation of model internals, validated through targeted ablation and intervention.

4 Methodology

We reframe calibration as a problem of causal intervention on model heads. Specifically, our framework identifies and suppresses causal heads responsible for overconfident errors, while removing sources of noise in otherwise correct predictions. This perspective provides a mechanistic view of miscalibration, linking reliability to identifiable structural components within the model. An overview of **CoCHI** is shown in Figure 1.

4.1 Problem Setup

Let $f_\theta(x)$ denote a pretrained language model that maps an input x to a probability distribution $p_\theta(\cdot \mid x)$ over a fixed answer space $\mathcal{Y}$. The model prediction is defined as:

$$\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x). \tag{3}$$

The corresponding confidence is defined as:

$$c(x) = \max_{y \in \mathcal{Y}} p_\theta(y \mid x). \tag{4}$$

Let y^* denote the ground-truth label. A prediction is correct if $\hat{y}(x) = y^*$. We focus on the failure mode of *overconfident errors*, defined as incorrect predictions with high confidence:

$$\hat{y}(x) \neq y^*, \quad c(x) \geq \tau_c, \tag{5}$$

where τ_c is a predefined confidence threshold.

Our goal is to design an inference-time intervention that reduces such errors by targeting their underlying internal causes, thereby improving calibration while preserving predictive accuracy.

4.2 Calibration-Class Decomposition

Rather than treating calibration as a global property of output probabilities, we decompose model behaviour into four calibration classes based on correctness and confidence:

$$\mathcal{C} = \{\texttt{correct_high}, \texttt{correct_low}, \texttt{wrong_high}, \texttt{wrong_low}\}.$$

Formally, this defines a partition over predictions induced by correctness ($\hat{y}(x) = y^*$ or $\hat{y}(x) \neq y^*$) and confidence ($c(x) \geq \tau_c$ or $c(x) < \tau_c$). This decomposition isolates overconfident errors (`wrong_high`) as a distinct subset, enabling us to analyse the structural causal head that give rise to miscalibrated confidence, rather than conflating them with other types of errors.

4.3 Class-Conditioned Causal Head Attribution

Following Causal Head Gating (CHG) [17], we attribute model behaviour to individual attention heads by estimating their causal contribution. Let $\mathcal{H}$ denote the set of all attention heads across l layers. For each head $h \in \mathcal{H}$, CHG computes two scores with respect to a target behaviour defined on a dataset $\mathcal{D}$: a *necessary score* s_h^{nec}, measuring the drop in performance when the head is ablated, and a *sufficient score* s_h^{suf}, measuring how well the head alone can reproduce the behaviour when isolated.

In our setting, we apply CHG in a class-conditioned manner. For each calibration class $c \in \mathcal{C}$, we construct a subset $\mathcal{D}_c$ consisting of inputs whose predictions fall into class c, and perform CHG using the model's own predictions as the target signal. This yields a *necessary score* $s_h^{\text{nec}}(c)$ and a *sufficient score* $s_h^{\text{suf}}(c)$ for each head. The necessary score measures how strongly a head persists when the optimisation is encouraged to minimise the number of active heads, whereas the sufficient score measures how strongly a head remains selected when the optimisation is encouraged to retain all heads compatible with good task performance.

CoCHI is trained in three sequential stages. First, an unregularised optimisation stage ($\lambda = 0$) is performed to obtain a shared task-relevant initialisation:

$$z_{l,h}^0 \xrightarrow{\text{stage 1}} \hat{z}_{l,h}. \tag{6}$$

The resulting logits $\hat{z}_{l,h}$ are then copied into two separate optimisation branches:

$$\hat{z}_{l,h} \to z_{l,h}^+, \quad \hat{z}_{l,h} \to z_{l,h}^-, \tag{7}$$

corresponding to the positive-L1 and negative-L1 stages, respectively.

During each forward pass, the learned logit is converted into a multiplicative gate using a sigmoid function and applied directly to the attention head output:

$$\tilde{x}_{l,h} = \sigma(z_{l,h}) \cdot x_{l,h}, \tag{8}$$

where $x_{l,h} \in \mathbb{R}^{T \times d_h}$ denotes the output of head h at layer l, and T is the sequence length. A gate value close to 0 suppresses the head, while a value close to 1 preserves its contribution.

The optimisation objective is defined as:

$$\mathcal{L} = -\frac{1}{|S|} \sum_{t \in S} \log p_\theta(x_{t+1} \mid x_{\leq t}) - \lambda \cdot \frac{1}{LH} \sum_{l,h} \text{clamp}(z_{l,h}, -4, 4), \tag{9}$$

where the first term is the next-token negative log-likelihood and the second term is the regularisation objective. The logits are clamped to $[-4, 4]$ before regularisation to stabilise optimisation and encourage near-binary gate values.

Two complementary optimisation stages are then performed. In the positive-L1 stage ($\lambda = +0.1$), the optimiser is encouraged to suppress unnecessary heads, so only heads whose removal significantly increases the NLL retain large gate values. These heads correspond to necessary heads. In the

negative-L1 stage ($\lambda = -0.1$), the optimiser instead encourages larger gate values, identifying heads that are sufficient to maintain low task loss.

At the end of optimisation, the final sigmoid gate values are recorded as:

$$s_h^{\text{nec}}(c) = \sigma(z_{l,h}^{+})^{(\text{final})}, \tag{10}$$

$$s_h^{\text{suf}}(c) = \sigma(z_{l,h}^{-})^{(\text{final})}, \tag{11}$$

which define the necessary and sufficient scores, respectively.

A binary threshold $\tau = 0.5$ is then applied to derive discrete causal roles:

$$\text{role}_{l,h} = \begin{cases} \textbf{facilitating}, & s_h^{\text{nec}}(c) \geq \tau \ \wedge \ s_h^{\text{suf}}(c) \geq \tau, \\ \textbf{interfering}, & s_h^{\text{nec}}(c) < \tau \ \wedge \ s_h^{\text{suf}}(c) \geq \tau, \\ \textbf{irrelevant}, & \text{otherwise.} \end{cases} \tag{12}$$

A head is considered *facilitating* for class c if both scores exceed a threshold τ, indicating that it is both necessary and sufficient for producing the corresponding behaviour; otherwise, it is treated as non-facilitating, encompassing both interfering and irrelevant heads.

This procedure yields two sets per class, $\mathcal{H}_c^{\text{fac}}$ and $\mathcal{H}_c^{\text{non-fac}}$. Crucially, this class-conditioned decomposition associates specific structural causal head with distinct calibration behaviours and forms the basis for targeted intervention in subsequent steps.

4.4 Canonical-Union Mask Construction

We construct a targeted structural causal mask for calibration by combining two complementary types of heads. First, from the `wrong_high` class, we extract the facilitating heads:

$$\mathcal{H}^{\text{err}} = \mathcal{H}^{\text{fac}}_{\texttt{wrong_high}}, \tag{13}$$

which contribute to generating high-confidence incorrect predictions. These heads can be interpreted as forming the targeted structural causal mask that amplify erroneous outputs.

Second, from the `correct_high` class, we extract the non-facilitating heads:

$$\mathcal{H}^{\text{noise}} = \mathcal{H}^{\text{non-fac}}_{\texttt{correct_high}}, \tag{14}$$

which do not support correct predictions and may introduce variability in confidence estimation.

We then define the candidate mask set as the union $\mathcal{H}^{\text{mask}} = \mathcal{H}^{\text{err}} \cup \mathcal{H}^{\text{noise}}$. To avoid excessive intervention, we rank heads by their attribution strength:

$$r(h) = \max\left(|s_h^{\text{nec}}|, |s_h^{\text{suf}}|\right), \tag{15}$$

and select the top-k heads to form a compact mask $\mathcal{H}^{\text{top-}k}$. This ranking focuses the intervention on the most influential components while preserving the overall functionality of the model.

4.5 Inference-Time Intervention

Given the identified set of influential heads $\mathcal{H}^{\text{top-}k}$, we perform a structured intervention on the model's internal computation at inference time. Rather than modifying model parameters or post-processing outputs, our approach directly alters the forward pass by selectively suppressing specific attention heads.

Let $z_h(x)$ denote the output of attention head h for input x. We define a binary gating operator $g_h \in \{0, 1\}$ for each head as:

$$g_h = \begin{cases} 0, & \text{if } h \in \mathcal{H}^{\text{top-}k}, \\ 1, & \text{otherwise,} \end{cases} \tag{16}$$

and obtain the modified head output:

$$\tilde{z}_h(x) = g_h \cdot z_h(x). \tag{17}$$

This yields a perturbed model $f_\theta^{\text{masked}}(x)$.

This operation can be interpreted as a form of *causal editing* on the model's computation graph, where selected edges (i.e., head outputs) are removed to disrupt specific functional circuits. In particular, masking heads in $\mathcal{H}^{\text{err}}$ suppresses pathways responsible for overconfident errors, while masking heads in $\mathcal{H}^{\text{noise}}$ reduces spurious variation in otherwise correct predictions.

The intervention is both *local* and *minimal*, modifying only a small subset of heads and thus preserving most of the model's representational capacity. The calibration classes are used exclusively during mask construction. At inference time, the resulting global mask is applied uniformly to all test inputs without requiring ground-truth labels or sample-specific class identification. From a computational perspective, the overhead is negligible, as the intervention involves only element-wise gating during the forward pass. Since no parameters are updated, the method can be seamlessly applied to pretrained models in a plug-and-play manner.

5 Experiments

5.1 Implementation Details

In practice, we apply CoCHI on a small subset of the data (typically fewer than 50 samples), with balanced sampling across calibration classes to ensure stable attribution estimates. We use a fixed confidence threshold $\tau_c = 0.5$ to define calibration classes, and a threshold τ to identify facilitating heads. The number of masked heads k is treated as a hyperparameter; we find that small values (e.g., $k \in [6, 8]$) provide a favourable trade-off between calibration improvement and accuracy preservation. The penalty coefficient was empirically tuned across 15 model–dataset runs, yielding a stable trade-off that limits degradation on `correct_high` calibration ($\Delta\text{ECE} < 0.015$) while still effectively suppressing overconfident errors on incorrect predictions. We evaluate our method on three representative open-source large language models: `LLaMA-3.1-8B` [7], `Gemma-2-9B` [48], and `Mistral-7B-v0.1` [49]. We adopt `LLaMA-3.1-8B` as the primary backbone for ablation and hyperparameter studies. All experiments are conducted in a zero-shot setting, ensuring that improvements in calibration are solely attributable to the proposed inference-time intervention. Further experimental details are provided in Appendix A.

5.2 Evaluation Protocols

Datasets. We evaluate our method on five widely used multiple-choice question answering benchmarks spanning diverse domains, including general knowledge, science, and medicine, to assess both generalisation and domain-specific robustness of calibration. Measuring Massive Multitask Language Understanding (MMLU) [50] is a large-scale multi-domain benchmark covering 57 subjects such as STEM, humanities, and social sciences, designed to evaluate general knowledge and reasoning ability. MedMCQA [51] is a large medical QA dataset derived from real-world entrance examinations, focusing on domain-specific knowledge in healthcare. OpenBookQA [52] is a science QA benchmark requiring multi-hop reasoning over a small set of core science facts. MedQA-USMLE [53] is a challenging medical benchmark based on USMLE exams, evaluating clinical reasoning and professional knowledge. SciQ [54] is a science QA dataset constructed from educational materials, focusing on factual understanding and basic reasoning. All datasets are formatted as four-choice multiple-choice tasks, and we adopt a unified evaluation protocol.

Baselines. We compare our method against several representative calibration approaches, including both post-hoc and inference-time techniques. Temperature Scaling (TS) [16] is a standard post-hoc method that learns a single scalar temperature T on a held-out validation set to rescale logits. Inference-Time Intervention (ITI) [36] modifies model activations during inference using a fixed intervention strength α, enabling reliability improvements without retraining. Thermometer (Therm.) [30] is a learned calibration method that trains a lightweight MLP-based calibrator on top of model outputs to better align predicted confidence with empirical accuracy. Adaptive Temperature Scaling (ATS) [29] extends TS by dynamically adapting the temperature based on input features, allowing more flexible calibration across samples.

Metrics. We evaluate performance in terms of both accuracy and calibration. Prediction quality is measured by **accuracy**, defined as the proportion of correctly predicted answers. Calibration is assessed using **Expected Calibration Error (ECE)** [16], which measures the discrepancy between predicted confidence and empirical accuracy across confidence bins; **Negative Log-Likelihood**

Table 1: Results on four benchmarks (MMLU, MedMCQA, OpenBookQA, and MedQA-USMLE). Best results are highlighted in bold.

	MMLU				MedMCQA			
Method	Accuracy ↑	ECE ↓	NLL ↓	BS ↓	Accuracy ↑	ECE ↓	NLL ↓	BS ↓
Gemma-2-9B								
TS	0.6653±0.0050	0.0660±0.0173	0.5060±0.0335	0.1683±0.0125	0.5607±0.0186	0.0556±0.0261	0.6114±0.0151	0.2122±0.0066
ITI	0.6693±0.0060	0.0605±0.0062	0.5029±0.0233	0.1675±0.0105	0.5673±0.0200	0.0451±0.0140	0.6085±0.0154	0.2122±0.0070
Therm.	0.6653±0.0050	0.1831±0.0567	1.8729±1.1455	0.2302±0.0429	0.5607±0.0186	0.2849±0.0106	2.9242±0.4188	0.3293±0.0165
ATS	0.6653±0.0050	0.3784±0.0052	0.9191±0.0066	0.3553±0.0030	0.5607±0.0186	0.3534±0.0783	4.1559±1.3606	0.3702±0.0396
CoCHI	**0.6730±0.0023**	**0.0601±0.0011**	**0.4982±0.0141**	**0.1654±0.0056**	**0.5873±0.0023**	**0.0380±0.0263**	**0.5983±0.0064**	**0.2070±0.0027**
LLaMA-3.1-8B								
TS	0.6007±0.0061	0.0569±0.0040	0.5205±0.0138	0.1936±0.0065	0.5753±0.0297	0.0557±0.0124	0.6173±0.0142	0.2144±0.0052
ITI	0.5853±0.0685	0.2014±0.1191	0.7161±0.1590	0.2580±0.0643	0.3747±0.0893	0.3766±0.2287	1.2043±0.7682	0.3831±0.1870
Therm.	0.6007±0.0061	0.1717±0.0778	1.9024±0.7247	0.2415±0.0354	0.5753±0.0297	0.2327±0.0356	1.8116±0.8973	0.2907±0.0208
ATS	0.6007±0.0061	0.3265±0.0257	1.7325±1.5251	0.3330±0.0157	0.5753±0.0301	0.3307±0.0727	2.5476±2.9442	0.3495±0.0501
CoCHI	**0.6120±0.0626**	**0.0490±0.0089**	**0.5168±0.0236**	**0.1743 ±0.0008**	**0.5773±0.0081**	**0.0365±0.0090**	**0.6064±0.0032**	**0.2114±0.0012**
Mistral-7B								
TS	0.5987±0.0121	0.0902±0.0182	0.5754±0.0130	0.1974±0.0037	**0.4867±0.0081**	0.0604±0.0308	0.6424±0.0339	0.2249±0.0133
ITI	0.3980±0.1152	0.3693±0.2666	1.2168±0.8397	0.3845±0.2056	0.3293±0.0939	0.3985±0.3280	1.4918±1.2039	0.4317±0.2469
Therm.	0.5987±0.0121	0.2988±0.0449	1.4277±0.9584	0.3177±0.0274	**0.4867±0.0081**	0.2668±0.0776	2.6632±0.6919	0.3385±0.0372
ATS	0.5987±0.0121	0.3379±0.0290	2.5765±1.4883	0.3462±0.0184	**0.4867±0.0081**	0.3817±0.1376	3.6701±2.4950	0.4032±0.0945
CoCHI	**0.6100±0.0040**	**0.0854±0.0015**	**0.5649±0.0008**	**0.1941±0.0005**	0.4847±0.0023	**0.0383±0.0035**	**0.6168±0.0012**	**0.2146±0.0006**

	OpenBookQA				MedQA-USMLE			
Method	Accuracy ↑	ECE ↓	NLL ↓	BS ↓	Accuracy ↑	ECE ↓	NLL ↓	BS ↓
Gemma-2-9B								
TS	0.6080±0.0087	0.0686±0.0119	0.5410±0.0076	0.1848±0.0054	0.7714±0.0124	0.1782±0.0188	0.5109±0.0377	0.1829±0.0118
ITI	**0.6100±0.0111**	0.0682±0.0064	0.5438±0.0070	0.1848±0.0043	**0.7927±0.0151**	0.1725±0.0143	0.5091±0.0163	0.1708±0.0072
Therm.	0.6080±0.0087	0.2635±0.0266	2.4829±0.5178	0.2930±0.0223	0.7810±0.0114	0.1767±0.0298	1.7598±0.2096	0.1757±0.0205
ATS	0.6080±0.0087	0.3350±0.0165	1.9401±1.8405	0.3399±0.0090	0.7740±0.0164	0.1863±0.0131	2.2281±0.2476	0.1866±0.0141
CoCHI	0.6090±0.0766	**0.0434±0.0116**	**0.5353±0.0268**	**0.1840±0.0092**	0.7893±0.0223	**0.1723±0.0070**	**0.5081±0.0046**	**0.1694±0.0019**
LLaMA-3.1-8B								
TS	**0.6827±0.0170**	0.0618±0.0106	0.5776±0.0125	0.1853±0.0066	**0.7360±0.0087**	0.1198±0.0200	0.5603±0.0394	0.1750±0.0123
ITI	0.5200±0.1618	0.2272±0.2848	0.9131±0.6744	0.2952±0.2009	0.5387±0.0908	0.1611±0.0761	0.6333±0.0739	0.2226±0.0332
Therm.	**0.6827±0.0170**	0.1654±0.0405	1.5188±0.5910	0.2180±0.0302	**0.7360±0.0087**	0.1954±0.0372	1.9680±0.4174	0.2124±0.0213
ATS	**0.6827±0.0170**	0.3346±0.0658	2.6951±1.5194	0.3195±0.0403	**0.7360±0.0087**	0.2432±0.0119	3.1290±0.1205	0.2430±0.0082
CoCHI	0.6700±0.0040	**0.0523±0.0193**	**0.5069±0.0090**	**0.1730±0.0035**	0.7307±0.0186	**0.0961±0.0169**	**0.5066±0.0121**	**0.1700±0.0048**
Mistral-7B								
TS	0.5100±0.0100	0.0673±0.0383	0.6016±0.0130	0.2076±0.0033	**0.6707±0.0064**	0.1489±0.0077	0.5720±0.0115	**0.1987±0.0089**
ITI	0.3567±0.0996	0.3583±0.2315	1.0117±0.3433	0.3684±0.1309	0.5193±0.0574	0.1522±0.0597	0.6011±0.0402	0.2089±0.0195
Therm.	0.5100±0.0100	0.2388±0.0647	2.0444±0.3463	0.3042±0.0360	**0.6707±0.0064**	0.2214±0.0398	2.0876±0.3632	0.2511±0.0222
ATS	0.5100±0.0100	0.4317±0.0040	4.2157±0.0519	0.4266±0.0012	**0.6707±0.0064**	0.2904±0.0064	3.0443±0.0280	0.2896±0.0046
CoCHI	**0.5120±0.0040**	**0.0316±0.0073**	**0.5912±0.0029**	**0.2047±0.0014**	0.6700±0.0139	**0.1432±0.0108**	**0.5634±0.0068**	0.2028±0.0035

(NLL), which evaluates the quality of predicted probability distributions by penalising overconfident incorrect predictions; and the **Brier Score (BS)**, which computes the mean squared error between predicted probabilities and ground-truth labels. Lower values of ECE, NLL, and Brier Score indicate better calibration. Together, these metrics provide a comprehensive assessment of both predictive correctness and the reliability of confidence estimates.

5.3 Main Results

The results in Table 1 show that **CoCHI** consistently improves calibration across all datasets and models while maintaining competitive accuracy. On datasets such as *MMLU* and *MedMCQA*, CoCHI achieves clear reductions in ECE together with improved NLL and Brier Score. Similar trends are observed on *OpenBookQA* and *MedQA-USMLE*, where CoCHI attains the lowest or near-lowest calibration errors without sacrificing predictive performance. In several cases, CoCHI also improves accuracy, particularly on LLaMA-3.1-8B and Mistral-7B.

In contrast, existing methods show limited or unstable behaviour across datasets. Post-hoc methods such as TS and ATS provide only marginal improvements and can severely degrade calibration in some settings, while Thermometer often increases NLL, indicating unreliable probability estimates. ITI also exhibits high variance across models and datasets. The advantage of CoCHI is especially evident on Mistral-7B, where it consistently achieves lower ECE together with strong NLL and Brier Score performance. Overall, these results suggest that intervening on targeted causal attention heads

provides more effective and stable calibration than conventional output-level adjustment methods. Additional results are provided in Appendix B.1.

5.4 Ablation Study

The ablation results in Table 2 highlight the importance of structured, causally-informed masking. Naive strategies, such as masking only wrong-high facilitating heads or applying a broad canonical-union mask, severely degrade both accuracy and calibration (accuracy 0.2044/0.2189; ECE 0.4745/0.3850), suggesting that indiscriminate head removal disrupts essential model functionality. More targeted variants also fail to achieve a satisfactory trade-off: masking correct-high interfering heads lowers ECE to 0.0694 but substantially reduces accuracy to 0.4267, while random size-matched masking performs poorly (accuracy 0.2111, ECE 0.4330). In contrast, **CoCHI** achieves the best balance, with accuracy 0.6120 and ECE 0.0490. These results show that effective intervention depends not only on identifying problematic heads, but also on selecting a compact set of high-impact components through canonical-union design and top-k ranking.

Table 2: Ablation study of different masking strategies on MMLU.

Mask Rule	Masked Acc ↑	Masked ECE ↓
Wrong-high facilitating only	0.2044±0.0038	0.4745±0.1477
Correct-high non-facilitating only	0.2389±0.0398	0.1518±0.0474
Correct-high interfering only	0.4267±0.0462	0.0694±0.0488
Canonical union (broad)	0.2189±0.0271	0.3850±0.1816
Random size-matched	0.2111±0.0077	0.4330±0.2736
CoCHI	**0.6120±0.0626**	**0.0490±0.0089**

5.5 Sensitivity Analysis

Table 3 presents the sensitivity analysis of the confidence threshold τ_c. Overall, the proposed method remains relatively stable across different threshold settings, indicating that its performance is not highly sensitive to the choice of τ_c. Among all configurations, $\tau_c = 0.5$ achieves the best overall trade-off, yielding the lowest ECE, while also obtaining competitive Accuracy, NLL, and Brier Score values. In contrast, very high thresholds (e.g., $\tau_c = 0.7$ and 0.9) lead to noticeable degradation in calibration quality, reflected by increased ECE and NLL. Lower thresholds such as 0.2 provide reasonable calibration performance but slightly reduce predictive accuracy. These results suggest that moderate confidence thresholds are more effective for balancing classification performance and uncertainty calibration.

Table 3: Sensitivity analysis of the confidence threshold τ_c on MMLU.

τ_c	Accuracy ↑	ECE ↓	NLL ↓	BS ↓
0.2	0.5933 ±0.0217	0.0555 ±0.0096	0.5385 ±0.0075	0.1808 ±0.0032
0.3	0.6060 ±0.0059	0.0643 ±0.0102	0.5319 ±0.0043	0.1781 ±0.0017
0.4	0.5860 ±0.0118	0.0665 ±0.0108	0.5276 ±0.0058	0.1761 ±0.0025
0.5	0.6120±0.0626	0.0490±0.0089	0.5168±0.0236	0.1743 ±0.0008
0.7	0.5873 ±0.0377	0.0705 ±0.0099	0.5548 ±0.0212	0.1873 ±0.0083
0.9	0.5933 ±0.0081	0.0696 ±0.0277	0.5255 ±0.0059	0.1748 ±0.0022

5.6 Efficiency Analysis

Table 4 presents an efficiency comparison across calibration methods, highlighting a clear trade-off between calibration cost and inference overhead. CoCHI requires GPU computation during a one-time calibration phase (69.9s), but produces an extremely compact artifact (40 B) and introduces zero inference overhead, as the intervention is implemented by directly zeroing selected attention heads. In contrast, methods such as ITI and Thermometer incur additional computation at every forward pass (e.g., $O(k \cdot d)$ and $O(H \cdot C)$, respectively), leading to ongoing runtime costs and significantly larger artifacts (up to 1.0 MB). Lightweight post-hoc approaches such as TS and ATS have minimal calibration cost (0.5s and 25.8s) and negligible inference overhead, but do not modify the model's internal behaviour and therefore lack structural impact.

Table 4: Efficiency comparison of calibration methods using 40 calibration samples on MMLU.

Method	GPU	Cal. time (s) ↓	Artifact size ↓	Inf.	Prunable
TS	No	0.5	4 B	Negligible	No
ATS	No	25.8	16 B	Negligible	No
ITI	Yes	18.4	16 KB	$O(k \cdot d)$	No
Therm.	No	19.4	1.0 MB	$O(H \cdot C)$	No
CoCHI	Yes	**69.9**	**40 B**	**None**	**Yes**

A key advantage of CoCHI is its *prunability*: the masked heads can be permanently removed, yielding a smaller model with no runtime wrapper. Overall, CoCHI shifts the computational burden from continuous inference to a one-time structural calibration step, offering a more deployment-efficient solution for improving model reliability. More results are available in Appendix B.2.

5.7 Case Study

Figure 2 presents a per-class calibration comparison between the unmasked baseline and our method across five datasets, focusing on correct predictions with low confidence (left) and wrong predictions with high confidence (right). While our method is primarily designed to address `wrong_high` errors, we observe consistent improvements in both regimes. In the `correct_low` subset, our method reduces ECE across all datasets (e.g., 0.10 on MMLU and 0.09 on MedMCQA), indicating improved confidence alignment for underconfident correct predictions. In the more challenging `wrong_high` subset, our method also consistently lowers ECE, albeit with smaller margins. This suggests that suppressing causally relevant heads not only mitigates overconfident errors but also improves overall confidence stability.

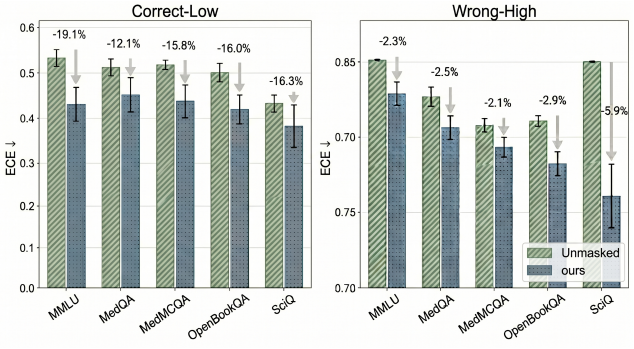


Figure 2: ECE comparison between the unmasked model and CoCHI on `correct_low` and `wrong_high` subsets.

6 Conclusion

In this work, we study overconfident errors in large language models and argue that miscalibration should be addressed not only at the output level, but also through the internal mechanisms that give rise to unreliable confidence estimates. We propose CoCHI, a class-oriented causal head intervention framework that decomposes predictions by correctness and confidence, identifies attention heads associated with distinct calibration behaviours, and performs targeted inference-time masking without updating model parameters. Across multiple LLMs and five multiple-choice QA benchmarks, CoCHI consistently improves calibration metrics while maintaining competitive predictive accuracy. The ablation and sensitivity analyses further show that reliable calibration depends on structured causal selection rather than indiscriminate head removal. Moreover, while CoCHI primarily targets `wrong_high` errors, it also yields improvements in `correct_low` cases. These findings support a mechanism-aware view of calibration, linking model reliability to identifiable internal structures rather than treating it solely as a post-hoc output adjustment problem.

Limitations. This work focuses on multiple-choice question answering, where confidence can be naturally defined over a fixed answer space. Extending CoCHI to open-ended generation, long-form reasoning, and multi-turn interaction requires further investigation, especially when correctness and confidence are harder to define. In addition, our current intervention operates at the attention-head level, which provides an efficient and interpretable unit for masking, but may not capture all sources of miscalibration arising from MLP layers, residual streams, or cross-layer interactions.

Acknowledgements

We acknowledge the RMIT Advanced Compute Environment (RACE) for providing the high-performance computing resources and infrastructure that supported this research. We also thank Patrick Taylor for his valuable technical support and assistance.

References

[1] Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM: iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6aebba00fff5b6de7b488e496f80edd7-Abstract-Conference.html.

[2] Mark Steyvers, Heliodoro Tejeda Lemus, Aakriti Kumar, Catarina G. Belém, Sheer Karny, Xinyue Hu, Lukas William Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025. doi: 10.1038/S42256-024-00976-7. URL https://doi.org/10.1038/s42256-024-00976-7.

[3] Jing Ren, Jiapeng Du, Bowen Li, Ziqi Xu, Xin Zheng, Hong Jia, Suyu Ma, Xiwei Xu, and Feng Xia. Spiking graph predictive coding for reliable OOD generalization. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 9056–9067. ACM, 2026. doi: 10.1145/3774904.3793044. URL https://doi.org/10.1145/3774904.3793044.

[4] Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, and Xiuzhen Zhang. Hallucinate less by thinking more: Aspect-based causal abstention for large language models. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 32555–32563. AAAI Press, 2026. doi: 10.1609/AAAI.V40I38.40532. URL https://doi.org/10.1609/aaai.v40i38.40532.

[5] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[6] DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

[7] Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

[8] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24128–24156. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1176. URL https://doi.org/10.18653/v1/2025.acl-long.1176.

[9] Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. Halogen: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1402–1425. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.71. URL https://doi.org/10.18653/v1/2025.acl-long.71.

[10] Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal LLM hallucination via bottom-up holistic reasoning. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 8460–8468. AAAI Press, 2025. doi: 10.1609/AAAI.V39I8.32913. URL https://doi.org/10.1609/aaai.v39i8.32913.

[11] Akira Okutomi. Stable but miscalibrated: A kantian view on overconfidence from filters to large language models. *CoRR*, abs/2510.14925, 2025. doi: 10.48550/ARXIV.2510.14925. URL https://doi.org/10.48550/arXiv.2510.14925.

[12] Yaqi Yang, Ziqi Xu, Jie Li, Chenglong Ma, Jeffrey Chan, Mark Sanderson, Xin Zheng, and Yongli Ren. Mitigating bias in large language model based question answering through causal front door prompting. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, pages 2208–2218. ACM, 2026. doi: 10.1145/3805712.3809679. URL https://doi.org/10.1145/3805712.3809679.

[13] Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15440–15459. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.824. URL https://doi.org/10.18653/v1/2024.acl-long.824.

[14] Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. Cats: Calibrated test-time scaling for efficient LLM reasoning. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9f9ecbf4062842df17ec3f4ea3ad7f54-Abstract-Conference.html.

[15] Jing Ren, Bowen Li, Ziqi Xu, Xikun Zhang, Haytham M. Fayek, and Xiaodong Li. When to trust: A causality-aware calibration framework for accurate knowledge graph retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 3859–3869. ACM, 2026. doi: 10.1145/3774904.3792358. URL https://doi.org/10.1145/3774904.3792358.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

[17] Andrew Nam, Henry Conklin, Yukang Yang, Tom Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a7f530e11fa19e9551b7a51dbd0f336f-Abstract-Conference.html.

[18] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

[19] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

[20] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. Linking model intervention to causal interpretation in model explanation. *Pattern Recognit.*, 173:112814, 2026. doi: 10.1016/J.PATCOG.2025.112814. URL https://doi.org/10.1016/j.patcog.2025.112814.

[21] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Ke Wang. Disentangled representation for causal mediation analysis. In *Thirty-Seventh AAAI Conference on Artificial Intelligence,*

AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 10666–10674. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26266. URL https://doi.org/10.1609/aaai.v37i9.26266.

[22] David John Cameron Mackay. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

[23] Alex Graves. Practical variational inference for neural networks. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2348–2356, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/7eb3c8be3d411e8ebfab08eba5f49632-Abstract.html.

[24] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1613–1622. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/blundell15.html.

[25] José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1861–1869. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/hernandez-lobatoc15.html.

[26] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 681–688. Omnipress, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[28] Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, and Benjamin Roth. Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3740–3761. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.188. URL https://doi.org/10.18653/v1/2025.acl-long.188.

[29] Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18128–18138. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1007. URL https://doi.org/10.18653/v1/2024.emnlp-main.1007.

[30] Maohao Shen, Subhro Das, Kristjan H. Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 44687–44711. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/shen24c.html.

[31] Xin Liu, Muhammad Khalifa, and Lu Wang. Litcab: Lightweight language model calibration over short- and long-form responses. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jH67LHVOIO.

[32] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.330. URL https://doi.org/10.18653/v1/2023.emnlp-main.330.

[33] Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12073–12086. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.652. URL https://doi.org/10.18653/v1/2024.acl-long.652.

[34] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5506–5524. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.335. URL https://doi.org/10.18653/v1/2023.emnlp-main.335.

[35] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/TACL_A_00494. URL https://doi.org/10.1162/tacl_a_00494.

[36] Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

[37] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *CoRR*, abs/2505.20045, 2025. doi: 10.48550/ARXIV.2505.20045. URL https://doi.org/10.48550/arXiv.2505.20045.

[38] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[39] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

[40] Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=41HlN8XYM5.

[41] Judea Pearl. *Causality*. Cambridge university press, 2009.

[42] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

[43] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *Foundations and Trends in Optimization*, 9(1-2):1–247, 2025. doi: 10.1561/2400000052. URL https://doi.org/10.1561/2400000052.

[44] Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4288–4295. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.318. URL https://doi.org/10.18653/v1/2022.naacl-main.318.

[45] Xinshu Li, Shiyi Yang, Ziqi Xu, Feng Xia, Quan Z. Sheng, and Lina Yao. Causal abstraction learning for multi-modal grounded planning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD 2026, Jeju Island, Korea, August 9-13, 2026*, pages 2790–2801. ACM, 2026. doi: 10.1145/3770855.3817995. URL https://doi.org/10.1145/3770855.3817995.

[46] Tianyi Zhao, Yinhan He, Wendy Zheng, Yujie Zhang, and Chen Chen. Wired for overconfidence: A mechanistic perspective on inflated verbalized confidence in llms. *CoRR*, abs/2604.01457, 2026. doi: 10.48550/ARXIV.2604.01457. URL https://doi.org/10.48550/arXiv.2604.01457.

[47] Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. Calibration across layers: Understanding calibration evolution in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 14686–14714. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.742. URL https://doi.org/10.18653/v1/2025.emnlp-main.742.

[48] Gemma Team. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL https://doi.org/10.48550/arXiv.2408.00118.

[49] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

[50] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[51] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

[52] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1260. URL https://doi.org/10.18653/v1/d18-1260.

[53] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

[54] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 94–106. Association for Computational Linguistics, 2017. doi: 10.18653/V1/W17-4413. URL https://doi.org/10.18653/v1/w17-4413.

A Dataset Preparation and Experimental Pipeline

This section details the data preparation and experimental pipeline to ensure reproducibility.

A.1 Datasets

We use five publicly available multiple-choice question answering datasets (as shown in Table 5) from HuggingFace, covering general knowledge, science, and medical domains. All datasets are formatted as four-option multiple-choice questions with a single correct answer.

Table 5: Statistical summary of the datasets.

Dataset	Domain	# Questions
MMLU	Multi-domain knowledge	14,042
MedQA-USMLE	Medical exam	11,451
OpenBookQA	Science	∼4,200
SciQ	Science	13,679
MedMCQA	Medical exam	187,005

A.2 Data Cleaning and Standardisation

All datasets are converted into a unified format containing: question ID, dataset source, question text, four answer choices (A–D), and the correct label. We remove samples that: (i) do not contain exactly four options, or (ii) have missing or invalid answer labels.

SciQ Adjustment. In SciQ, the correct answer appears in a fixed position. To avoid positional bias, we cyclically rotate answer options to ensure a uniform distribution.

A.3 Prompt Formatting

All inputs follow a unified zero-shot prompt:

> Answer the following multiple choice question with a single letter: A, B, C, or D.
> Question: [question text]
> A. [choice 1]
> B. [choice 2]
> C. [choice 3]
> D. [choice 4]
> Answer:

No demonstrations or additional context are provided.

A.4 Training Example Selection

From the fit pool, we sample a balanced calibration subset:

- up to 10 examples per class,
- maximum of 40 samples per dataset.

This simulates a low-resource calibration setting.

A.5 Top-k Head Selection

We determine the number of masked heads k via grid search over $k \in \{1, \ldots, 20\}$. Selection is based on minimising Expected Calibration Error (ECE), subject to an accuracy drop constraint of at most 1% relative to the unmodified model. A single k is selected per model and applied across all datasets to capture model-level calibration properties. The selected values are $k = 10$ for LLaMA-3.1-8B, and $k = 8$ for Gemma-2-9B and Mistral-7B.

Table 6: Results on SciQ. Best results are highlighted in bold.

Method	Acc↑	ECE↓	NLL↓	BS↓
		Gemma-2-9B		
TS	0.9173±0.0012	0.0579±0.0184	0.4699±0.4226	0.0641±0.0094
ITI	0.9253±0.0083	0.1027±0.0091	0.2337±0.0076	0.0617±0.0025
Therm.	0.9223±0.0021	0.0609±0.0066	1.0836±0.1824	0.0599±0.0068
ATS	0.9173±0.0012	0.0607±0.0067	1.1578±0.1183	0.0602±0.0065
CoCHI	**0.9353±0.1069**	**0.0563±0.0199**	**0.2309±0.1276**	**0.0548±0.0472**
		LLaMA-3.1-8B		
TS	0.8790±0.0210	0.0354±0.0109	0.2377±0.0145	0.0721±0.0045
ITI	0.8647±0.0311	0.0359±0.0176	0.3032±0.0329	0.0921±0.0139
Therm.	0.8920±0.0131	0.0738±0.0112	0.9514±0.2711	0.0807±0.0065
ATS	0.8790±0.0210	0.0932±0.0219	1.4970±0.2712	0.0928±0.0195
CoCHI	**0.9020±0.0020**	**0.0330±0.0010**	**0.2266±0.0012**	**0.0676±0.0004**
		Mistral-7B		
TS	0.8667±0.0150	0.0764±0.0048	0.3057±0.0402	0.0944±0.0045
ITI	0.5720±0.1512	0.2095±0.1429	0.6363±0.1672	0.2249±0.0680
Therm.	0.8467±0.0310	0.0986±0.0084	1.4162±0.0701	0.0978±0.0073
ATS	0.8667±0.0150	0.1056±0.0139	1.6920±0.2413	0.1049±0.0134
CoCHI	**0.8773±0.0046**	**0.0730±0.0116**	**0.3030±0.0144**	**0.0903±0.0046**

B Additional Results

B.1 Main results

Table 6 confirms the effectiveness of our method on SciQ, demonstrating consistent improvements in calibration across all models while maintaining strong predictive performance. On Gemma-2-9B, our method achieves the best overall calibration, reducing ECE and improving both NLL and Brier Score , while matching the highest accuracy. On LLaMA-3.1-8B, the gains are even more pronounced: our method simultaneously achieves the best performance across all four metrics, improving accuracy, lowering ECE, and reducing both NLL and Brier Score. Similar trends are observed on Mistral-7B, where our method achieves the best results across all metrics, outperforming the strongest baseline (ATS) in both calibration and accuracy. These results indicate that the proposed intervention consistently improves confidence alignment without sacrificing predictive quality.

In contrast, baseline methods exhibit unstable and inconsistent behaviour. ITI achieves competitive accuracy on Gemma-2-9B but significantly worsens calibration, while on Mistral-7B it severely degrades both accuracy and calibration. Post-hoc methods such as TS and ATS provide moderate or inconsistent calibration improvements but often fail to jointly optimise NLL and Brier Score, with ATS in particular showing substantial degradation in probabilistic metrics (e.g., NLL exceeding 1.0 across models). Thermometer similarly leads to large increases in NLL, indicating unreliable probability estimates. Overall, these results reinforce our central claim: calibration is better improved

Table 7: Calibration time (s; mean ± std) of CoCHI across three model families, averaged over 15 runs per model. Costs include class-conditioned forward passes and exclude model loading, and scale with model size while remaining practical.

Model	Cal. time (s) ↓	Min	Max	Peak GPU (GiB)
Mistral-7B-v0.1	39±21	21	80	28.1
Llama-3.1-8B	58±20	41	93	30.4
Gemma-2-9B	88±23	72	159	46.2

by targeting internal causal mechanisms rather than relying on output-level adjustments, leading to more stable and reliable performance across diverse models.

B.2 Efficiency Analysis

Table 7 shows that the calibration cost of our method scales predictably with model size while remaining practical. The average calibration time increases from 39 s for Mistral-7B to 58 s for LLaMA-3.1-8B and 88 s for Gemma-2-9B, reflecting the expected growth in forward-pass complexity, yet all models complete calibration within roughly 1–1.5 minutes. The variance is moderate (std $\approx$ 20–23 s), indicating stable computational behaviour. Peak GPU memory ranges from 28.1 GiB to 46.2 GiB, increasing with model size but remaining within the capacity of standard high-end GPUs. Notably, these costs include multiple class-conditioned forward passes, which are essential to our method, yet remain comparable to inference-time intervention baselines. Overall, the results indicate that the calibration procedure introduces only moderate overhead, scales with model size, and is practical for real-world deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the contributions and scope of the paper, including the proposed method, its motivation, and empirical validation across multiple benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the key limitations in the final section. First, our method relies on a small calibration set, and its effectiveness may vary under distribution shifts. Second, the intervention is limited to attention head-level manipulation and does not consider other components such as MLP layers or residual streams. Third, experiments are conducted on multiple-choice QA tasks, and further validation is required for open-ended generation and multi-turn reasoning.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not focus on formal theoretical guarantees or proofs, but instead provides a mechanism-driven empirical framework supported by extensive experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide sufficient details to reproduce the experiments, including datasets, preprocessing steps, data splits, model configurations, calibration procedures, and evaluation metrics. Additional details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide detailed implementation descriptions and experimental protocols to ensure reproducibility. Implementation details and code are provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental settings are clearly specified, including dataset preparation, calibration protocol, hyperparameters (e.g., τ_c, k), and evaluation procedures. Additional implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results are reported as mean $\pm$ standard deviation over three random seeds, capturing variability due to stochastic data sampling and model inference.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about computational resources, including GPU usage, memory consumption, and runtime for both calibration and inference stages.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It uses publicly available datasets and models and does not involve sensitive or restricted data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: On the positive side, our method improves calibration and reduces over-confident errors, enhancing reliability in high-stakes applications. On the negative side, improved calibration may increase user trust, potentially leading to over-reliance if users do not critically evaluate model outputs. We emphasise responsible deployment and user awareness.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not introduce new models or datasets with high risk of misuse. It relies on publicly available resources.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and models are properly cited and used in accordance with their respective licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or models are introduced. Implementation details and code are provided in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve human subjects or crowdsourcing.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve human subjects and does not require IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM assistance was limited to writing, editing, or formatting, and did not affect the core method, experiments, scientific rigor, or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.