Calibration Across Layers: Understanding Calibration Evolution in LLMs

Abhinav Joshi **Areeb Ahmad** **Ashutosh Modi**
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur (IIT Kanpur)
{ajoshi,areeb,ashutoshm}@cse.iitk.ac.in

Abstract

Large Language Models (LLMs) have demonstrated inherent calibration capabilities, where predicted probabilities align well with correctness, despite prior findings that deep neural networks are often overconfident. Recent studies have linked this behavior to specific components in the final layer, such as entropy neurons and the unembedding matrix's null space. In this work, we provide a complementary perspective by investigating how calibration evolves throughout the network's depth. Analyzing multiple open-weight models on the MMLU benchmark, we uncover a distinct *confidence correction phase* in the upper/later layers, where model confidence is actively recalibrated after decision certainty has been reached. Furthermore, we identify a low-dimensional *calibration direction* in the residual stream whose perturbation significantly improves calibration metrics (ECE and MCE) without harming accuracy. Our findings suggest that calibration is a distributed phenomenon, shaped throughout the network's forward pass, not just in its final projection, providing new insights into how confidence-regulating mechanisms operate within LLMs.

1 Introduction

Large Language Models (LLMs) have demonstrated strong generalization across a wide variety of tasks (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), yet it is challenging to understand how they manage and express uncertainty. Understanding the internal mechanisms by which LLMs regulate confidence is becoming increasingly important, especially as these models are deployed in settings where overconfidence can be costly. *Model calibration*, the alignment between a model's confidence and its accuracy, has emerged as a key axis for interpreting model behavior. Although deep neural networks have historically been found to be miscalibrated or overconfident (Guo et al., 2017), recent empirical studies suggest that LLMs exhibit surprisingly well-calibrated behavior across multiple tasks (Kadavath et al., 2022; Achiam et al., 2023; Plaut et al., 2024). This has sparked growing interest in uncovering architectural or representational mechanisms that support/cause calibration in LLMs.

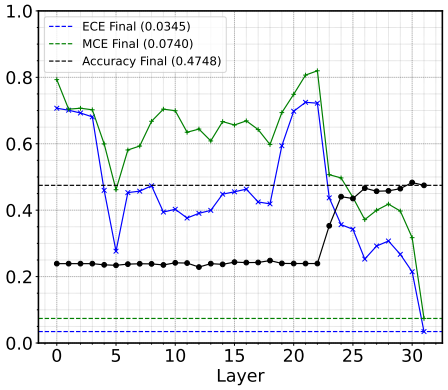


Figure 1: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the phi-2 model on the MMLU Humanities dataset. We observe that the model performance remains near-random (25%, 4-options) for initial layers and starts to rise from layer 22 and saturates at layer 26, with minor changes in the 26-31 layers. However, the ECE and MCE scores first rise (layers 25-28) and then decline (layers 28-31), highlighting the model calibration changing in the later layers.

Recent investigations have made notable progress in this direction. For instance, entropy neurons in the final layers have been shown to adjust the uncertainty of model predictions while minimally affecting the output distribution (Stolfo et al., 2024; Gurnee et al., 2024). These neurons modulate the entropy of the output distribution by operating in the null space of the unembedding matrix, effectively influencing confidence without altering accuracy. Such findings provide compelling evidence that calibration is an active, structured process, and that LLMs contain specialized components for reg-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14698–14726

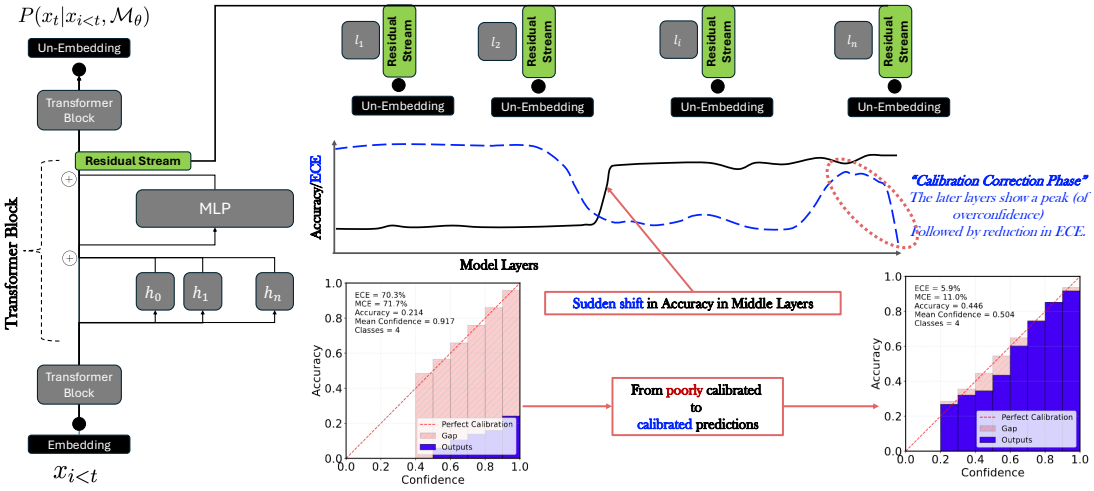


Figure 2: The figure provides an overview of the performed study. The Residual stream signals from each of the layers are projected back to the vocabulary space using the unembedding matrix. The obtained predictions are inspected for accuracy and model calibration scores (ECE/MCE). The models show a sudden peak arising in the middle layers, after which the performance remains saturated. Interestingly, the model goes into a calibration correction phase where the ECE first rises and then reduces, while maintaining the same accuracy, i.e. going from a poorly calibrated predictions to calibrated predictions (shown as reliability diagrams, also see Figure 12).

ulating confidence.

In this work, we build on this line of research by exploring whether such mechanisms are also present *within the intermediate layers* of the model. Specifically, we investigate how calibration evolves across the full depth of transformer-based language models. While prior work has illuminated the role of final-layer structures, the calibration dynamics of earlier layers remain less well understood. We examine multiple popular open-weight models (Phi-2 (Javaheripi et al., 2023), LLaMA-3 (Grattafiori et al., 2024), LlaMa-2 (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023)), on multiple real-world benchmarks (see App. C) with a special focus on the MMLU benchmark (Hendrycks et al., 2021), inspecting the confidence and prediction behavior across layers via the residual stream (Elhage et al., 2021). Our analysis reveals that calibration is not restricted to the model's final stages. Instead, we find that LLMs undergo a clear *confidence correction phase* in their later layers, where confidence is actively adjusted, even after prediction accuracy has stabilized. Further, we identify a low-dimensional direction in the residual stream that is consistently aligned with changes in model confidence. Perturbing this direction improves calibration metrics (ECE and MCE) without degrading accuracy, weakly suggesting the existence of a meaningful calibration subspace distributed across layers. In a nutshell, we make the following contributions:

- We provide a layerwise analysis of calibration dynamics in transformer-based LLMs, showing that confidence is not simply correlated with accuracy but evolves through a distinct *confidence correction phase*, where models become temporarily overconfident before self-adjusting in later layers. (see Figure 2 for an overview and Figure 1 for calibration changing across later layers.)
- We identify a "calibration direction" in the residual stream that governs confidence modulation, and demonstrate that small perturbations along this direction improve calibration metrics (ECE and MCE) without sacrificing accuracy. (see Figure 5 and App. Figure 21 for generalization across datasets)
- We provide a complementary perspective to existing work on final-layer calibration mechanisms by revealing distributed calibration behavior across the network's depth, especially in intermediate layers that have received limited attention in prior studies.

In summary, our findings contribute to a more complete picture of how calibration is implemented within LLMs. Rather than being an isolated property of the final output layer, calibration appears to be a dynamic and distributed process. We hope this perspective encourages further work toward interpretable and controllable confidence modulation in language models. Our code is publicly accessible at `https://github.com/Exploration-Lab/LLM-Calibration-Mechanism`.

2 Related Work

Understanding and regulating uncertainty in neural networks has long been a foundational challenge in machine learning. Early studies revealed that modern deep networks tend to be overconfident and poorly calibrated (Guo et al., 2017), prompting the development of theoretical frameworks for uncertainty estimation, including Bayesian approximations and dropout-based techniques (Gal, 2016). While much of this work focused on vision models or shallow classifiers, recent attention has shifted toward calibration in LLMs, where the stakes of miscalibrated predictions can be higher. Several recent studies have revealed that LLMs, despite their size and complexity, often exhibit surprisingly strong calibration properties. Kadavath et al. (2022) and Achiam et al. (2023) showed that LLM token-level probabilities correlate well with accuracy, suggesting an emergent form of self-knowledge. This idea has been extended by Yin et al. (2023) and Xiong et al. (2024), who further investigated how well LLMs can express and act on their uncertainty in downstream tasks. However, Kapoor et al. (2024) cautioned that such behavior may not generalize without explicit training signals, raising questions about when and how such calibration arises. A complementary line of work has sought to uncover the mechanisms underlying these behaviors. Notably, Stolfo et al. (2024) identified final-layer "confidence regulation neurons" that influence the entropy of the model's output distribution without significantly changing its predictions. Similarly, Cancedda (2024) explored how the spectral properties of the unembedding matrix contribute to calibration, emphasizing the importance of low-energy directions previously overlooked. Other studies like Sharma et al. (2024) demonstrated that a large fraction of the unembedding space is redundant and can be compressed without performance loss, which may affect how uncertainty is encoded. These findings suggest that LLMs use sophisticated mechanisms at their output layers to regulate confidence. However, most prior analysis have centered exclusively on final-layer phenomena, entropy neurons, attention patterns, and projection bottlenecks, while neglecting how confidence emerges and evolves across the depth of the network. Tools like the Logit Lens (nostalgebraist, 2020) and residual stream analyses (Elhage et al., 2021) have made it possible to study intermediate representations, but their use in the context of calibration remains limited.

In parallel, some recent work has explored prompting strategies and fine-tuning methods for improving confidence estimation in LLMs (Tian et al., 2023), while broader surveys (Geng et al., 2024; Gawlikowski et al., 2023) have documented a wide variety of calibration techniques, from temperature scaling to Bayesian ensembling. Yet, these methods often treat the model as a black box, providing little insight into the internal computations shaping confidence.

In contrast, our work provides a mechanistic, layer-wise perspective on calibration, complementing prior studies by tracking how uncertainty evolves throughout the forward pass. We identify a *confidence correction phase* in the later layers and a *calibration direction* in the residual stream, demonstrating that confidence is explicitly modulated across depth, not just at the output.

3 Background

In this section, we review essential background on transformer-based language modeling and model calibration. We focus on aspects most relevant to our study, which include token-level prediction in decoder-only transformers and how calibration metrics quantify model uncertainty.

Transformer-based Language Modeling: Transformer-based language models (LMs) are typically trained to predict the next token in a sequence, modeling the conditional probability distribution $P(x_t \mid x_1, \ldots, x_{t-1})$ over a vocabulary $\mathcal{V}$ (Modern language models commonly use vocabularies of size $|\mathcal{V}| \geq 50{,}000$ (Radford et al., 2019; Liu et al., 2019).) These models are implemented as deep neural networks parameterized by θ, denoted $\mathcal{M}_\theta$, and trained in an autoregressive fashion. Given a token sequence $x = [x_1, \ldots, x_{t-1}] \in \mathcal{V}^{t-1}$, the model outputs a vector of logits $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$, where each component corresponds to the unnormalized log-probability of a vocabulary token. Applying the softmax function to $\mathbf{z}_t$ yields the probability distribution over the next token. Internally, decoder-only models consist of a stack of transformer blocks $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_L}$, which process the input sequence via self-attention and feedforward layers. These blocks operate on and update a shared residual stream, with skip connections facilitating gradient flow and information propagation (Elhage et al., 2021). At each layer, token representations are refined

until the final hidden state is projected to the vocabulary space. We focus on the representation corresponding to the last input token x_{t-1}, as this token is typically responsible for generating the next-token prediction. The final residual vector for this token is first normalized by a LayerNorm module, then projected to the vocabulary space via a learned weight matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$, often referred to as the unembedding matrix (Elhage et al., 2021).

Layer Normalization (LayerNorm) (Ba et al., 2016) plays a critical role in stabilizing training and enhancing convergence in transformer models. Given an input vector $\mathbf{z}_t \in \mathbb{R}^{d_{\text{model}}}$, LayerNorm transforms it as:

$$\texttt{LayerNorm}(\mathbf{z}_t) = \frac{\mathbf{z}_t - \mu_{\mathbf{z}_t}}{\sqrt{\texttt{Var}(\mathbf{z}_t) + \epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$$

Here, $\mu_{\mathbf{z}_t}$ and $\texttt{Var}(\mathbf{z}_t)$ denote the mean and variance of the vector components, and $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_{\text{model}}}$ are learned scale and shift parameters. This operation standardizes the input and enables better gradient flow across layers.

Model Calibration: In machine learning models *calibration* refers to the alignment between a model's predicted confidence and the actual likelihood of being correct. A model is said to be well-calibrated, if across many predictions, tokens predicted with a given probability p are correct approximately p fraction of the time. Formally, for a given input prompt $x = [x_1, \ldots, x_{t-1}]$, let $y = x_t$ denote the true next token, and let $\hat{y} = \arg\max_{v \in \mathcal{V}} P_{\mathcal{M}_\theta}(v \mid x)$ be the model's predicted token. The model's *confidence* is given by $p = P_{\mathcal{M}_\theta}(\hat{y} \mid x)$, while its *accuracy* is defined as $a = \mathbb{I}(\hat{y} = y)$, where $\mathbb{I}$ is the indicator function. To assess calibration over a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, predictions are grouped into M bins $\{B_m\}_{m=1}^{M}$ based on their confidence scores (e.g., into intervals such as [0.0, 0.1), [0.1, 0.2), etc.). For each bin B_m, we define:

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p^{(i)},$$

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} a^{(i)},$$

where $p^{(i)}$ and $a^{(i)}$ are the confidence and accuracy for the i-th prediction. The *Expected Calibration Error* (ECE) aggregates the absolute difference between confidence and accuracy over bins, weighted by the number of samples per bin:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|.$$

A related metric, the *Maximum Calibration Error* (MCE), captures the worst-case bin-level deviation between accuracy and confidence:

$$\text{MCE} = \max_{m \in \{1,\ldots,M\}} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|.$$

Both metrics are minimized (i.e., equal to zero) when the model is perfectly calibrated. For a comprehensive treatment of calibration techniques in language models, please refer to Pavlovic (2025).

Reliability Diagrams: To visualize model calibration, we use *reliability diagrams* (Guo et al., 2017) (see App. Fig 12), which plot predicted confidence against empirical accuracy for different confidence intervals. In a perfectly calibrated model, the points lie on the diagonal $y = x$, indicating that predicted probabilities align with observed correctness. Deviations below the diagonal suggest overconfidence, while points above the diagonal indicate underconfidence. Reliability diagrams provide an intuitive, qualitative assessment of how model confidence corresponds to actual performance.

4 Experimental Setup

In this work, we study the calibration behavior of transformer-based language models (LLMs) by analyzing their performance on multiple-choice question answering (MCQA) tasks. The model predicts the next token conditioned on the context (input query). Our experimental setup focuses on assessing how the input structure, query framing, and model components impact calibration.

Task Setup: We evaluate LLMs on a task where a query is presented in the form of a multiple-choice question answering (MCQA) prompt. The input consists of two primary components: 1) *Query Information* (x_{query}): This contains the specific question or context associated with a dataset instance. 2) *Choice Set* (x_{options}): This includes the set of answer options provided for each instance. The number of options depends on the dataset (e.g., four in the case of MMLU). Specifically, given the input structure as:

$$P(x_t | x_{i<t}, \mathcal{M}_\theta) = P(x_t | x_{\text{query}}, x_{\text{options}}, x_\epsilon, \mathcal{M}_\theta)$$

where, x_{query} represents the query information (specific question or context), x_{options} denotes the set

of available answer choices (e.g., A, B, C, D), x_ϵ represents the set of prompt templates used for MCQA, $\mathcal{M}_\theta = \{f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_L}\}$ represents the language model with parameters θ across L layers. The model is expected to generate the correct answer choice token as the next token in the output sequence, which we evaluate for performance. Additionally, to ensure diversity and mitigate potential position biases in the answer choices, we randomize the order of the answer options, where:

$$x_{\text{options}} \leftarrow \{\text{A. } o_{\text{correct}}, \text{B. } o_{\text{wrong}}\}$$

This formulation allows us to test the model's ability to handle different question structures while monitoring its confidence across various layers of the transformer. In this work, we stick to reasoning captured using multiple-choice question answering (MCQA)-style prompts (Robinson et al., 2023; Joshi et al., 2024). The MCQA setup provides a principled and constrained setting for investigating the internal decision-making processes of LLMs (Wiegreffe et al., 2025; Joshi et al., 2025). Unlike open-ended or cloze-style generation, MCQA structures the task as a selection among discrete alternatives, thereby reducing confounding factors related to token frequency, length bias, and linguistic fluency (Brown et al., 2020). This format enables precise analysis of the transition from contextual representation to decision, making it well-suited for quantifying/measuring calibration changing across intermediate representations.

The use of structured MCQA format helps in consistent evaluation of model calibration across layers, as it requires the model to make a discrete decision among a fixed set of alternatives. Unlike open-ended generation or cloze-style completion, MCQA also has less ambiguity in output interpretation by constraining the prediction space, allowing us to more directly isolate and measure model confidence. Moreover, the use of standardized evaluation using metrics such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) can be established in a straightforward fashion, which is more difficult to apply in generative settings. Additionally, because the model's output is evaluated on a fixed set of tokens, MCQA avoids the stochasticity introduced by sampling strategies (e.g., temperature sampling or top-k decoding), which often confound confidence analysis in generation tasks. In contrast, open-ended generation introduces several challenges for layer-wise calibration analysis, i.e., the ambiguity of token-level correctness, the absence of well-defined calibration metrics for full sequences, and the non-determinism inherent in decoding strategies. For these reasons, we specifically adopt MCQA as a controlled and interpretable framework for understanding internal confidence dynamics and calibration behavior in large language models.

Datasets and Prompt Templates: For our experiments, we use multiple real-world datasets (see App. C for details) with a primary focus on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), which spans 57 diverse subjects across STEM, humanities, social sciences, and other fields. This benchmark is designed to evaluate general knowledge acquired during pretraining of language models. Each question is paired with four answer choices, and model performance is evaluated based on the correctly predicted choice. Notably, the nature of the answer choices varies across MMLU categories. In some categories (e.g., logical reasoning or high school computer science), the choices are relatively formulaic and repeat across multiple instances (e.g., "True"/"False"). In contrast, other categories (e.g., medical or legal domains) present unique, context-dependent options for each question. This variation introduces different levels of reasoning complexity and lexical diversity, making calibration analysis more nuanced. We include a representative prompt template in App. Figure 8. This diversity introduces varying degrees of lexical and semantic complexity, which we believe provides an overall generalization of the experimental findings regarding calibration that we further explore in our experiments.

Monitoring Layer Performance: To understand how calibration varies across layers, we take inspiration from the approach by Logit Lens (nostalgebraist, 2020), computing accuracy at different layers of the transformer. After each transformer block, we extract the residual stream representation z_t and project it onto the vocabulary space using the unembedding matrix $\mathbf{W}_U$ as follows:

$$\text{logits}(z_t) = \mathbf{W}_U \, \text{LayerNorm}(z_t)$$

We then compute the accuracy of the logits at each layer and track the changes in performance as the information propagates through the layers. This helps us pinpoint which layer's representations are most decisive for the model's predictions. Addi-

tionally, we calculate the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) at each layer to quantify the calibration at different stages of the model. These measurements provide insight into both representational quality and the internal emergence of confidence across depth (model layer internals). Also see App. A.1 for details on the residual stream computations.

Analyzing the Prediction (Unembedding and Confidence Dynamics): To further analyze the role of the unembedding matrix, we apply Singular Value Decomposition (SVD) to it. The unembedding matrix $\mathbf{W}_U$ is decomposed as:

$$\mathbf{W}_U = \mathbf{U}_U \Sigma_U \mathbf{V}_U^T$$

This decomposition helps us separate the projection of the residual stream onto the prediction space, allowing us to study the significance of different components of the matrix. We find that the singular values exhibit a consistent pattern, where the initial values are large, followed by a long tail with decreasing values. Recent research (Sharma et al., 2024) suggests that this decomposition can be used to improve model performance by pruning less significant components, but the last few singular values, especially those in the tail, play a crucial role in the model's predictions and calibration (Cancedda, 2024). (also see App. Figure 13 for singular values of Unembedding Matrix in Phi-2 and Llama-3-8B models showing null space)

Notes from Prior Work: Recent studies have revealed intriguing structural properties of transformer models, particularly in their final layers. For instance, the unembedding matrix $\mathbf{W}_U$ often exhibits a characteristic spectral pattern when decomposed via Singular Value Decomposition (SVD), i.e., a handful of large singular values followed by a long tail and a sharp drop in the final 5% of the spectrum. Sharma et al. (2024) shows that substantial portions of these component matrices can be pruned (via SVD) without hurting, sometimes even improving, model performance. In contrast, Cancedda (2024) argue that the final singular modes, often dismissed as unimportant, in fact carry signals critical to prediction accuracy. Complementing this, Stolfo et al. (2024) propose that the model deliberately shapes this low-rank null space to regulate prediction confidence, effectively influencing model calibration.

While these findings highlight how late-stage components influence model confidence and calibration, less is known about the evolution of calibration within the model, especially across intermediate layers. Our study addresses this gap by analyzing how uncertainty and confidence emerge and evolve throughout the model's depth. Specifically, we measure both predictive performance and calibration metrics (ECE, MCE) layer-wise to localize where in the transformer stack the model becomes "confident" and how reliably that confidence reflects correctness. This layerwise perspective allows us to identify where in the model, confidence stabilizes and to what extent it is calibrated across depth.

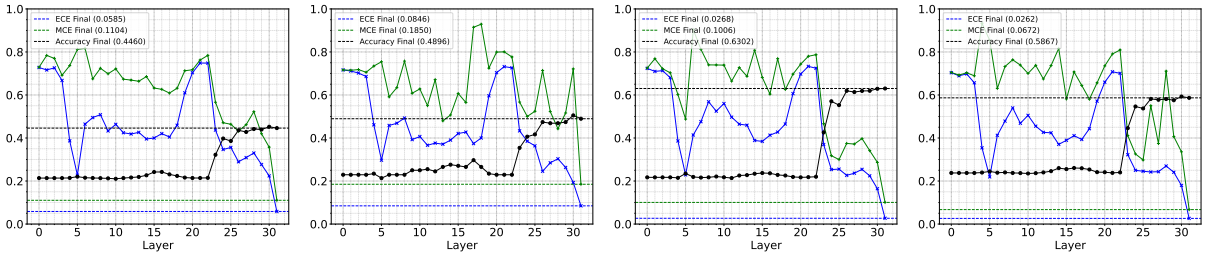
5 Results and Analysis

We present our findings in three stages: **1) How calibration evolves** across layers of a transformer, **2) The role of the unembedding** matrix's low-rank components, and **3) The discovery of a direction** in activation space that appears to regulate model calibration.

Layerwise Calibration Dynamics: We begin by analyzing how calibration and prediction performance vary across transformer layers in the phi-2 model. Each transformer block modifies the residual stream, which we project into the vocabulary space using the unembedding matrix $\mathbf{W}_U$ (see §4). At each layer, we compute predictive performance (via Accuracy), Expected Calibration Error (ECE), and Maximum Calibration Error (MCE).

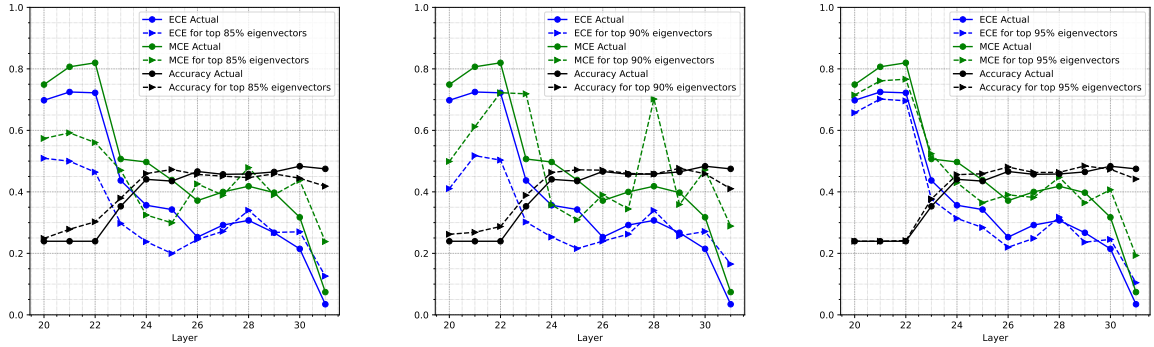
Across multiple datasets, a consistent trend emerges: accuracy begins to rise significantly from layer 22 and stabilizes by layer 26. However, the calibration behavior follows a different trajectory, ECE and MCE scores increase after layer 25, peaking around layer 28, before declining toward the final layers. This suggests that even after the model has become sufficiently accurate, it undergoes a phase of overconfidence before recalibrating its predictions. We refer to this as a "*confidence correction phase*" in the final layers.

To visualize this phenomenon, Figure 3 shows how reliability improves across the final layers. Reliability diagrams (App. Figure 12) further confirm this trend, revealing a widening and then narrowing gap between model confidence and accuracy. This denotes that the residual stream in the later layers is affected/modified in such a way that modulates the model calibration with no/minor change in the model performance (black line). The upper/later layers show the presence of ***calibration correction***



(a) MMLU STEM (b) MMLU Humanities (c) MMLU Social Science (d) MMLU Others

Figure 3: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Phi-2 model on the different datasets. We observe that the model performance starts to rise from layer 22 and saturates at layer 25/26, with minor changes in the 26-31 layers. However, the ECE and MCE scores first rise (layers 26-28) and then decline (layers 29-31), highlighting calibration changing in the later layers, with meager changes in the model performance. The upper/later layers show the presence of ***calibration correction phase***. Similar trends are found for other models (Llama-3-8B Figure 15, Mistral-7B Figure 16, and Llama-2-7B Figure 17).



(a) top 85% (b) top 90% (c) top 95%

Figure 4: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the phi-2 model computed by reconstructing the unembedding matrix, using only top-85%, top-90% and top-95% singular values in (a), (b), and (c), respectively. Overall, we observe the ECE scores with minor fluctuations, pointing towards a small contribution of lower singular values in model calibration.

phase. Similar trends are found for other models (Llama-3-8B Figure 15, Mistral-7B Figure 16, and Llama-2-7B Figure 17).

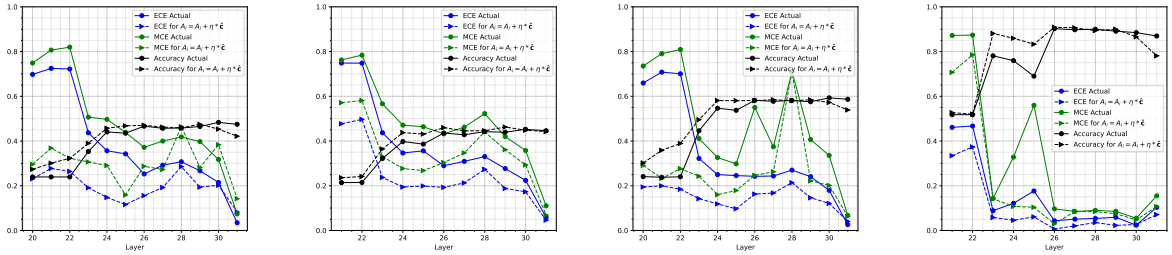
Effect of Unembedding Null Space: Prior work (Cancedda, 2024; Stolfo et al., 2024) suggests that the lower-rank (small singular value) components of the unembedding matrix may be involved in calibration, particularly via "entropy neurons" writing into its null space. To test this, we decompose the unembedding matrix $\mathbf{W}_U = \mathbf{U}_U \Sigma_U \mathbf{V}_U^T$ and reconstruct it by discarding the smallest 5%, 10%, and 15% of singular values:

$$\hat{\mathbf{W}}_U = \mathbf{U}_U \, \Sigma_U^{[:k]} \, \mathbf{V}_{U_{[:k]}}^T$$

where, $k \in \{85\%, 90\%, 95\%\}$. This intervention limits the influence of the null space on model outputs. Figure 4 shows the results using these truncated matrices. Accuracy remains largely unchanged, indicating that most predictive capacity lies in the dominant singular vectors. However, we observe fluctuations in calibration metrics, especially MCE, supporting the hypothesis that the null space plays a supporting role in calibration.

Interestingly, we find that these effects are in both directions (increasing and decreasing calibration) for middle layers, pointing towards no clear indication of ECE/MCE being increased when null space is removed. Null-space sensitivity appears across the upper layers of the network, suggesting that calibration is mediated by distributed subspaces throughout the model.

Discovery of a Calibration Subspace: Given the observations above, we ask: **Is there a specific direction in representation space that the model uses to recalibrate predictions?** From layer-wise activation traces, we identify significant represen-



(a) MMLU Humanities (b) MMLU STEM (c) MMLU Others (d) TruthfulQA

Figure 5: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the phi-2 model on the different datasets when adding the calibration direction to the residual stream. The added calibration direction to the residual stream helps shift the calibration scores to lower values, validating the impact of the calibration direction. Interestingly, the direction found using MMLU Humanities works well for other datasets like MMLU Others. (Due to space constraints, we move the results on other datasets to the App. Figure 21)

tation changes starting from layer 28, precisely when calibration begins to improve while accuracy plateaus (also see App. Figure 20). We use a simple strategy and define the *calibration direction* $\hat{\mathbf{c}}$ as the mean of the normalized differences between successive layer outputs in the final three layers:

$$\hat{\mathbf{c}} = \frac{1}{3}(\vec{c}_{29} + \vec{c}_{30} + \vec{c}_{31}), \quad \vec{c}_i = \frac{A_i - A_{i-1}}{\|A_i - A_{i-1}\|}$$

here, A_i denotes the residual stream output after layer i. This direction captures the internal shift the model undergoes to improve calibration, without affecting prediction correctness. We also verify that this calibration direction is *not* aligned with the null space (low singular values) of $\mathbf{W}_U$, as shown in Figure 6, indicating it arises from a distinct mechanism.

Modulating Calibration via Subspace Intervention: To test the functional role of the calibration direction, we perturb the residual stream along $\hat{\mathbf{c}}$ during inference:

$$A'_i = A_i + \eta\hat{\mathbf{c}}, \quad \eta > 0$$

Figure 5 shows that this small intervention leads to lower ECE and MCE scores without harming classification accuracy. Remarkably, the effect generalizes across datasets: the direction $\hat{\mathbf{c}}$ computed on the MMLU-Humanities split improves calibration on other MMLU subsets as well (including other datasets like TruthfulQA, Figure 21 (d)). This suggests the existence of a task-agnostic calibration subspace, i.e., distinct from the prediction subspace, that the model uses to regulate confidence.

6 Discussion

Our findings suggest that model calibration is not merely a byproduct of prediction accuracy but a distinct representational property shaped by specific components within the network. The emergence of calibration improvements in the final transformer layers, despite minimal accuracy gains, points to a dedicated phase in the model's forward pass where confidence is explicitly regulated. The fact that interventions in the residual stream using the found direction $\hat{\mathbf{c}}$ can improve calibration without affecting accuracy further supports the hypothesis that calibration resides in a separate, manipulable subspace. While the identified calibration directions show promising results within individual models and some datasets, they are not directly found across different architectures, and more investigations would be needed on similar lines (see §7 for more details). This lack of generalization (of the found calibration direction) suggests that the directions are partly model- and domain-specific, and motivates future work to identify more universal confidence-modulating features in these layers. We see this work as an initial step toward uncovering/understanding the mechanisms of confidence regulation in LLMs, with future research needed to evaluate generalization across generative tasks, domains, and training regimes. Additionally, the limited but non-negligible role of the unembedding null space reinforces insights from prior work (Cancedda, 2024; Stolfo et al., 2024), but our layerwise analysis shows that this effect is not isolated to the final projection step. Rather, it is distributed, suggesting a broader calibration mechanism involving intermediate representations.

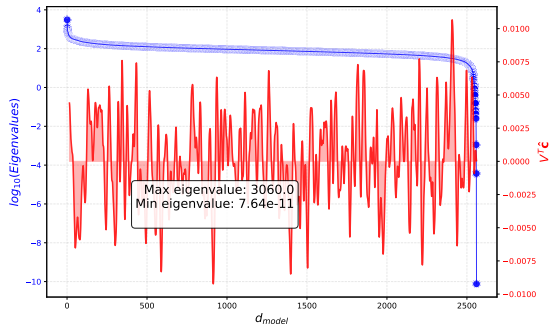


Figure 6: The figure shows the log of eigenvalues for each principal direction and its alignment with calibration direction $\hat{\mathbf{c}}$, denoting the writing not only in the null space (tail towards the right), but throughout.

Taken together, these results open up promising avenues for controllable calibration in LLMs through geometric interventions and motivate deeper exploration into how representations across layers encode not just "what" the model predicts but "how confident" it should be.

Several parallel lines of research have also investigated calibration dynamics across intermediate representations in deep neural networks, particularly in vision tasks using architectures like ResNets (He et al., 2015) and VGG (Simonyan and Zisserman, 2015). A recent work by Wang and Zhang (2024) identifies a calibration bottleneck in the middle layers of vision models, using linear probes trained on hidden representations. Their analysis reveals a U-shaped trend across layers, i.e., model predictions are more calibrated in the middle layers, with miscalibration increasing again toward the final layers, which they attribute to overcompression of information in the later layers. In contrast, we observe a different trend in the open-weight transformer-based language models that we experimented with, where the model's calibration is improved/regulated in the later/upper layers before the final predictions are made. A noteworthy difference between the other studies and our experimental setup is the use of logit lens-style probing, i.e., we directly project residual stream representations into the unembedding space to analyze the model's own prediction distribution, without training any additional classifiers, which we believe avoids introducing supervision and better reflects the model's internal confidence dynamics. These differences, in both architecture (transformers (Vaswani et al., 2017) vs. CNNs (He et al., 2015; Simonyan and Zisserman, 2015)) and methodology (unsupervised probing (nostalgebraist, 2020) vs. trained classifiers (Belinkov, 2022)), highlight the need for more domain- and architecture-specific analyses when studying calibration behavior in deep models. We believe our findings contribute to this growing literature by presenting an unsupervised, layerwise view of calibration dynamics in large-scale language models, which needs further investigation.

7 Conclusion and Future Directions

In this work, we conduct an investigation into how large transformer-based language models regulate their confidence across layers. Our analysis uncovers a structured three-phase calibration pattern: an initial *decision formation* phase, a subsequent phase of *overconfidence*, and a final *confidence correction phase* in the upper/later layers. Specifically, in the Phi-2 model, we observe that while accuracy plateaus beyond layer 24, calibration metrics such as ECE and MCE continue to fluctuate, first worsening, then sharply improving, revealing that model confidence is actively corrected even after predictions have stabilized. This phenomenon points to a dynamic internal mechanism modulating uncertainty across depth. We identify a low-dimensional *calibration direction* in the residual stream that weakly appears to underlie this correction phase. Perturbing this direction improves calibration across layers without degrading accuracy, suggesting that confidence regulation is not confined to the output layer but is instead distributed and tunable throughout the model's forward pass.

These findings extend prior work on entropy neurons by showing that confidence correction is a deliberate, multi-layer process, and that calibration can emerge progressively rather than being finalized at the prediction head. Our work provides some insights for probing this behavior. Practically, our results raise caution in relying on intermediate layers for downstream decision-making, as they may exhibit high accuracy but poor calibration. However, the ability to adjust confidence post-hoc via the calibration direction suggests new opportunities that need further investigations for early exiting and efficient inference that maintain reliability. Looking ahead, future work may explore how these layerwise calibration mechanisms arise during pretraining, whether similar correction dynamics generalize across model families and sizes, and how these phenomena can be explicitly modeled or optimized for applications requiring reliable confidence estimates.

Limitations

One of the primary limitations of this study is that this work restricts its analysis to a single-token classification setting, focusing on multiple-choice question answering tasks. While this setup may not reflect the full complexity of generative language modeling, it allows for a clean and controlled examination of model confidence and calibration using well-established metrics such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). This framing avoids the confounding effects introduced by autoregressive decoding, ensuring the interpretability of the results. Extending this analysis to multi-token generation remains an exciting direction for future work, where a deeper understanding of calibration over token sequences and temporal dynamics could be developed.

Our method for identifying the calibration direction in the residual stream is currently model- and dataset-specific. Although the discovered direction in Phi-2 leads to meaningful calibration improvements without degrading accuracy, it does not generalize to other models such as Mistral or LLaMA-2. This limitation highlights interesting differences in how confidence is regulated across architectures and invites further investigation into whether model-specific inductive biases or training schemes influence the emergence of such calibration structures. Similarly, the confidence correction phase we report is most evident in knowledge acquisition tasks like MMLU, where performance saturates in mid-to-late layers. In contrast, reasoning-based datasets exhibit gradually increasing accuracy, making it harder to isolate calibration behavior independently from prediction quality. We view this as an opportunity to refine analysis tools that can disentangle calibration from competence in such settings.

Finally, while our approach to defining the calibration direction is based on a simple difference between layerwise residuals, it lays the groundwork for richer strategies. More principled methods, such as those based on optimization, gradient sensitivity to ECE loss, or attribution techniques, could uncover more robust and generalizable directions. We see our current results as a strong proof of concept that invites further methodological development and broader application across tasks and architectures. We believe this line of investigation opens up a promising path toward mechanistically understanding calibration in the coming future.

Ethical Considerations

This paper aims to advance the field of machine learning, focusing specifically on model calibration and confidence regulation in transformer architectures. While we do not foresee any immediate ethical concerns arising from the research presented, it is essential to recognize that the broader implications of developing more reliable models include both positive and potentially negative societal consequences. Future applications of these techniques could affect areas such as fairness, bias mitigation, and decision-making in systems built upon LLMs, and it will be critical to assess and address these issues in subsequent research and applications.

Acknowledgments

We would like to thank the anonymous reviewers and the meta-reviewer for their insightful comments and suggestions. This research work was partially supported by the Research-I Foundation of the Department of CSE at IIT Kanpur.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Yarin Gal. 2016. Uncertainty in deep learning.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, 56(Suppl 1):1513–1589.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew

Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

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. 2023. Mistral 7b.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. 2024. COLD: Causal reasoning in closed daily activities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Abhinav Joshi, Areeb Ahmad, Divyaksh Shukla, and Ashutosh Modi. 2025. Towards quantifying commonsense reasoning with mechanistic insights. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9633–9660, Albuquerque, New Mexico. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19260–19268.

nostalgebraist. 2020. interpreting GPT: the logit lens, LessWrong. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. [Accessed 26-01-2025].

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Maja Pavlovic. 2025. Understanding model calibration – a gentle introduction and visual exploration of calibration and the expected calibration error (ece).

Benjamin Plaut, Khanh Nguyen, and Tu Trinh. 2024. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *CoRR*, abs/2402.13213.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2024. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. 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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. 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)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Deng-Bao Wang and Min-Ling Zhang. 2024. Calibration bottleneck: Over-compressed representations are less calibratable. In *Forty-first International Conference on Machine Learning*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Appendix

Table of Contents

List of Figures

A Additional Computation Details

A.1 Residual Stream Computations

The transformer architecture operates by reading from and writing to a residual stream across different layers (Elhage et al., 2021). Each layer applies various transformations (e.g., LayerNorm, Multi-Head Attention, FeedForward) to the residual stream. Mathematically, the operation at each transformer layer can be described as:

$$z_i = f_i \left(z_0 + \sum_{j=1}^{i-1} z_j \right)$$

where z_0 is the embedding vector from the embedding matrix $\mathbf{W}_E$, and f_i represents the function applied by the i-th transformer block. These operations modify the residual stream, which ultimately affects the prediction of the model. This residual stream formulation is central to mechanistic interpretability approaches, providing a lens into how information is incrementally composed between layers (Elhage et al., 2021).

B Prompt Templates

For all our experiments, we follow a standard prompt template. This section provides the details of the prompt templates used in our multiple-choice question answering (MCQA) evaluations of autoregressive open-weight language models (e.g., `LLaMA(-2)`, `Phi-2`). All prompts follow a unified format to ensure consistency across tasks and models. Figure 7 presents the general template used. Each prompt begins with an instruction directing the model to select the correct answer from a set of multiple-choice options. In few-shot or in-context learning settings, this instruction is optionally followed by a set of in-context examples. The question is prefaced by a task-specific description, and followed by a list of labeled answer choices (A, B, C, etc.). The final line of the prompt contains the prefix `Answer:`, which serves as the model's response cue. For evaluation, we extract the model's next-token prediction probabilities at this position over the answer option tokens (e.g., `A`, `B`), which we treat as the model's predicted distribution over choices.

Figure 8 illustrates a fully instantiated example using a question from the MMLU dataset. The correct answer is provided at the end of the prompt and underlined. The colored annotations in the figure denote fixed template components (in black) and variable elements drawn from the dataset (in orange and teal).

This templated format is applied uniformly across datasets and experimental configurations, enabling controlled comparisons of model behavior across domains and prompting setups.

C Dataset Details

We evaluate model calibration and confidence dynamics across a diverse set of NLP benchmarks, selected to cover a range of reasoning, linguistic, and factual understanding capabilities. Below, we detail each dataset used in our experiments:

MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021). MMLU is a comprehensive benchmark designed to test knowledge and problem-solving ability across 57 diverse subjects, spanning STEM, humanities, social sciences, law, medicine, and more. Each instance consists of a multiple-choice question with four answer options. Unlike many standard MCQA datasets, MMLU introduces lexical and semantic complexity by using dynamically varying answer choices, which increases the challenge for models to generalize and calibrate effectively across categories. This benchmark is widely adopted for evaluating pretraining quality in LLMs.

CoLA (Corpus of Linguistic Acceptability) (Warstadt et al., 2019): CoLA is a binary classification task that requires models to judge whether a given English sentence is grammatically acceptable. The dataset is drawn from linguistic publications and includes a broad spectrum of syntactic phenomena, making it a strong test of a model's grasp of grammatical rules. Performance is typically measured using Matthews Correlation Coefficient (MCC) and accuracy, providing a nuanced view of linguistic acceptability modeling.

COPA (Choice of Plausible Alternatives) (Gordon et al., 2012): COPA is a commonsense reasoning benchmark where the task is to select the most plausible cause or effect given a premise. Each instance presents two alternatives, and the model must determine which one best explains or results from the premise. The task challenges causal inference and contextual reasoning and is part of the SuperGLUE benchmark suite (Wang et al., 2020). Accuracy is used as the primary evaluation metric.

Rotten Tomatoes (Pang and Lee, 2005): This is a sentiment classification benchmark consisting

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: A generalised statement pertaining to the task -: question/statement
A. choice1
B. choice2
Answer: A
```

Figure 7: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models , (e.g., `llama(-3)`, `Phi-2`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the datasets which contains either a review or a statement or a question. The `teal text` is a template comment describing the task, which changes according to the dataset The next-token prediction probabilities of the option IDs at the `red text` is used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Mars has an atmosphere that is almost entirely carbon dioxide. Why isn't there a
strong greenhouse effect keeping the planet warm?
A: the atmosphere on Mars is too thin to trap a significant amount of heat
B: There actually is a strong greenhouse effect and Mars would be 35oC colder than it is now
without it.
C: Mars does not have enough internal heat to drive the greenhouse effect
D: the greenhouse effect requires an ozone layer which Mars does not have
Answer: A
```

Figure 8: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `Phi-2`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **MMLU dataset**. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

of short movie reviews labeled as positive or negative. Reviews are often limited to a sentence or short paragraph, focusing on fine-grained lexical and compositional sentiment cues. This dataset is widely used to benchmark sentiment understanding and general text classification performance in LLMs.

TruthfulQA (Lin et al., 2022): TruthfulQA evaluates a model's ability to provide factually correct and non-deceptive answers. It includes questions across multiple domains, such as science, health, and politics, carefully crafted to elicit plausible but incorrect responses from language models trained on internet-scale data. The benchmark serves as a diagnostic tool for hallucination and misinformation. Evaluations are conducted using truthfulness and informativeness scores, often involving human or model-based judgment. In this work, we specifically use a binary version of the dataset, framing an MCQA query for quantifying calibration.

Details of all the prompt templates and the MCQA formulations, for all the datasets used, are provided in our codebase.

D Extended Model Evaluations

D.1 Calibration Dynamics in Mistral and LLaMA-2 7B

To assess the generality of our findings, we evaluate calibration behavior in two additional open-weight transformer models: Mistral-7B and LLaMA-2-7B, along with LLaMA-3-8B. Across all three models, we observe an interesting pattern consistent with Phi-2: a *confidence correction phase* emerges in the later layers, characterized by stabilization of accuracy and a sharp improvement in calibration metrics such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

This suggests that the confidence correction behavior is not unique to Phi-2 but may reflect a broader inductive bias of transformer-based language models, where model confidence is actively adjusted after the prediction has converged. We observe this phenomenon in both Mistral and LLaMA variants, although the sharpness and layerwise extent of the correction vary slightly across models and tasks.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Select the suitable option for the following statement -: enchanted with low-life
tragedy and liberally seasoned with emotional outbursts . . . what is sorely missing, however,
is the edge of wild, lunatic invention that we associate with cage's best acting .
A: Negative
B: Positive
Answer: A
```

Figure 9: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models , (e.g., `llama(-3)`, `Phi-2`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **Rotten Tomatoes dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Which of the following events (given as options A or B) is a more plausible effect
of the event -: 'The woman betrayed her friend.'?
A: Her friend sent her a greeting card.
B: Her friend cut off contact with her.
Answer: B
```

Figure 10: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models , (e.g., `llama(-3)`, `Phi-2`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **COPA dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

However, our interventional experiments using a learned *calibration direction*, a low-dimensional vector in the residual stream found to modulate confidence in Phi-2, did not generalize well to other models. Attempts to extract and apply similar directions in Mistral-7B, LLaMA-2-7B, and LLaMA-3-8B yielded inconsistent results and failed to produce consistent improvements in calibration metrics. This suggests that while the calibration phase itself may be general, the specific encoding of confidence control within the residual stream may vary significantly across architectures and training regimes.

These findings point toward a promising avenue for future research: understanding how architectural or training factors give rise to shared calibration dynamics, and why certain models encode more "steerable" calibration subspaces than others.

E Reliability Diagrams Across Layers

To visualize calibration quality across model depth, we plot layerwise reliability diagrams for Phi-2, Mistral-7B, and LLaMA-2-7B on representative tasks (refer to our codebase for all the reliability diagrams). In Phi-2, we observe a clear overconfidence pattern in middle layers (see Figure 12), followed by a significant correction in later layers, where predicted confidence aligns more closely with empirical accuracy. (also see ECE/MCE patterns in Mistral-7B Figure 16, LLaMA-2-7B Figure 17, LLaMA-3-8B Figure 15 Phi-2 Figure 3, all showing a common ***calibration correction phase*** in the later/upper layers of the model where the calibration error first increases and then decreases, keeping the prediction accuracy intact.)

F Dataset-Level Calibration Trends

Beyond MMLU, we examine calibration behavior across four additional datasets: CoLA, COPA, Rotten Tomatoes, and TruthfulQA. For each dataset, we compute accuracy, ECE, and MCE across transformer layers. (see Figure 18 and Figure 19)

In contrast to MMLU, where accuracy stabilizes in later layers, these datasets exhibit gradual performance improvements throughout the model depth. This continuous gain limits our ability to isolate a calibration correction phase, as improvements in calibration may be confounded with accuracy

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Select the suitable option for the following statement -: The cat was bitten the
mouse.
A: Unacceptable
B: Acceptable
Answer: A
```

Figure 11: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models , (e.g., `llama(-3)`, `Phi-2`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **CoLA dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

Dataset	Task Type	# Samples	Avg. Prompt Length (tokens)	# Choices
MMLU (STEM)	Subject Knowledge (STEM)	3,018	149.09	4
MMLU (Humanities)	Subject Knowledge (Humanities)	4,705	535.10	4
MMLU (Social Sciences)	Subject Knowledge (Soc. Sci.)	3,077	116.35	4
MMLU (Other)	Subject Knowledge (Misc.)	3,242	163.32	4
CoLA	Grammatical Acceptability	1,043	41.83	2
COPA	Causal Commonsense Reasoning	1,000	34.89	2
Rotten Tomatoes	Sentiment Classification	1,066	115.52	2
TruthfulQA (Binary)	Factual Knowledge	790	159.21	2

Table 1: **Dataset Overview.** Summary of datasets used in our evaluation. We report the type of reasoning or knowledge tested, number of samples used, average prompt length (in tokens), and number of answer choices. MMLU categories are grouped based on domain.

refinement.

On CoLA and COPA, calibration remains noisy across layers, likely due to the small size and binary structure of the tasks. On Rotten Tomatoes, calibration improves steadily but without a sharp correction pattern. On TruthfulQA, we observe persistent underconfidence, with predicted probabilities often falling below empirical correctness, especially in earlier layers.

These observations highlight the complexity of measuring calibration when models have not yet saturated in performance.

G Interventional Experiments

To probe the functional role of calibration dynamics, we identify a low-dimensional "calibration direction" in Phi-2's residual stream using linear probes aligned with calibration error. We perform targeted interventions by adding scaled versions of this direction to intermediate residual representations.

Our experiments reveal that adding this direction at select layers (e.g., layers 22–32, Figure 5 and Figure 21) consistently reduces ECE and MCE without degrading accuracy. This confirms that the calibration signal is encoded in the residual stream and can be modulated independently of the model's final decision.

However, attempts to extract and apply similar directions in Mistral-7B and LLaMA-2-7B were unsuccessful. These models either lacked a distinct calibration direction or showed no calibration improvement upon intervention. This suggests that confidence regulation in Phi-2 is likely facilitated by an architectural or representational property not shared across models.

H Future Work and Discussion

In this section, we outline several promising directions to extend and deepen our current findings on calibration mechanisms in LLMs.

Generalizing Calibration Directions

Our current method for identifying a calibration direction in the residual stream, based on layerwise differences, provides useful insights but may lack generalizability across datasets and models. Future work can explore gradient-based approaches, such as computing the derivative of calibration metrics (e.g., ECE, MCE) with respect to residual activations. This could identify directions more causally linked to calibration. Furthermore, veri-

fying whether such directions align with low-rank or null-space structures in the unembedding matrix (as suggested by Cancedda, 2024) could provide a more principled mechanistic explanation.

Model Internals and Interpretability

The calibration direction could also serve as a tool for mechanistic interpretability. Specifically, it would be valuable to identify neurons or components (e.g., attention heads or MLP submodules) that significantly contribute to confidence modulation along this direction. Additionally, examining whether calibration behavior propagates through residual stream updates (indirect effects) may reveal compositional mechanisms behind calibration.

Robust Calibration Across Task Types

Our current findings are most robust in knowledge-centric datasets like MMLU, where prediction accuracy plateaus, enabling clear identification of calibration phases. In contrast, reasoning-focused datasets (e.g., COPA) exhibit gradual accuracy gains across layers, making it harder to isolate calibration-specific dynamics. Understanding how reasoning and knowledge acquisition tasks differentially affect confidence modulation is an important avenue for future research.

Calibration in In-Context Learning Settings

The current study is limited to zero-shot MCQA prompts. Extending this analysis to in-context learning (ICL) settings, such as few-shot or chain-of-thought prompting, may reveal how calibration dynamics change when more contextual supervision is available. However, this may prove more fruitful for reasoning tasks than factual knowledge tasks, where ICL often yields limited gains.

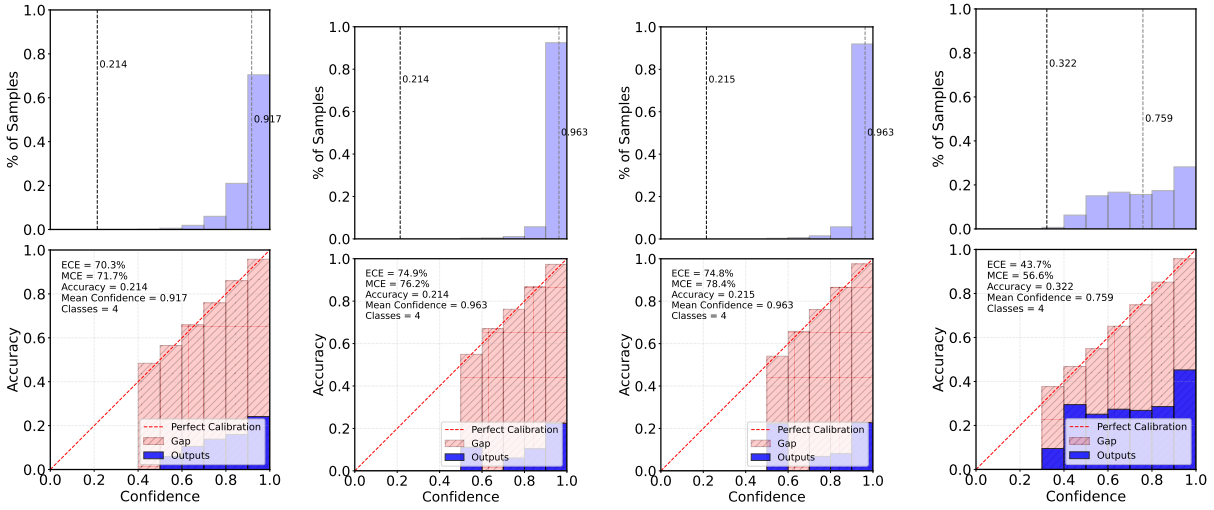
Cross-Model and Dataset Transferability

While the confidence correction phase is observed across various models (e.g., Phi-2, LLaMA-3-8B), the calibration direction identified in Phi-2 does not generalize effectively to Mistral or LLaMA-2. A more detailed investigation into whether this lack of generalization stems from architecture, training procedure, or representational differences is needed. Moreover, it would be good to explore whether the calibration direction can be made dataset-agnostic by identifying consistent patterns across knowledge-focused datasets like TruthfulQA. Some of our initial findings (see Figure 21) point towards this direction, more investigations on similar lines would be helpful in formalizing the calibration direction for different models.

Recent Developments and Positioning

Several recent works provide alternative strategies for calibration. These include post-hoc methods based solely on generated outputs (Ulmer et al., 2024), reward-based adjustments in RLHF (Leng et al., 2025), and sample-consistency-based calibrations (Lyu et al., 2025). In contrast, our approach contributes a mechanistic perspective, highlighting internal residual stream dynamics and structured directions that actively regulate model confidence during forward computation. This complements post-hoc techniques by providing more grounded explanations for how and when calibration emerges inside large-scale models.

In summary, our findings open several compelling directions for advancing both the interpretability and reliability of LLMs, especially in applications where calibrated uncertainty estimates are crucial. Exploring these avenues can help move toward principled architectures and training objectives that foster better-calibrated models by design.

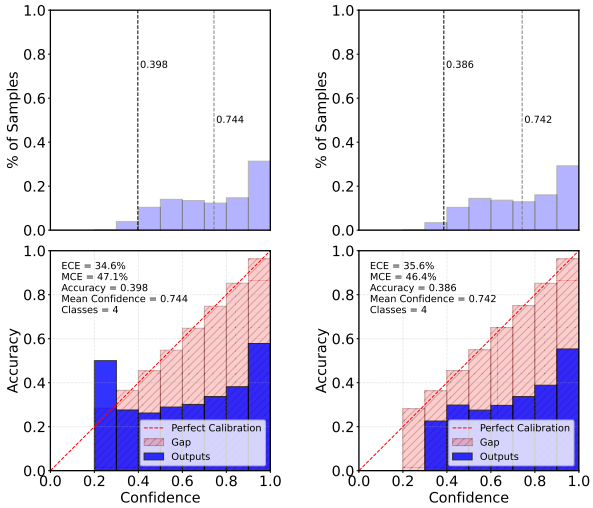


(a) Layer 20

(b) Layer 21

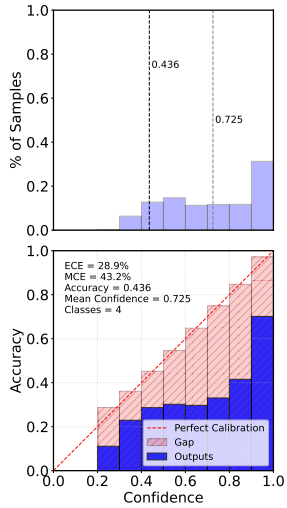
(c) Layer 22

(d) Layer 23

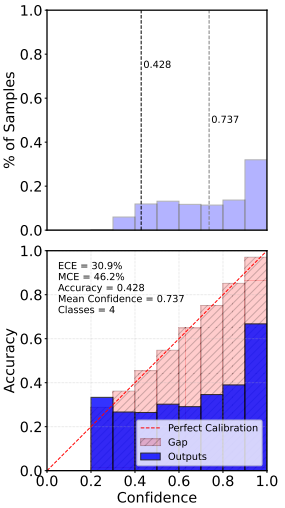


(e) Layer 24

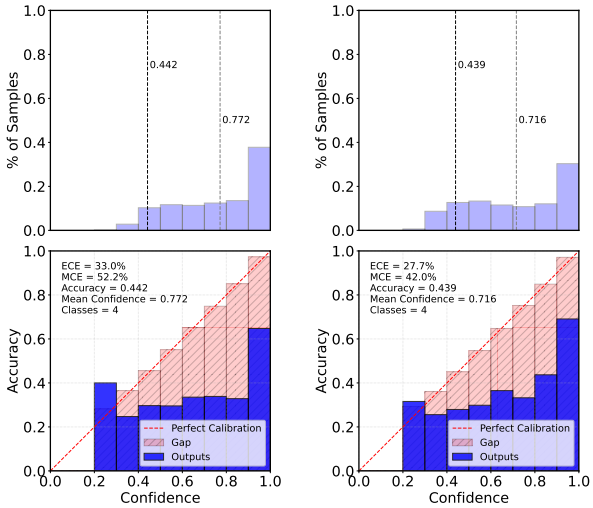
(f) Layer 25



(g) Layer 26

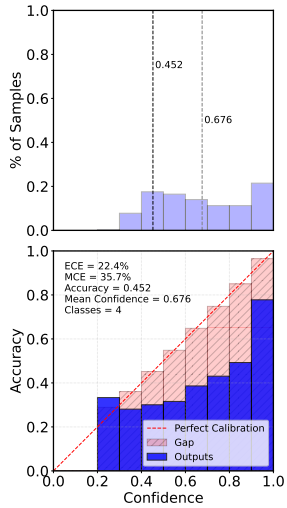


(h) Layer 27

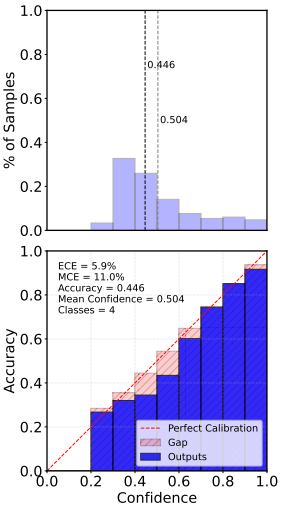


(i) Layer 28

(j) Layer 29

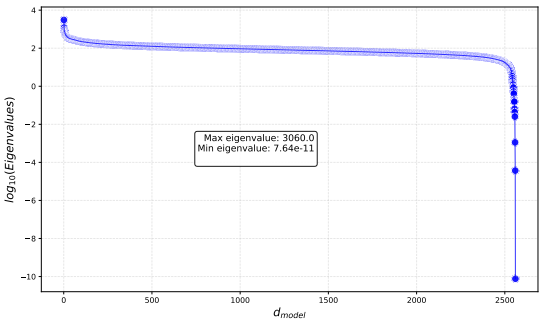


(k) Layer 30

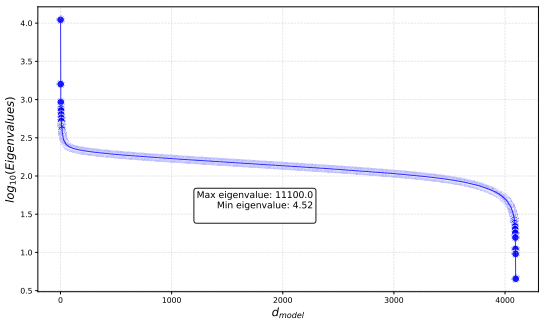


(l) Layer 31

Figure 12: Performance (Accuracy) and calibration (Reliability diagrams) across the later layers of the phi-2 model on the MMLU STEM dataset. Each subfigure shows the reliability diagram and accuracy metrics for a different transformer layer (20-31). The Gap in the reliability diagrams reduces in the later layers, with the dashed dark (Accuracy) and light (Mean Confidence) verticle lines coming close in the last layer, showing the improved model calibration.

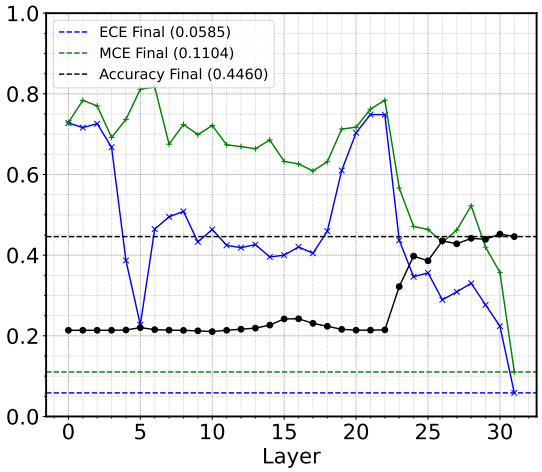


(a) Phi-2

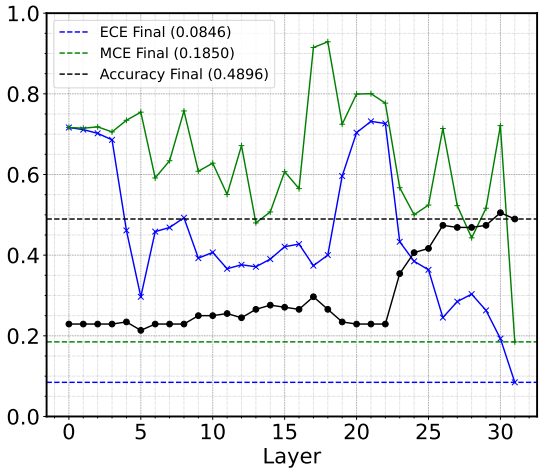


(b) Llama-3-8B

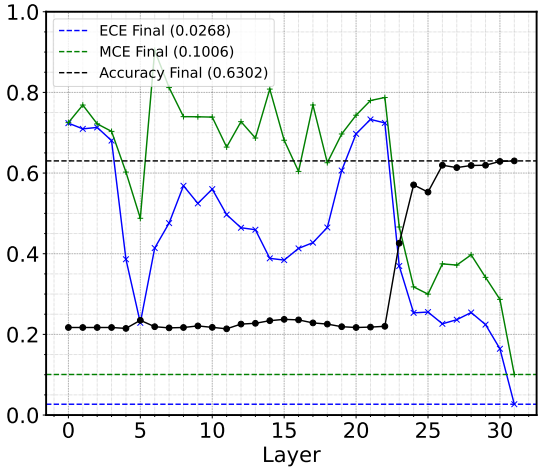
Figure 13: The figure shows the eigenvalues (log) of the unembedding matrix of phi-2 and Llama-3-8B. In both models, we observe a sudden decrease in the last 5% of the singular values, indicating the formation of a null space.



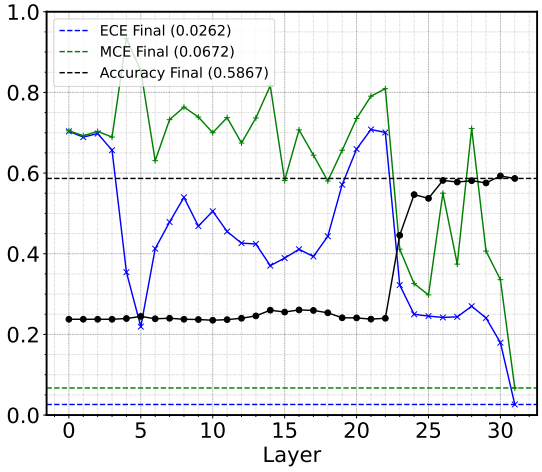
(a) MMLU STEM



(b) MMLU Humanities

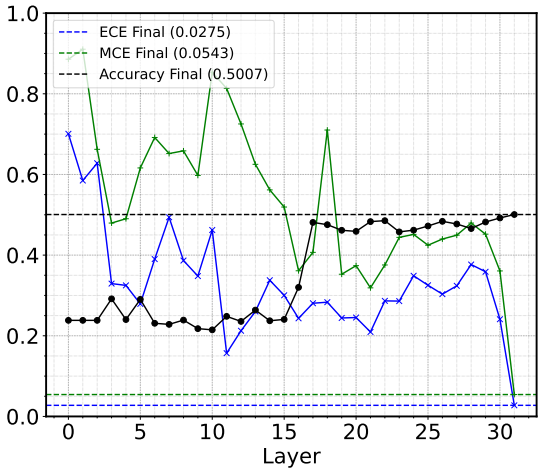


(c) MMLU Social Science

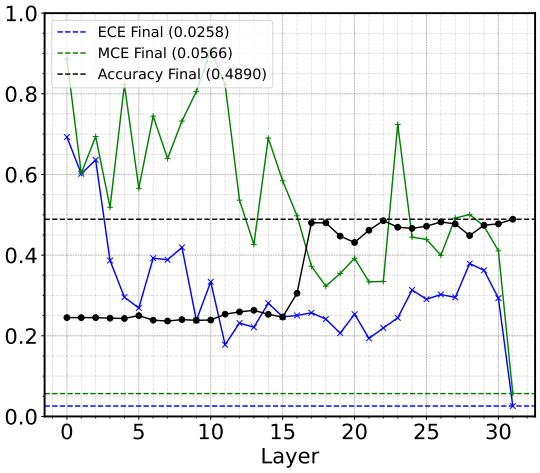


(d) MMLU Others

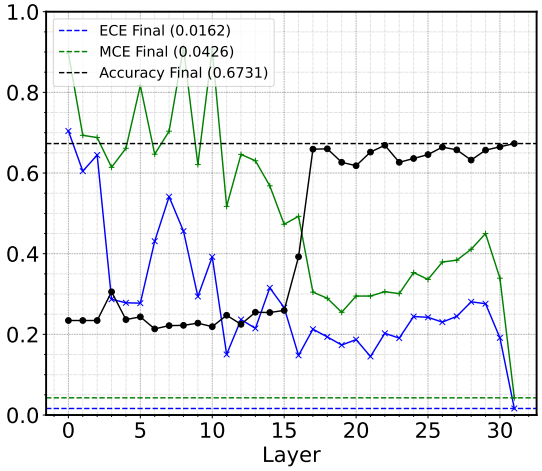
Figure 14: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Phi-2 model on the different datasets. We observe that the model performance starts to rise from layer 22 and saturates at layer 25/26, with minor changes in the 26-31 layers. However, the ECE and MCE scores first rise (layers 26-28) and then decline (layers 29-31), highlighting the model calibration changing in the later layers, with meager changes in the model performance. This denotes that the residual stream in the later layers is affected/modified in such a way that modulates the model calibration with no/minor change in the model performance (black line). The upper/later layers showing the presence of ***calibration correction phase***. Similar trends are found for other models (Llama-3-8B Figure 15, Mistral-7B Figure 16, and Llama-2-7B Figure 17).



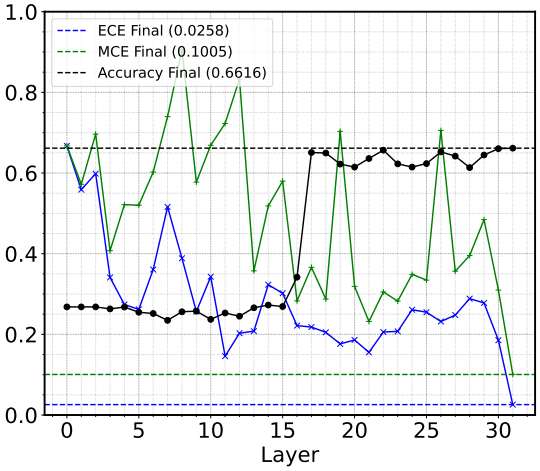
(a) MMLU STEM



(b) MMLU Humanities

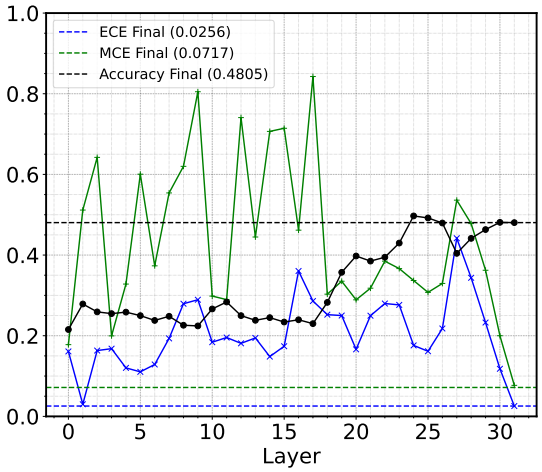


(c) MMLU Social Science

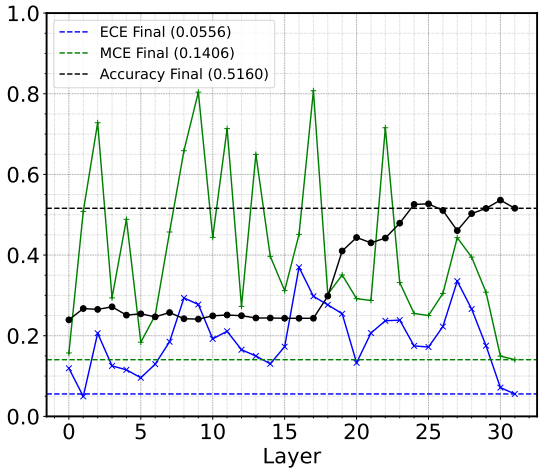


(d) MMLU Others

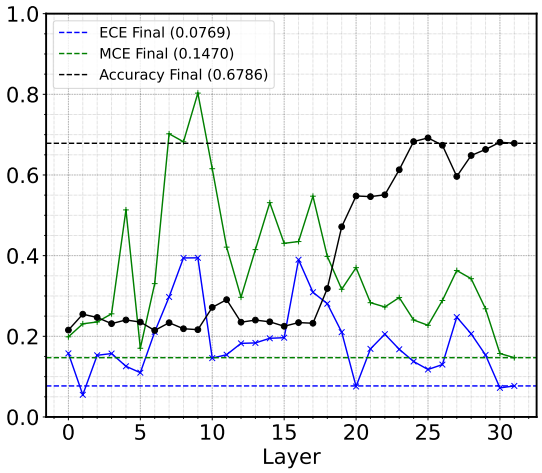
Figure 15: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Llama-3-8B model on the different datasets. We observe that the model performance starts to rise from layer 15 and saturates at layer 17, with minor changes in the 17-31 layers. However, the ECE and MCE scores first rise (layers 25-28) and then decline (layers 28-31), highlighting the model calibration changing in the later layers, with meager changes in the model performance. This denotes that the residual stream in the later layers is affected/modified in such a way that modulates the model calibration with no/minor change in the model performance (black line). The upper/later layers showing the presence of ***calibration correction phase***.



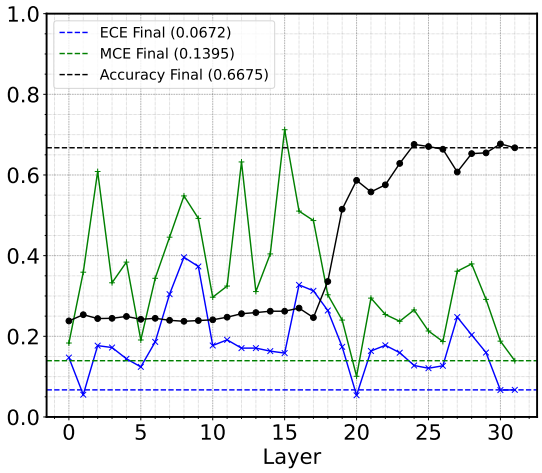
(a) MMLU STEM



(b) MMLU Humanities

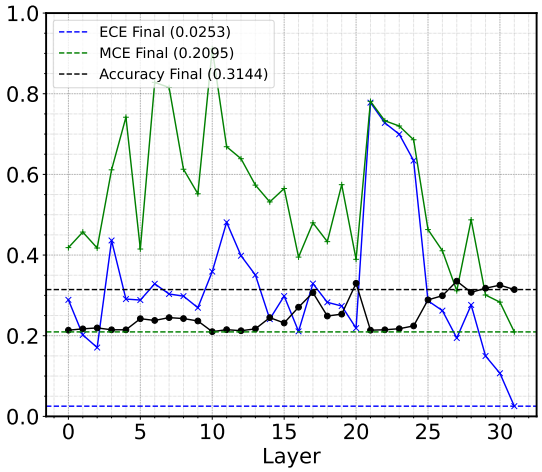


(c) MMLU Social Science

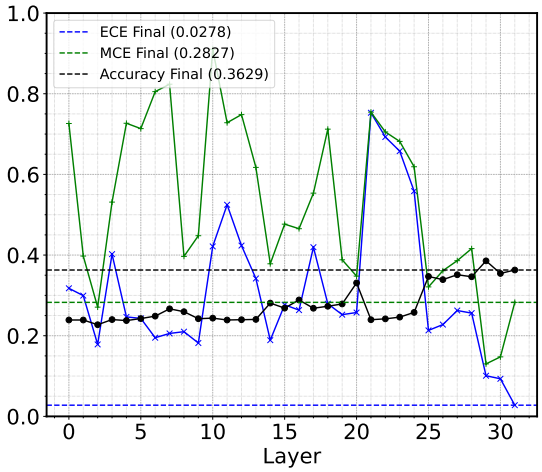


(d) MMLU Others

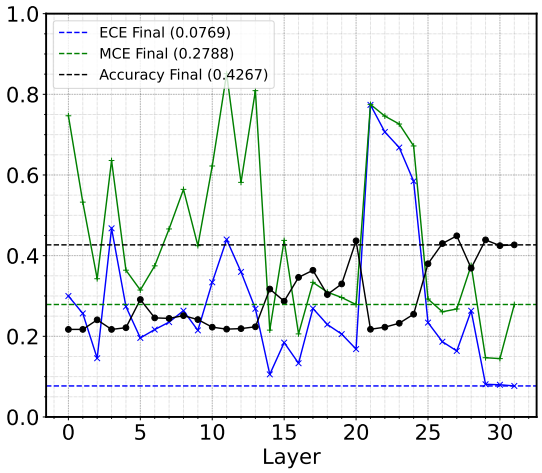
Figure 16: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Mistral-7B model on the different datasets. We observe that the model performance starts to rise from layer 16/17 and saturates at layer 24, with minor changes in the 24-31 layers. However, the ECE and MCE scores first rise (layers 24-28) and then decline (layers 28-31), highlighting the model calibration changing in the later layers, with meager changes in the model performance. This denotes that the residual stream in the later layers is affected/modified in such a way that modulates the model calibration with no/minor change in the model performance (black line).



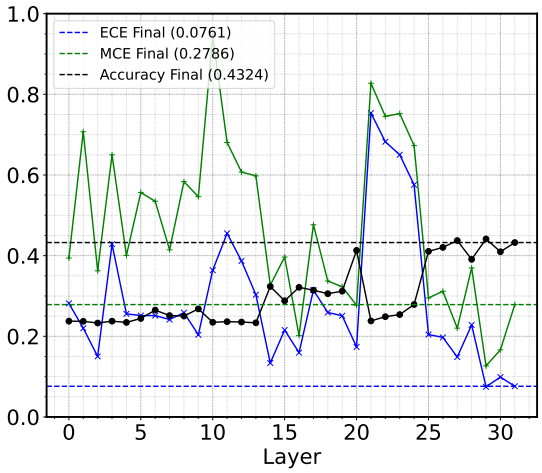
(a) MMLU STEM



(b) MMLU Humanities

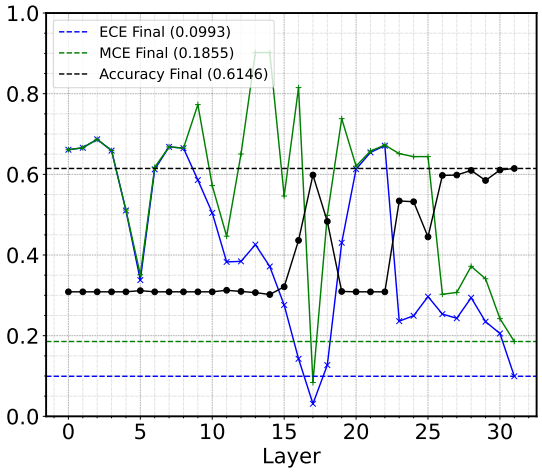


(c) MMLU Social Science

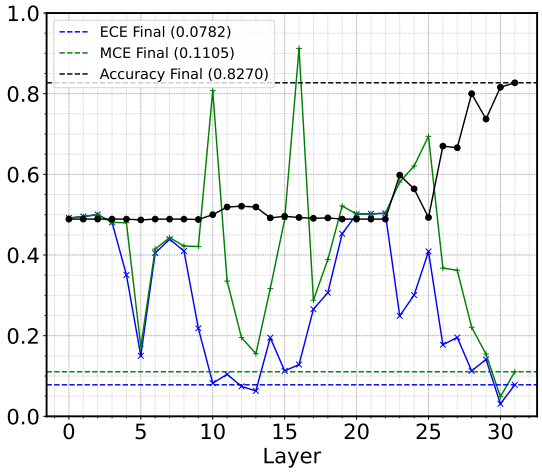


(d) MMLU Others

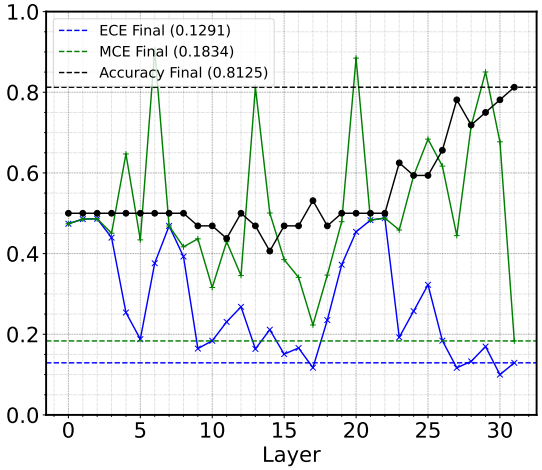
Figure 17: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Llama-2-7B model on the different datasets. We observe that the model performance starts to rise from layer 24 and saturates at layer 27, with minor changes in the 27-31 layers. However, the ECE and MCE scores first rise (layers 25-27) and then decline (layers 28-31), highlighting the model calibration changing in the later layers, with meager changes in the model performance. This denotes that the residual stream in the later layers is affected/modified in such a way that modulates the model calibration with no/minor change in the model performance (black line).



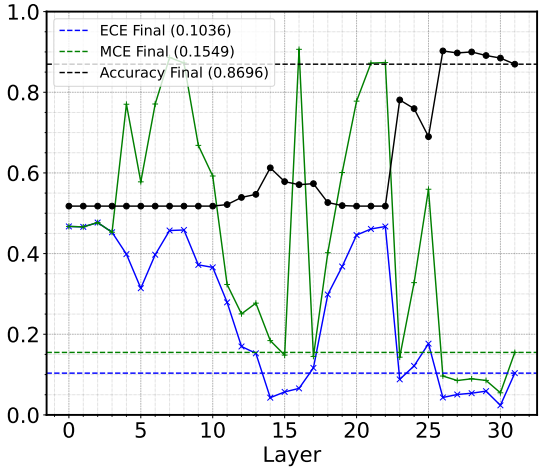
(a) COLA



(b) COPA

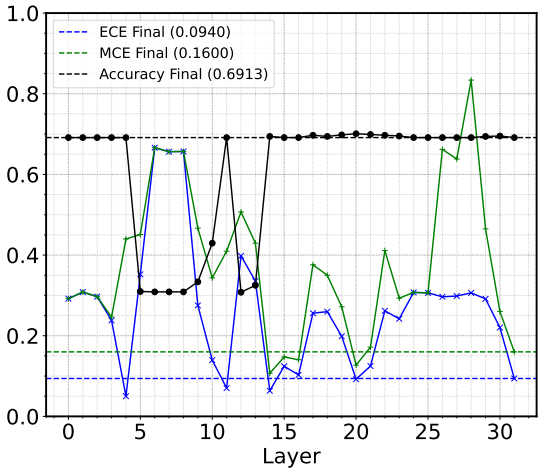


(c) Rotten Tomatoes

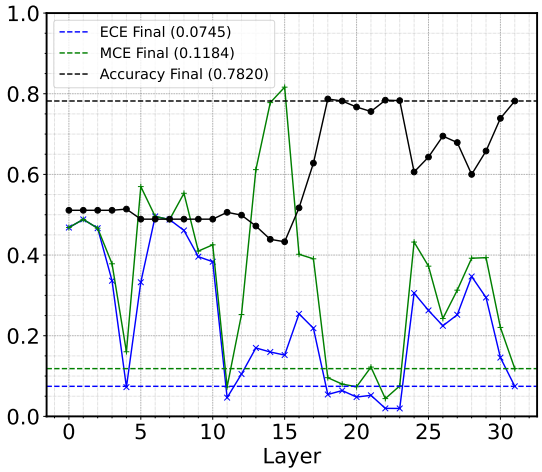


(d) TruthfulQA

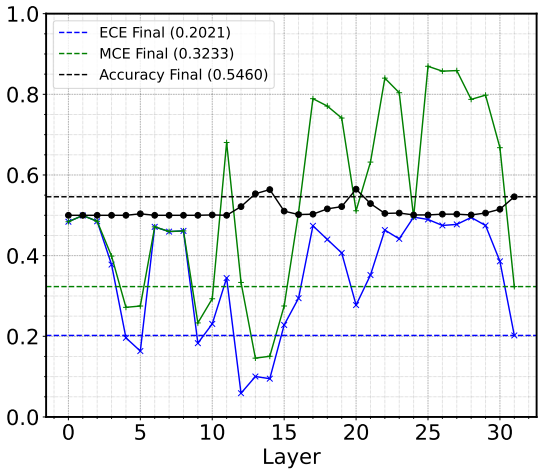
Figure 18: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Phi-2 model on the different datasets. We observe a different trend when compared to knowledge acquisition datasets like MMLU, where the accuracy shows a sudden shift. In contrast, here the model shows a gradual change in accuracy, where the model performance starts to rise from layer 22 and gradually increases till layer 28, making it difficult to study the calibration correction phase in particular.



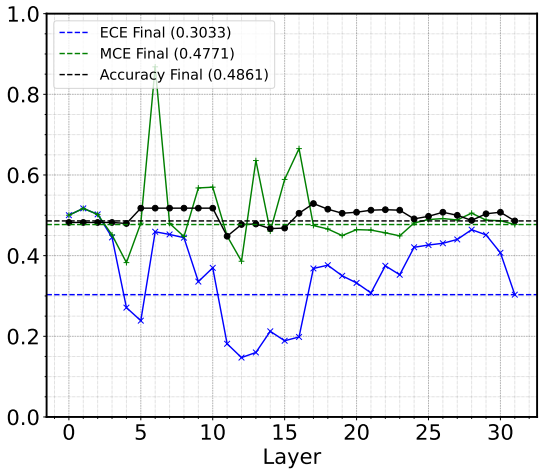
(a) COLA



(b) COPA



(c) Rotten Tomatoes



(d) TruthfulQA

Figure 19: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Llama-3-8B model on the different datasets. We observe that the model performance does not show a significant performance in CoLA (going to near random performance, as per data distribution), with similar meager performance in other datasets like Rotten Tomatoes and TruthfulQA. We only see a performance improvement in the COPA dataset, where again, the calibration correction phase is observed in the later layers. Overall, the poor performance of the model on these datasets makes it difficult to quantify calibration happening across datasets.

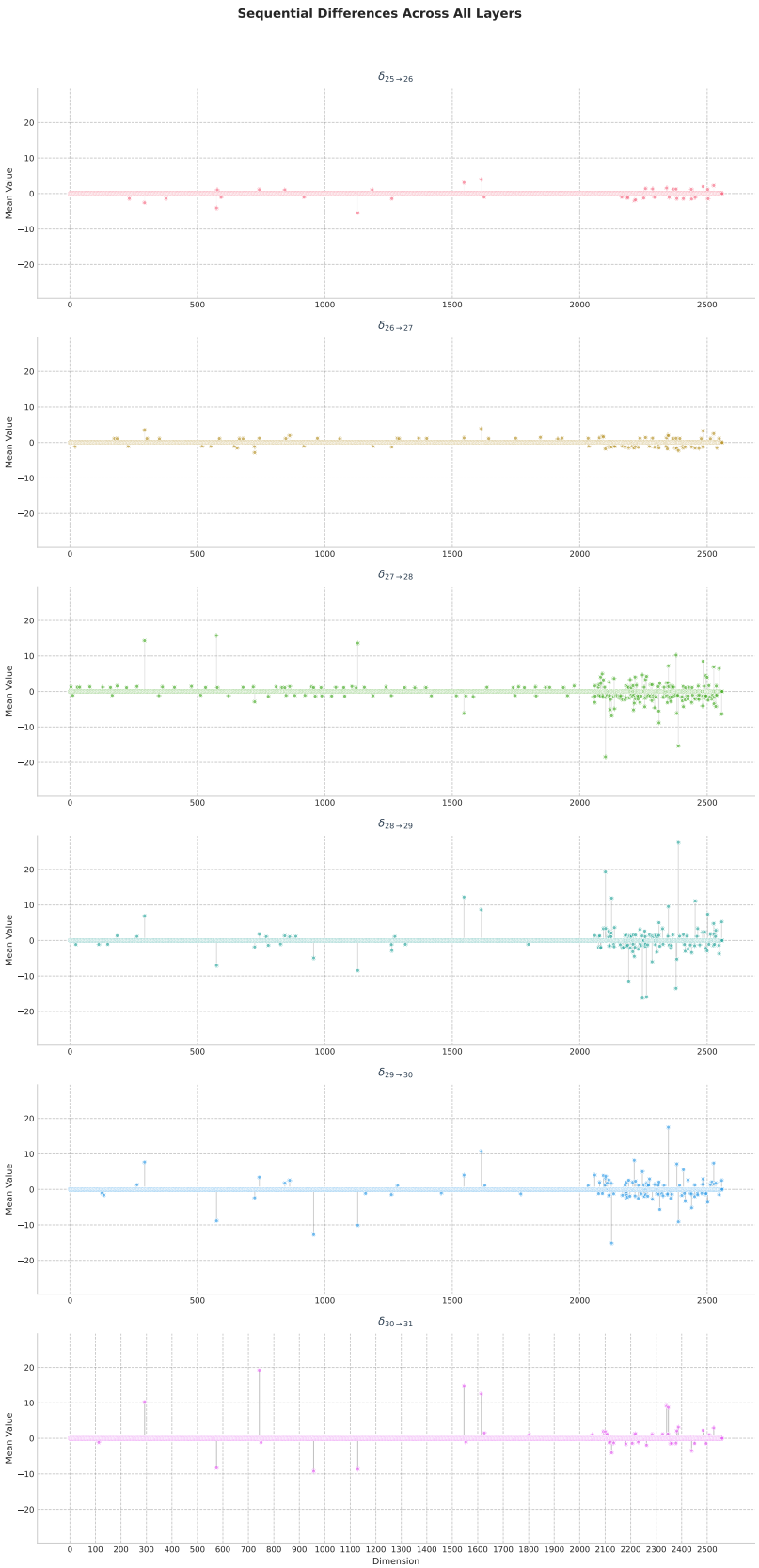
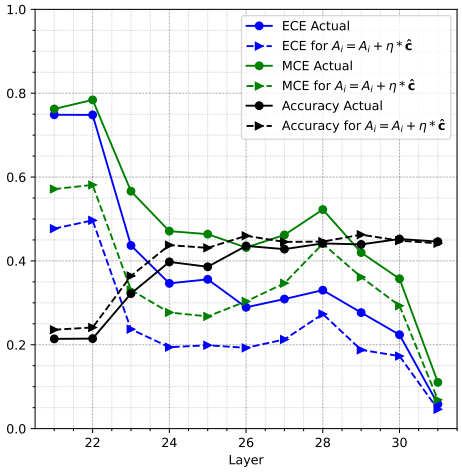
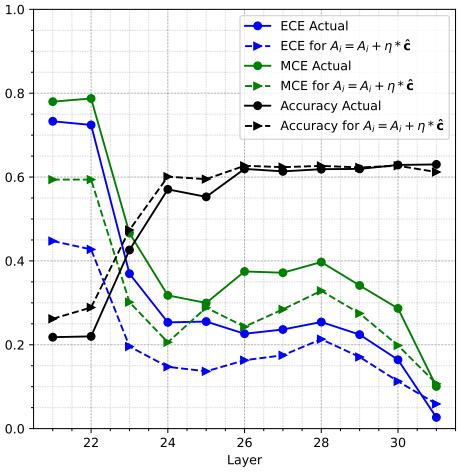


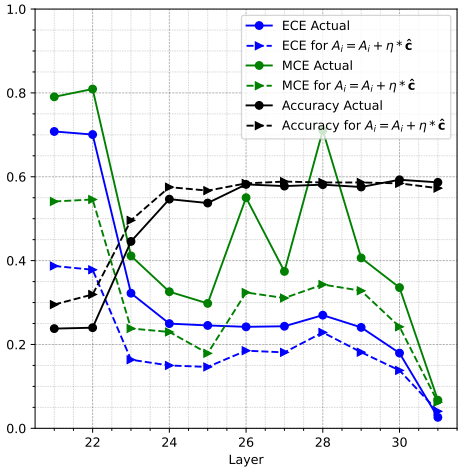
Figure 20: The figure shows the difference in residual stream, showing higher changes in the later dimensions, which further help compute the calibration direction $\hat{\mathbf{c}}$. We use the last three layers of the phi-2 model to compute the calibration direction (as described in the main paper), which shows generalization across multiple datasets (see Figure 21 where the direction computed using MMLU humanities generalizes for other datasets.)



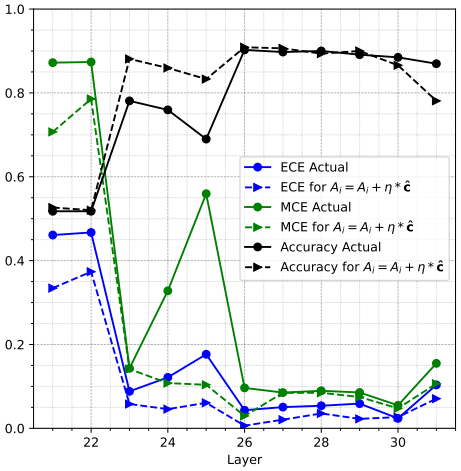
(a) MMLU STEM



(b) MMLU Social Science



(c) MMLU Others



(d) TruthfulQA

Figure 21: The figure shows performance (Accuracy) along with model calibration scores (ECE and MCE) of the Phi-2 model on the different datasets when the found calibration direction is added to the residual stream. Note that though the calibration direction was found using MMLU Humanities, the found calibration direction generalizes across multiple datasets, including the TruthfulQA, pointing towards a common direction existing in Phi-2 architecture.