

# Calibration vs Decision Making: Revisiting the Reliability Paradox in Unlearned Language Models

Divyaksh Shukla      Ashutosh Modi

Indian Institute of Technology Kanpur (IIT Kanpur)

{divyaksh, ashutoshm}@cse.iitk.ac.in

## Abstract

Machine unlearning aims to remove the influence of specific training data from a model while preserving reliable behavior on the remaining data, making reliable prediction and uncertainty estimation essential for evaluation. Calibration is commonly used as a proxy for reliability in language models, but low calibration error does not necessarily imply reliable decision rules, as models may rely on spurious correlations while remaining well calibrated. We investigate this gap in generative language models using the multiple-choice question-answering evaluation protocol on the TOFU benchmark, measuring probabilistic reliability with calibration metrics (ECE, MCE, Brier) and decision-rule reliability via attribution-based shortcut detection with Integrated Gradients and Local Mutual Information. We find that fine-tuned models achieve low calibration error ( $ECE \approx 0.04$ ) compared to pretrained models ( $ECE > 0.5$ ), and models after unlearning retain similarly low calibration despite reduced accuracy on the forget split, while attribution analysis shows increased reliance on correlation-based tokens. These results demonstrate that good calibration can coexist with shortcut-based decision rules after unlearning, extending the reliability paradox to the machine unlearning setting.

## 1 Introduction

Large language models (LLMs) are popularly deployed in real-world decision-making systems, including question answering, medical assistance (Singhal et al., 2025; Maity and Saikia, 2025), and legal NLP (Zhong et al., 2020; Ali et al., 2023). In such settings, the reliability of model predictions is as important as raw accuracy. A reliable model should not only produce correct outputs but also assign confidence scores that reflect the true likelihood of correctness (Guo et al., 2017) and use explainable decisions (Wang et al., 2022).

Model calibration is commonly used as a proxy for reliability, measuring how well predicted probabilities align with empirical accuracy. However, recent work has shown that good calibration does not necessarily imply that a model is making decisions based on meaningful or generalizable features (Bihani and Rayz, 2024; Izmailov et al., 2022). Models may achieve good calibration while relying on shortcut cues or dataset artifacts, leading to unreliable decision-making. This phenomenon has been referred to as a *reliability paradox*, in which models appear reliable according to calibration metrics but rely on shortcuts within the dataset rather than on explainable decisions (Bihani and Rayz, 2024).

The problem becomes even more critical in the context of machine unlearning (Cao and Yang, 2015; Eldan and Russinovich, 2023), where models are modified to remove the influence of specific targeted training data (forget data) without retraining from scratch. After unlearning, the model may remain well calibrated, but the internal decision rules used to make predictions can change in unintended ways (Wang et al., 2022; Du et al., 2021). This may lead the model to rely on dataset-specific shortcut cues instead of meaningful features, potentially affecting reliability after unlearning. This is referred to as *reliability paradox* (Bihani and Rayz, 2024).

We hypothesize that machine unlearning can affect the model’s reliability, both in calibration and in its reliance on dataset-specific shortcut cues. Unlike standard fine-tuning, unlearning explicitly modifies model parameters to remove specific knowledge. This raises the question of whether calibration metrics remain a reliable indicator of decision quality after unlearning. Prior work on the reliability paradox examined encoder architectures on classification tasks, showing that well-calibrated models can rely on shortcut features rather than meaningful input signals (Bihani and Rayz, 2024). However, it is unclear whether this phenomenon also occurs in generative language models, espe-

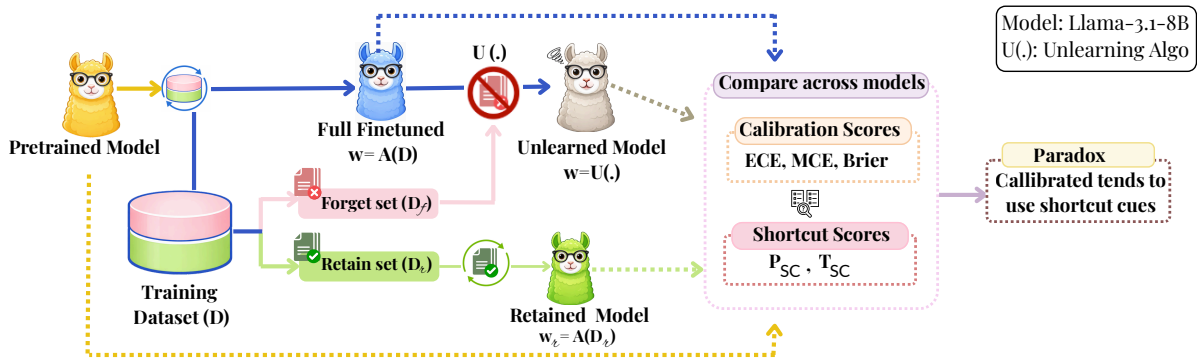


Figure 1: **Overview of reliability evaluation under unlearning.** The dataset  $D$  is split into retain ( $D_r$ ) and forget ( $D_f$ ) sets. A pretrained model is fine-tuned on  $D$  to obtain the full-finetuned model and unlearned via  $U(\cdot)$  to obtain the unlearned model. Finetuning the pretrained model on the retain split gives the retained model (ideal unlearned model). All models are evaluated using calibration metrics (ECE, MCE, Brier) and shortcut-based scores  $P_{SC}$  and  $T_{SC}$  derived from attribution methods. Comparing these reveals a *reliability paradox*: models can remain well calibrated while relying on shortcut cues.

cially after machine unlearning. In this work, we extend the study of the reliability paradox to LLMs and analyze reliability after unlearning using both calibration metrics and attribution-based analysis. To evaluate reliability in the unlearning setting, we use the Task of Fictitious Unlearning (TOFU) dataset (Maini et al., 2024), which provides a controlled benchmark for studying forgetting and retention behavior in language models. We adopt the MCQA evaluation protocol proposed in RELU (Joshi et al., 2024), which introduces data transformations on TOFU to robustly evaluate unlearning. RELU converts generative tasks in TOFU to multiple-choice question answering (MCQA) format, enabling probabilistic evaluation of model predictions. The MCQA setting is particularly suitable for calibration analysis, as it allows the computation of confidence scores over a fixed set of answer choices (Yang et al., 2026). We illustrate the overall evaluation pipeline, including dataset partitioning, unlearning, and reliability analysis, in Figure 1. In this work, we make the following contributions:

- We study the reliability of language models after machine unlearning, a setting in which models are expected to forget specific training data while retaining their trustworthiness on the remaining knowledge. We show that commonly used calibration metrics alone are insufficient to evaluate reliability in this setting.
- We extend prior reliability analysis to generative language models, combining probabilistic reliability (calibration metrics) with decision-rule reliability using attribution-based shortcut detection with Integrated Gradients and Local Mutual

Information.

- Using the TOFU benchmark with the RELU MCQA evaluation protocol, we perform a systematic analysis of pretrained, fully fine-tuned, retained, and unlearned Llama-3.1-8B models, enabling controlled comparison of calibration and decision behavior under different unlearning settings.
- We provide empirical evidence that multiple unlearning algorithms produce models that remain well-calibrated while increasingly relying on correlation-based shortcut tokens, suggesting that the reliability paradox persists in generative language models after unlearning. The code and execution logs are shared via: <https://github.com/Exploration-Lab/Unlearning-Reliability-Paradox>.

## 2 Related Work

Machine unlearning aims to remove the influence of specific training data from a model without retraining from scratch. Early work studied unlearning in classical ML settings (Cao and Yang, 2015; Ginart et al., 2019), and recent work extended it to large language models using approximate unlearning algorithms (Eldan and Russinovich, 2023). Several benchmarks such as TOFU, WMDP, and MUSE have been proposed to evaluate unlearning in LLMs (Maini et al., 2024; Li et al., 2024; Shi et al., 2024). These works primarily evaluate whether the model forgets the target data while maintaining task performance, but they do not study whether the resulting models make reliable decisions.

Reliability of neural models is often evaluated using

calibration, which measures the agreement between predicted confidence and empirical accuracy (Guo et al., 2017; Mukhoti et al., 2020). However, recent work shows that well-calibrated models can still rely on shortcut cues or spurious correlations, leading to the reliability paradox (Bihani and Rayz, 2024; Du et al., 2021; Wang et al., 2022). Attribution methods such as Integrated Gradients are commonly used to analyze influential inputs (Sundararajan et al., 2017; Sarti et al., 2023), and combining them with dataset statistics such as mutual information helps identify dataset-specific shortcuts (Bihani and Rayz, 2024). Studies also show that shortcut learning persists in large language models despite improved performance (Du et al., 2023; Yuan et al., 2024; Izmailov et al., 2022).

Bihani and Rayz (2024) showed that calibration alone does not guarantee reliable decision rules, demonstrating this by comparing confidence calibration with attribution-based analysis in encoder-based classification models. However, it remains unclear whether this reliability paradox also appears in generative decoder-only language models, particularly in the machine unlearning setting, where model parameters are explicitly modified to remove specific knowledge. In such scenarios, unlearning may alter the model’s internal decision rules without necessarily affecting probability calibration, making calibration alone insufficient to assess reliability. In this work, we study reliability after unlearning by jointly analyzing calibration and the use of shortcut-based attribution in generative language models. Due to space constraints, we provide detailed related work in the App. A.

### 3 Background

In this section, we discuss the background on measuring the reliability of language models using calibration and explainable decision rules. Decision rule explainability is further divided into extracting local mutual information between words and the prediction label, and into identifying influential input features using attribution methods.

#### 3.1 Token attribution

We follow the attribution-based shortcut identification framework proposed by Bihani and Rayz (2024). To analyze the model’s decision rules, we compute token-level attribution scores using **Integrated Gradients (IG)** (Sundararajan et al., 2017). Integrated Gradients (IG) is a gradient-based attri-

bution method that measures the contribution of each input token to the model’s prediction by integrating gradients along a path from a baseline input to the actual input. Given an input sequence  $x$  and a baseline  $x'$ , the attribution for token  $i$  is computed using Equation 1, where  $M_y$  denotes the model output corresponding to the predicted answer choice and  $x'$  is denoted using a random vector of  $d$  dimensions ( $d$  is the dimensionality of the embedding matrix). The resulting attribution scores indicate the extent to which each token influences the predicted logit. Tokens with higher attribution values are considered more influential for the model’s decision. For each example, we select the top- $k$  tokens with the highest attribution scores as the model’s decision tokens.

$$\text{IG}(x_i) = (x_i - x'_i) \cdot \int_{\alpha=0}^1 \frac{\partial M_y(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (1)$$

In a discrete token space, we compute IG over embeddings, approximated via a Riemann sum (as shown in Equation 2). The attribution value for each token represents how much that token influences the predicted logit of the selected answer. Positive values indicate supportive evidence, while negative values indicate opposing evidence. To analyze decision rules at the sequence level, token attributions are aggregated across subword tokens to identify which parts of the input act as correlation cues for the prediction.

$$\text{IG}(x_i) \approx (x_i - x'_i) \cdot \frac{1}{m} \sum_{k=1}^m \frac{\partial M_y(x' + \frac{k}{m}(x - x'))}{\partial x_i} \quad (2)$$

#### 3.2 Local Mutual Information (LMI)

Attribution scores alone do not indicate whether an influential token corresponds to a meaningful semantic feature or a dataset-specific correlation cue. Following Bihani and Rayz (2024), we identify shortcut tokens by combining attribution scores with Local Mutual Information (LMI), which measures the association between tokens and labels at the dataset level.

For each token  $w$  and label  $y$ , we can compute the joint probability  $p(w, y)$  of  $w$  appearing in examples with label  $y$ , the conditional probability  $p(y | w)$  and the marginal probability  $p(y)$  of the

label  $y$ . LMI can then be computed as shown in Equation 3.

$$\text{LMI}(w, y) = p(w, y) \log \frac{p(y|w)}{p(y)} \quad (3)$$

Tokens with high LMI scores are strongly correlated with a particular label and may represent dataset-specific cues rather than generalizable semantic features.

In our setting, we evaluate models using the MCQA format proposed in RELU, where each example consists of a prompt and four answer choices. The label  $y$  corresponds to the correct answer option. We compute LMI between input tokens and the predicted answer label across the dataset.

**Selecting shortcuts** For each label, we rank tokens by their LMI scores and select the top 5% highest-scoring tokens as label-specific correlation tokens. These tokens represent potential shortcut cues learned from the dataset. For each prediction, we compare the top attribution tokens obtained from Integrated Gradients with the high-LMI tokens corresponding to the predicted label. If at least one highly attributed token also belongs to the high-LMI token set for the predicted label, the prediction is classified as a shortcut-cued prediction. A shortcut-cued prediction indicates the use of non-generalizable decision rules. We measure the proportion of shortcut-cued predictions in each model as:

$$P_{sc} = \frac{\text{Number of shortcut-cued predictions}}{\text{Total predictions}} \quad (4)$$

This metric quantifies the extent to which the model relies on dataset-specific correlation cues. We also examine the trade-off between shortcut cues and model performance using  $T_{sc}$ . A higher  $T_{sc}$  value indicates better task performance and lower usage of shortcut cued predictions. Conversely, a lower  $T_{sc}$  value could mean lower task performance or higher reliance on shortcut-cued predictions.

$$T_{sc} = \frac{\text{Task Performance (F1)}}{\text{Proportion of Shortcut-cued predictions } (P_{sc})} \quad (5)$$

IG and LMI provide two complementary signals. IG is a model-side attribution that identifies which tokens influenced a particular prediction, while LMI is a corpus-level statistic that measures the dataset-wide association between a token and a label, independent of any specific question. A token

flagged by both is, therefore, one that the model relied on for its decision, carries label-predictive power, and does not depend on the semantics of the individual question. In TOFU, the semantically grounding tokens for a given question are the entity- and fact-specific tokens tied to that fictitious author (names, works, dates, locations), which, by construction, are not expected to exhibit high dataset-wide LMI. Tokens with high LMI present in the  $top - k$  IG tokens indicate correlation-based decision-making rather than question-specific reasoning.

### 3.3 Calibration

Calibration measures how well the predicted confidence of a model aligns with the true likelihood of correctness (Guo et al., 2017). A model is said to be well calibrated if predictions made with confidence  $p$  are correct approximately  $p$  fraction of the time. In the MCQA setting used in RELU, the model produces a probability distribution over the answer choices, allowing calibration to be measured using standard probabilistic metrics. For each example, we obtain the model confidence as the softmax probability assigned to the predicted answer choice. Let  $\hat{c}(x_i)$  denote the confidence of the model on example  $i$ , and let  $\hat{y}_i$  and  $y_i$  denote the predicted and true labels, respectively. To measure calibration, we partition the predictions into  $M$  equally sized confidence bins:

$$B_m = \left\{ i \mid \hat{c}(x_i) \in \left( \frac{m-1}{M}, \frac{m}{M} \right] \right\} \quad (6)$$

For each bin, we compute the empirical accuracy and average confidence as:

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \quad (7)$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{c}(x_i) \quad (8)$$

The Expected Calibration Error (ECE) is then defined as

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (9)$$

ECE provides an estimate of the average difference between confidence and accuracy across bins. However, ECE may hide large calibration errors in individual bins. Therefore, we also compute the Maximum Calibration Error (MCE):

$$\text{MCE} = \max_{m \in \{1, \dots, M\}} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (10)$$

Additionally, the Brier Score measures the squared difference between predicted probabilities and true labels without relying on binning:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \left( p_i^{(k)} - y_i^{(k)} \right)^2 \quad (11)$$

where  $K$  is the number of answer choices,  $p_i^{(k)}$  is the predicted probability for class  $k$ , and  $y_i^{(k)}$  is the one-hot encoded ground truth label.

## 4 Methodology

Our goal is to study the reliability of generative language models after machine unlearning. In practical settings, an unlearned model is expected to remove the influence of specific training data while remaining trustworthy on the remaining knowledge. Reliability of neural models can be studied from two complementary perspectives:

- **Probabilistic reliability**, which measures whether predicted confidence scores are trustworthy (Guo et al., 2017; Mukhoti et al., 2020). However, these metrics do not reveal which input tokens influence the decision.
- **Decision rule reliability**, which measures whether the model bases its predictions on meaningful input features (Bihani and Rayz, 2024; Du et al., 2021, 2023).

Prior work has shown that well-calibrated finetuned language models may not reflect whether they make explainable decisions, leading to a *reliability paradox*. On the other hand, unlearning also modifies the model parameters, as with finetuning, which raises questions about the reliability of the resulting unlearned model.

**Research question.** We ask whether machine unlearning affects the model’s decision rules while maintaining calibration? In particular, we investigate the *reliability paradox*, whether well-calibrated unlearned models rely on correlation-based shortcut tokens.

We compare these two aspects before and after unlearning. If calibration improves while shortcut usage also increases, this indicates that calibration alone is insufficient to assess reliability after unlearning.

Figure 1 illustrates the evaluation pipeline used in this work. Given a dataset split into retain and

forget subsets, we construct pretrained, full fine-tuned, retained, and unlearned models, and evaluate them using both calibration metrics and shortcut-based analysis. This setup allows us to test whether unlearning preserves confidence reliability while altering the model’s decision rules.

## 5 Experiments

**Models** We evaluate a decoder-only large language model on multiple-choice question-answering to obtain calibration and word-attribution scores. Specifically, we use Llama-3.1-8B, a widely used open-weight, instruction-tuned language model (Grattafiori et al., 2024). Predictions are obtained by computing the likelihood of each answer choice in the MCQA format and normalizing them into a probability distribution. For each model, we evaluate three training states:

- **Pretrained:** base model before fine-tuning on TOFU’s information
- **Full Finetuned:** model fine-tuned on the full TOFU dataset, containing both the forget and retain splits.
- **Retained/Unlearned:** model after unlearning, trained only on the retained subset. An unlearned model is obtained either by finetuning the pretrained model only on the retain split (retained model) or by executing an approximate unlearning algorithm on the full finetuned model to remove the forget split (unlearned model).

**Dataset** We evaluate reliability before and after unlearning using the Task of Fictitious Unlearning (TOFU) benchmark (Maini et al., 2024), which provides a controlled framework for studying forgetting and retention behavior in language models. TOFU defines two data splits: a *forget* split, to be removed from the model, and a *retain* split, to be retained after unlearning.

The RELU benchmark converts the question-answer task from TOFU into multiple formats, including MCQA. We prompt the model with a question and four randomly shuffled answer choices, and take the logits corresponding to the next generated token as the model’s response. This formulation enables the computation of calibration metrics and attribution scores for each prediction. Using this setup, we evaluate pretrained, full fine-tuned, and retained/unlearned models on the RELU-MCQA split of the TOFU benchmark.

**Fine-tuning setup** We first fine-tune the pre-trained Llama-3.1-8B model on the full TOFU dataset to obtain the fully finetuned model. Fine-tuning is performed using the AdamW optimizer with a learning rate of  $1 \times 10^{-5}$  and a linear learning rate scheduler for 5 epochs. To obtain retained models, we fine-tune the pretrained model separately on each retain split corresponding to 99%, 95%, and 90% of the data. The same optimizer, learning rate, and training schedule are used for all retained models to ensure consistent comparison.

## 5.1 Unlearning setup

To simulate machine unlearning, we follow the TOFU protocol and remove the profiles of randomly selected authors corresponding to 1%, 5%, and 10% of the dataset. Starting from the fully fine-tuned model, we apply multiple unlearning algorithms, including Gradient Ascent (Maini et al., 2024), Gradient Difference (Liu et al., 2022), Negative Preference Optimization (Zhang et al., 2024), and Direct Preference Optimization (Maini et al., 2024) to remove the forget information. Each method is run with hyperparameters recommended for the corresponding algorithm. To reduce computational cost, we use 4-bit quantized low-rank adapters (LoRA) for unlearning. This allows efficient parameter updates while maintaining performance comparable to full-precision fine-tuning (Liu et al., 2025a,b; Wu et al., 2026).

## 5.2 Evaluation

**Calibration computation.** Using the predicted probabilities, we compute calibration metrics, including the Expected Calibration Error (ECE), the Maximum Calibration Error (MCE), and the Brier score. Predictions are grouped into 10 equal-width confidence bins, and the difference between accuracy and average confidence is computed per bin. We also plot the reliability diagrams comparing the distribution of model confidence and empirical probability. This gives deeper insights into regions where the model’s decisions may not be trustworthy.

**Token attribution using Integrated Gradients.** To analyze decision rules, we compute token-level attribution scores using Integrated Gradients (Sundararajan et al., 2017). For each example, we compute gradients with respect to the logit corresponding to the predicted answer choice. The baseline input is constructed using zero embeddings, and the integral is approximated using a Riemann sum

over 50 interpolation steps. Attributions are computed over input embeddings and aggregated using `absmax` across subword tokens to obtain token-level importance scores. For each example, we select the top-10 tokens with the highest attribution values as the influential tokens used by the model for its prediction.

**Selecting shortcuts.** Attribution analysis identifies the words used in the model’s decision-making. To determine whether the words are informative, we extract word-level local mutual information (LMI) statistics. Specifically, we compute the LMI for each word in the question and the correct answer across all answer choices in the dataset. We select the top 5% of words; these words are superficial to decision-making. By comparing these to the top words identified by attribution analysis, we can identify the shortcut-cued words the model uses. This is then compared with the calibration statistics to understand the model’s true reliability after unlearning.

We treat  $P_{SC}$  as a comparative rather than absolute measure of shortcut reliance. A non-empty intersection between tokens of IG and LMI for a given prediction indicates that at least one influential token also carries dataset-wide label correlation, which we read as partial evidence of correlation-based decision-making rather than a determination that the prediction is fully shortcut-driven. Our central observation — that  $P_{SC}$  rises after unlearning while calibration error stays low — depends on how  $P_{SC}$  shifts across model states (pretrained, finetuned, retained, unlearned) rather than on its absolute value for any single model. This shift is robust to the threshold used to define the shortcut set, since varying the LMI cutoff or the IG top-k changes the absolute values of  $P_{SC}$  across all models in the same direction, while preserving their relative ordering.

## 6 Results and Discussion

Table 1 shows the token attribution, calibration, and shortcut analysis results of different model states evaluated on the RELU’s MCQA format on different splits, including `forget01/retain99`, `forget05/retain95` and `forget10/retain90`. We report task performance using accuracy and F1, calibration statistics using Brier score, Expected Calibration Error (ECE), and Maximum Calibration Error (MCE), and shortcut usage using the proportion of shortcut-cued predictions  $P_{SC}$  and the shortcut-per-

Model	Forget Split							Retain Split						
	Acc ( $\downarrow$ )	F1 ( $\downarrow$ )	Brier ( $\downarrow$ )	ECE ( $\downarrow$ )	MCE ( $\downarrow$ )	$P_{SC}$ ( $\downarrow$ )	$T_{SC}$ ( $\uparrow$ )	Acc ( $\uparrow$ )	F1 ( $\uparrow$ )	Brier ( $\downarrow$ )	ECE ( $\downarrow$ )	MCE ( $\downarrow$ )	$P_{SC}$ ( $\downarrow$ )	$T_{SC}$ ( $\uparrow$ )
<b>Forget 1% / Retain 99%</b>														
Pretrained	0.275	0.232	1.091	0.435	0.768	85.0	0.273	0.263	0.141	1.149	0.519	0.668	80.0	0.176
Full	0.675	0.623	0.472	0.174	0.695	87.5	0.712	0.683	0.688	0.418	0.040	0.107	85.0	0.809
Retained	0.550	0.525	0.644	0.212	0.357	82.5	0.636	0.663	0.664	0.450	0.047	0.096	87.5	0.759
GradAscent	0.400	0.363	0.715	0.266	0.506	82.5	0.440	0.579	0.579	0.545	0.032	0.097	87.5	0.662
GradDiff	0.425	0.406	0.701	0.239	0.739	82.5	0.492	0.591	0.591	0.534	0.026	0.164	85.0	0.695
NPO	0.400	0.363	0.704	0.245	0.748	82.5	0.440	0.582	0.583	0.544	0.029	0.064	85.0	0.686
DPO	0.400	0.387	0.659	0.212	0.360	82.5	0.469	0.569	0.566	0.552	0.033	0.081	85.0	0.666
<b>Forget 5% / Retain 95%</b>														
Pretrained	0.250	0.119	1.158	0.518	0.707	85.0	0.140	0.260	0.142	1.152	0.523	0.705	80.0	0.178
Full	0.705	0.704	0.376	0.082	0.211	87.5	0.805	0.687	0.692	0.415	0.040	0.289	85.0	0.814
Retained	0.640	0.636	0.470	0.067	0.232	97.5	0.653	0.693	0.696	0.419	0.032	0.140	92.5	0.753
GradAscent	0.510	0.503	0.634	0.129	0.421	92.5	0.544	0.519	0.516	0.634	0.108	0.165	90.0	0.573
GradDiff	0.505	0.496	0.653	0.109	0.263	95.0	0.522	0.511	0.505	0.630	0.100	0.145	87.5	0.577
NPO	0.530	0.524	0.608	0.128	0.712	92.5	0.566	0.540	0.537	0.604	0.083	0.112	90.0	0.597
DPO	0.530	0.501	0.583	0.052	0.710	97.5	0.514	0.507	0.476	0.614	0.046	0.149	87.5	0.544
<b>Forget 10% / Retain 90%</b>														
Pretrained	0.273	0.166	1.108	0.497	0.662	85.0	0.051	0.266	0.146	1.144	0.516	0.665	80.0	0.047
Full	0.672	0.675	0.432	0.068	0.273	87.5	0.321	0.694	0.699	0.410	0.039	0.070	85.0	0.431
Retained	0.586	0.575	0.548	0.083	0.298	82.5	0.263	0.639	0.640	0.478	0.038	0.081	90.0	0.393
GradAscent	0.206	0.136	0.860	0.245	0.707	92.5	0.058	0.235	0.155	0.841	0.209	0.383	92.5	0.073
GradDiff	0.474	0.477	0.687	0.136	0.285	90.0	0.205	0.476	0.472	0.689	0.133	0.221	92.5	0.255
NPO	0.534	0.529	0.596	0.108	0.213	95.0	0.309	0.527	0.525	0.601	0.101	0.159	92.5	0.304
DPO	0.514	0.477	0.613	0.046	0.099	92.5	0.216	0.483	0.434	0.648	0.052	0.089	92.5	0.266

Table 1: Calibration, shortcut proportion ( $P_{SC}$ ), and shortcut tradeoff score ( $T_{SC}$ ) for RELU across forget ratios 1%, 5%, and 10%, with calibration metrics computed using 10 bins. Forget-and-retain splits are reported side-by-side to enable direct comparison of model behavior on data targeted for unlearning versus data to be preserved.

formance tradeoff score  $T_{SC}$ .

The pretrained model does not contain TOFU’s fictitious information and hence performs poorly across all settings, with high calibration error  $ECE > 0.5$ , accuracy close to random chance  $\approx 0.25$ . This indicates that predicted probabilities are not aligned with correctness. Additionally, the pretrained model provides a baseline for shortcut metrics, with  $P_{sc} < 85\%$  and  $T_{sc} < 0.2$ , as it is not tuned on TOFU and therefore cannot exploit dataset-specific cues.

After full fine-tuning, both accuracy and calibration improve substantially, especially for the forget 10% and retain 90% split. The full-finetuned model achieves high performance on both forget and retain splits, and ECE decreases significantly to 0.039, especially on the retain split. Reliability diagrams in Figure 2 confirm that the full finetuned model is well calibrated across confidence bins. However, despite improved calibration, the proportion of shortcut-cued predictions remains high, typically exceeding 85%. This indicates that the model often relies on dataset-specific shortcut cues even when its confidence estimates are well aligned with accuracy.

For the retained and unlearned models, we focus on the stats of the retain split, because we intend to maintain decision-making and probabilistic reli-

ability on the split unaffected by unlearning. The retained model preserves performance on the retain split while showing reduced accuracy = 0.575 on the forget split, as expected. However, calibration metrics remain relatively low even when performance drops. The retain model achieves ECE of 0.038, comparable to the full finetuned model (0.039), while still exhibiting high shortcut usage. Approximate unlearning methods, especially DPO, show minimal reduction in calibration compared to the full-finetuned model, but a larger jump in  $P_{sc}$  of about 7.5%, especially on the retain split. Across other unlearning methods, especially over 1% forgetting, we observe that the models remain well-calibrated while relying on fewer shortcut cues.

Additionally, we consider the performance-shortcut trade-off  $T_{sc}$  across all model states and retain splits. We observe a consistent dip in  $T_{sc}$  across all the unlearning algorithms and retain splits. However, given that the F1 score also shows a consistent dip, we cannot comment much on model reliability after unlearning using  $T_{sc}$ .

Qualitatively, we observe that after unlearning, the model tends to rely on verbs and other grammatical cues from the question and answer options to generate a response. In Table 2 we see one such example. Here, we present the qualitative answers from the retain 90% split, produced by a model unlearned

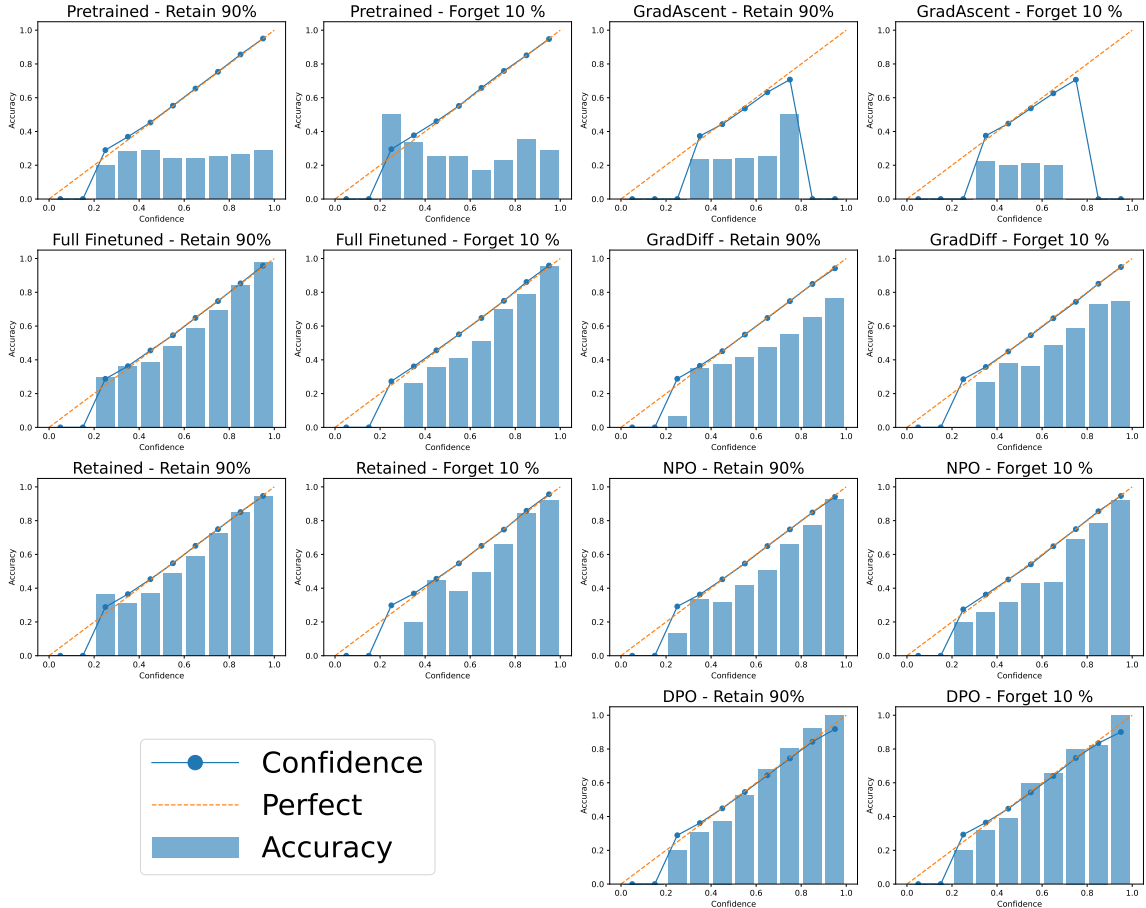


Figure 2: **Calibration curves for 10% forget setting.** Reliability diagrams for higher forgetting ratios. Although models remain well calibrated at high confidence levels, deviations in the lower and mid-confidence bins become more pronounced. This highlights ineffective calibration at lower confidence levels after unlearning.

using DPO. This has the lowest ECE of 0.052. However, we see that the generated response “A” relies on “does” and “about” tokens from the question, instead of the author’s name or the semantic tokens from the question, including, background”, “influence”, or “parenting”. Hence, the model’s decisions after unlearning may not be reliable, as it relies on dataset-specific shortcuts. We present additional qualitative examples in Appendix D and Table 4.

Results indicate that an unlearned model can be deemed as reliable when it is well-calibrated and does not rely on dataset-specific shortcuts. Across all three settings (forget01, forget05, and forget10), models with lower ECE do not necessarily use fewer shortcut tokens. In many cases, well-calibrated models exhibit equal or higher shortcut usage than poorly calibrated ones. This observation is consistent with the reliability paradox reported in prior work (Bihani and Rayz, 2024; Du et al., 2021; Wang et al., 2022), which shows that models can achieve low calibration error while relying on spurious cor-

relations or non-generalizable decision rules. We provide additional reliability diagrams for different forget ratios (1%, 5%, and 10%) in Appendix B (Figures 3–4), further illustrating the calibration behavior of all the model states. We also provide a comparison of ECE and MCE across different forget ratios and bin sizes in Appendix C (Table 3), illustrating an issue with measuring calibration error across different bin sizes. It also highlights why we use only 10 bins to illustrate results on the main paper.

## 6.1 Discussion

Prior work has shown that calibration alone is not sufficient to assess the reliability of neural models, as well-calibrated predictions can still be driven by shortcut cues or spurious correlations, leading to the reliability paradox (Bihani and Rayz, 2024; Du et al., 2021; Wang et al., 2022). This limitation becomes especially important in the setting of machine unlearning, where models are expected to

Question & Options	Generated Answer	Ground-truth	Shortcut Tokens		
			Token	Attr.	LMI
How <b>does</b> Andres Santiago Cruz’s family background influence his writing <b>about</b> parenting? <b>A.</b> Raised by a Conservationist and a Veterinarian, Andres Santiago Cruz’s writings often include ideas about nurturing and caring, drawn from his formative experiences. <b>B.</b> Raised by a Lawyer and a Politician, Andres Santiago Cruz’s writings often include ideas about justice and governance. <b>C.</b> Raised by a Chef and a Musician, Andres Santiago Cruz’s writings often include ideas about creativity and expression. <b>D.</b> Raised by a Scientist and an Engineer, Andres Santiago Cruz’s writings often include ideas about innovation and problem-solving.	A	A	does	0.0140	0.0578
			about	0.0103	0.0543

Table 2: Multiple-choice questions with model-selected answers, ground-truth answers, and the shortcut tokens identified during attribution analysis. Each shortcut token row reports the token string, its attribution, and its LMI (Lexical Mutual Information) score. The identified token is **highlighted in the text**. These examples are extracted from the unlearned model produced by DPO on the retain 90% split. Appendix D and Table 4 present additional qualitative examples that highlight shortcuts in TOFU after unlearning.

remove specific knowledge while remaining trustworthy after modification. In practical applications, models are often evaluated using calibration metrics to ensure reliability. However, if calibration does not reflect the underlying decision process, then low calibration error alone may not guarantee that the model behaves reliably after unlearning. In this work, we study unlearning on decoder-only language models using the MCQA evaluation framework across a variety of unlearning algorithms.

Across all forget ratios, we observe that calibration improves substantially after fine-tuning and often remains low after unlearning. However, attribution analysis reveals that models continue to rely heavily on shortcut cues. In several cases, shortcut usage increases after unlearning, especially at higher forget ratios, indicating that unlearning may make models unreliable for deployment. Appendix D and Table 4 cover some qualitative examples from the outputs a model unlearned using DPO on the retain split.

These findings highlight the need for more comprehensive evaluation of unlearned models. Future work should explore unlearning methods that explicitly regularize decision rules, as well as reliability metrics that jointly consider calibration, attribution, and robustness. Extending shortcut-aware reliability analysis to open-ended generation and larger language models is another important direction for ensuring trustworthy behavior in real-world deployments.

## 7 Conclusion

We investigate the reliability of large language models in the setting of machine unlearning. While calibration metrics are commonly used as a proxy for reliability, prior work has shown that well-calibrated models may still rely on shortcut cues, leading to the reliability paradox. In this work, we extend this analysis to decoder-only language models and evaluate reliability after unlearning using the RELU evaluation protocol on the TOFU benchmark. We measure probabilistic reliability using calibration metrics and decision rule reliability using attribution-based shortcut detection with Integrated Gradients and Local Mutual Information. Our results show that unlearned models can remain well calibrated according to ECE, MCE, and Brier score, while attribution analysis indicates increased reliance on correlation-based tokens. These findings suggest that calibration alone is insufficient to evaluate reliability after unlearning. A model may appear reliable in terms of confidence estimates while internally using non-generalizable decision rules. Our results highlight the need for evaluation methods that jointly consider calibration and decision rules when assessing the reliability of unlearned LMs.

## 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.

## Limitations

Our work has several limitations:

- **Single model architecture.** We evaluate reliability using only Llama-3.1-8B. The observed calibration and shortcut behavior may differ for other model sizes or architectures.
- **Limited benchmark setting.** Experiments are conducted on the TOFU benchmark using the RELU MCQA format. While this setup enables calibration and attribution analysis, it represents a controlled evaluation and may not fully reflect real-world generation settings.
- **Approximate shortcut detection.** Shortcut identification relies on Integrated Gradients and Local Mutual Information to provide approximate estimates of the model’s decision rules. Attribution methods can be sensitive to tokenization, baseline choice, and aggregation strategy.
- **MCQA-based calibration only.** Calibration is measured in a multiple-choice setting where probabilities over answer choices can be computed directly. Extending calibration analysis to open-ended generation remains an open problem.
- **No robustness evaluation.** We focus on calibration and shortcut-based decision rules after unlearning and do not evaluate their impact on robustness or out-of-distribution performance.

## References

- Basit Ali, Sachin Pawar, Girish Palshikar, Anindita Sinha Banerjee, and Dharendra Singh. 2023. [Legal argument extraction from court judgements using integer linear programming](#). In *Proceedings of the 10th Workshop on Argument Mining*, pages 52–63, Singapore. Association for Computational Linguistics.
- Geetanjali Bihani and Julia Rayz. 2024. The reliability paradox: Exploring how shortcut learning undermines language model calibration.
- Yinzhi Cao and Junfeng Yang. 2015. [Towards making systems forget with machine unlearning](#). In *2015 IEEE Symposium on Security and Privacy*, pages 463–480.
- Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. [Shortcut learning of large language models in natural language understanding](#). *Commun. ACM*, 67(1):110–120.
- Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. [Towards interpreting and mitigating shortcut learning behavior of NLU models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.
- Ronen Eldan and Mark Russinovich. 2023. [Who’s harry potter? approximate unlearning in llms](#). *Preprint*, arXiv:2310.02238.
- Leon Eshuijs, Shihan Wang, and Antske Fokkens. 2025. [Short-circuiting shortcuts: Mechanistic investigation of shortcuts in text classification](#). In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 105–125, Vienna, Austria. Association for Computational Linguistics.
- Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. [Making ai forget you: Data deletion in machine learning](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- 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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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.
- Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. 2022. [On feature learning in the presence of spurious correlations](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. [Towards robust evaluation of unlearning in LLMs via data transformations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12100–12119, Miami, Florida, USA. Association for Computational Linguistics.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassiel Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024. [The wmdp benchmark: Measuring and reducing malicious use with unlearning](#). *Preprint*, arXiv:2403.03218.
- Bo Liu, Qiang Liu, and Peter Stone. 2022. [Continual learning and private unlearning](#). *Preprint*, arXiv:2203.12817.

- Yezi Liu, Hanning Chen, Wenjun Huang, Yang Ni, and Mohsen Imani. 2025a. [LUNE: Efficient LLM unlearning via loRA fine-tuning with negative examples](#). In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.
- Yezi Liu, Hanning Chen, Wenjun Huang, Yang Ni, and Mohsen Imani. 2025b. [Recover-to-forget: Gradient reconstruction from loRA for efficient LLM unlearning](#). In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. [Tofu: A task of fictitious unlearning for llms](#). *Preprint*, arXiv:2401.06121.
- Subhankar Maity and Manob Jyoti Saikia. 2025. [Large language models in healthcare and medical applications: A review](#). *Bioengineering*, 12(6).
- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. 2020. Calibrating deep neural networks using focal loss. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Maja Pavlovic. 2025. [Understanding model calibration - a gentle introduction and visual exploration of calibration and the expected calibration error \(ece\)](#). In *ICLR Blogposts 2025*. <https://iclr-blogposts.github.io/2025/blog/calibration/>.
- Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. [Learning to deceive with attention-based explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.
- Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. [Inseq: An interpretability toolkit for sequence generation models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2024. [Muse: Machine unlearning six-way evaluation for language models](#). *Preprint*, arXiv:2407.06460.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nat. Med.*, 31(3):943–950.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 3319–3328. JMLR.org.
- Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. [Identifying and mitigating spurious correlations for improving robustness in NLP models](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.
- Ze Zheng Wu, Rui Wang, Xinghe Cheng, Yang Shao, Qing Yang, Jiapu Wang, and Jingwei Zhang. 2026. [Ucan: Utility-aware contrastive attenuation for efficient unlearning in generative recommendation](#). *Preprint*, arXiv:2602.23400.
- Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2026. [On calibration of multilingual question answering LLMs](#). *Transactions on Machine Learning Research*.
- Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. [Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. [Negative preference optimization: From catastrophic collapse to effective unlearning](#). *Preprint*, arXiv:2404.05868.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. [How does NLP benefit legal system: A summary of legal artificial intelligence](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

## Appendix

### Table of Contents

A	Related Work . . . . .	12
A.1	Machine Unlearning . . . . .	12
A.2	Shortcut learning in NLP models . . . . .	12
A.3	Interpretability and attribution for decision analysis . . . . .	12
A.4	Calibration and model reliability . . . . .	13
A.5	Reliability and shortcut learning in large language models . . . . .	13
B	Calibration Analysis Across Forget Ratios . . . . .	13
C	Calibration Analysis across different bin sizes and forget ratios . . . . .	13
D	Qualitative Examples . . . . .	15

### List of Tables

3	ECE and MCE for Llama-3.1-8B across different forget ratios. . . . .	16
4	Qualitative examples showing shortcut tokens . . . . .	17

### List of Figures

3	Calibration curves for 1% forget setting. . . . .	14
4	Calibration curves for 5% forget setting. . . . .	15

## A Related Work

### A.1 Machine Unlearning

Unlearning was first explored in recommendation systems and clustering algorithms (Cao and Yang, 2015; Ginart et al., 2019). Eldan and Russinovich (2023) extended unlearning to language models and laid the foundation of approximate unlearning algorithms. Various benchmarks and datasets, including TOFU, WMDP and MUSE, emerged to evaluate unlearning in LLMs (Maini et al., 2024; Li et al., 2024; Shi et al., 2024). However, these benchmarks measured effective removal of forget knowledge while maintaining model utility. However, they do not assess the reliability of unlearned models, which is necessary given that unlearning algorithms are increasingly used in deployed LLMs.

### A.2 Shortcut learning in NLP models

Shortcut learning refers to the tendency of neural models to rely on superficial correlations in the data instead of learning generalizable decision rules. Prior work has shown that NLU models often exploit dataset artifacts, leading to poor robustness under distribution shifts (Du et al., 2021; Wang et al., 2022; Pruthi et al., 2020). Such behavior has been observed across a variety of classification benchmarks, where models learn spurious lexical or syntactic cues rather than semantic features. Recent studies further show that shortcut learning persists in large language models, indicating that scale alone does not eliminate reliance on correlation-based features (Du et al., 2023; Yuan et al., 2024). Mechanistic analyses also suggest that shortcuts can emerge from specific internal features learned during training (Izmailov et al., 2022; Eshuijs et al., 2025).

### A.3 Interpretability and attribution for decision analysis

Understanding the decision rules used by language models often relies on attribution methods that identify influential input tokens. Gradient-based attribution methods such as Integrated Gradients have been widely used to analyze model behavior and detect spurious correlations (Sundararajan et al., 2017; Du et al., 2021). Recent toolkits such as Inseq provide practical frameworks for computing token-level attributions for sequence generation models, enabling interpretability analysis in large decoder-only architectures (Sarti et al., 2023). These methods allow researchers to study whether model pre-

dictions are based on meaningful evidence or shortcut cues.

#### A.4 Calibration and model reliability

Model calibration measures the alignment between predicted confidence and empirical accuracy, and is commonly used as a proxy for reliability in neural models. However, recent work has shown that calibration metrics alone may not reflect the true robustness of model decisions. In particular, models with low calibration error can still rely on non-generalizable features, leading to the reliability paradox, where statistically well-calibrated models make unreliable decisions (Bihani and Rayz, 2024). This suggests that evaluating reliability requires analyzing both confidence estimates and the underlying decision rules.

#### A.5 Reliability and shortcut learning in large language models

Recent work has begun to investigate whether large language models overcome shortcut learning. While scaling improves performance, studies show that LLMs can still rely on spurious correlations and dataset-specific cues, especially in structured evaluation settings (Yuan et al., 2024). These findings motivate the need for evaluation methods that jointly analyze calibration, attribution, and robustness in modern decoder-only models.

Prior work on the reliability paradox focuses on encoder-based classification models and does not consider the setting of machine unlearning, where model parameters are explicitly modified to remove specific training data. In the unlearning scenario, a model may maintain good calibration while internally changing its decision rules, potentially increasing reliance on shortcut cues. This behavior has not been systematically studied for large decoder-only language models.

In this work, we extend the analysis of calibration and shortcut learning to the unlearning setting. Using the RELU evaluation protocol on the TOFU benchmark, we evaluate reliability of pretrained, fully fine-tuned, and unlearned models using both calibration metrics and attribution-based shortcut detection. Our results show that well-calibrated models after unlearning can still rely on correlation-based tokens, demonstrating that the reliability paradox also appears in unlearned language models.

## B Calibration Analysis Across Forget Ratios

In this section, we present reliability diagrams for models evaluated under different forget ratios (1%, 5%, and 10%) in the RELU-MCQA setting. These plots visualize the relationship between model confidence and empirical accuracy, providing insight into probabilistic reliability.

Each figure shows calibration curves for multiple model variants, including pretrained, fully fine-tuned, retained, and unlearned models (e.g., GradAscent, GradDiff, NPO, DPO). The diagonal line represents perfect calibration, where predicted confidence matches observed accuracy. Deviations from this line indicate miscalibration.

Across all forget ratios, we observe that fine-tuned and unlearned models often remain well calibrated, with confidence closely tracking accuracy. However, as shown in the main paper, this apparent calibration does not necessarily reflect reliable decision-making, as these models may still rely on shortcut cues.

Figure 3 shows results for 1% forgetting (see :contentReference[oaicite:0]index=0), while Figures 4 and 2 present results for higher forget ratios.

## C Calibration Analysis across different bin sizes and forget ratios

Both Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) rely on a binning approach: predictions are first partitioned into bins based on confidence, and the gap between accuracy and average confidence is then aggregated within each bin. Consequently, the number of bins  $n$  directly influences the reported error—too few bins coarsen genuine miscalibration, while too many introduce sample-size bias as bins become sparsely populated. In this section, we examine the sensitivity of ECE and MCE to bin size ( $n \in \{10, 20, 33, 100\}$ ) across all forget ratios.

Table 3 shows that on retain splits (99%, 95%, 90%), ECE values remain largely unchanged across  $n = 10, 20,$  and  $33,$  with only a marginal increase at  $n = 100.$  For instance, the fully finetuned model on retain 90% reports ECE values of 0.039, 0.040, 0.041, and 0.051 across the four bin sizes. This stability is expected because the retain splits contain enough samples that even fine-grained binning yields reasonably populated buckets, keeping the binned ECE bias low. In contrast, ECE values on forget splits increase consistently with bin size. For

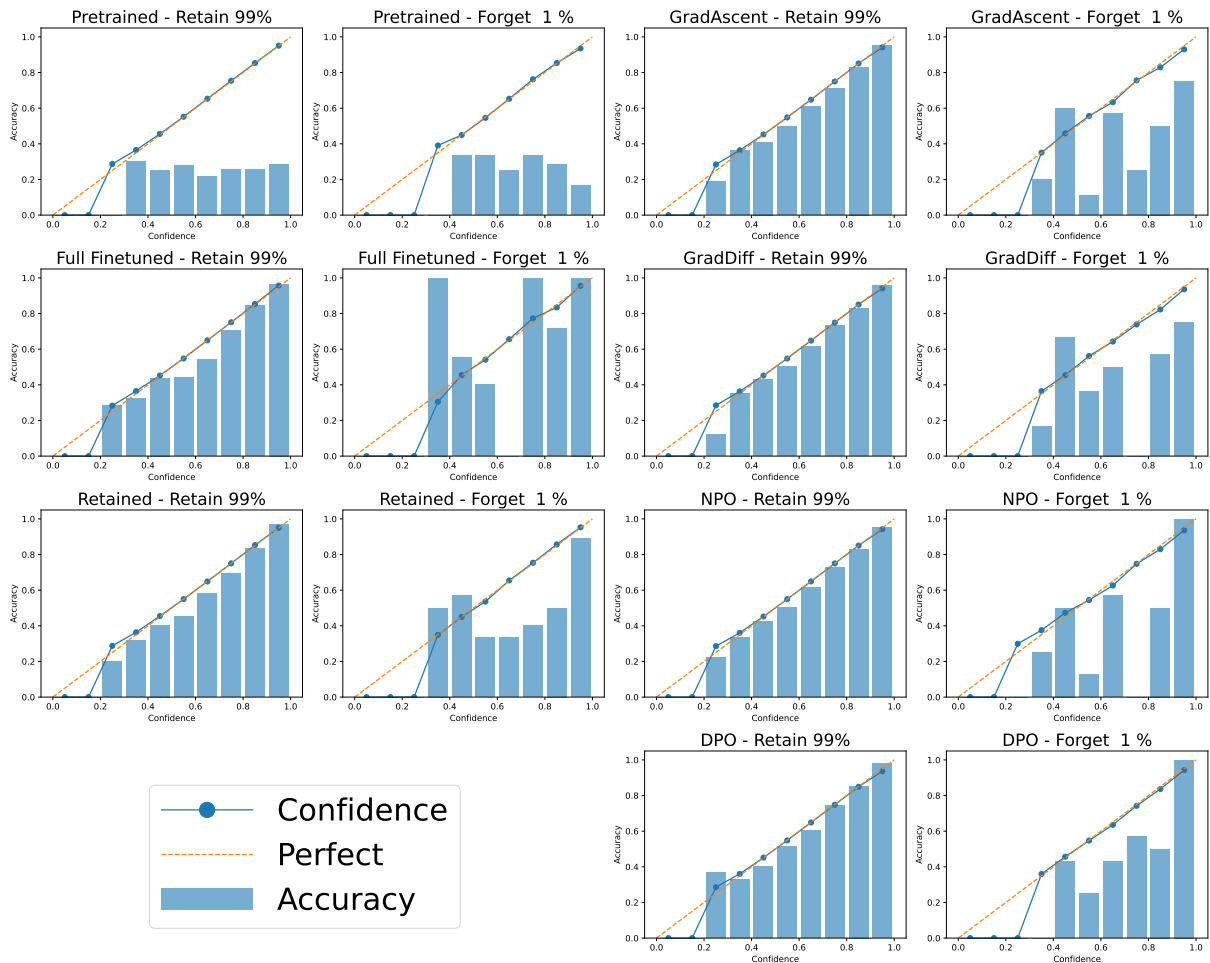


Figure 3: **Calibration curves for 1% forget setting.** Reliability diagrams for pretrained, fully fine-tuned, retained, and unlearned models under 1% forgetting. The diagonal line indicates perfect calibration. While fine-tuned and unlearned models exhibit low calibration error, several models show deviations in mid-confidence regions, suggesting imperfect alignment between confidence and accuracy.

example, on forget 1%, the fully finetuned model’s ECE rises from 0.174 ( $n = 10$ ) to 0.298 ( $n = 100$ ), and similar trends hold across all unlearning methods. We attribute this to the smaller size of the forget splits: as  $n$  grows, individual bins contain fewer datapoints, increasing both the variance and the bias of ECE. This is a well-known artifact of binned calibration estimators on small samples rather than a genuine change in model calibration (Pavlovic, 2025).

MCE shows more volatility across bin sizes than ECE, since a single sparsely populated bin can dominate the maximum. On most retain splits MCE increases monotonically with  $n$ , but the retained model on retain 95% deviates from this trend as the MCE jumps from 0.140 at  $n = 10$  to 0.729 at  $n = 33$ , then drops to 0.464 at  $n = 100$ . This non-monotonic behavior reflects the worst-case nature of MCE: as bins are redistributed, the bin contain-

ing the largest accuracy–confidence gap can shift abruptly, making MCE a less stable summary statistic when bin populations are uneven.

We adopt  $n = 10$  for the main results for two reasons. First, 10 bins yield enough datapoints per bucket on both retain and forget splits to produce a low-bias ECE estimate and is popular in previous studies on calibration estimation (Guo et al., 2017; Pavlovic, 2025). Second, we observe that the primary finding on reliability from the lens of calibration and shortcut usage is preserved across  $n = 10, 20$ , and 33 on the retain splits, indicating that the *reliability paradox* is not an artifact of a particular bin choice. These observations indicate that the calibration error of model before and after unlearning on retain splits is a robust property, irrespective of the bin size.

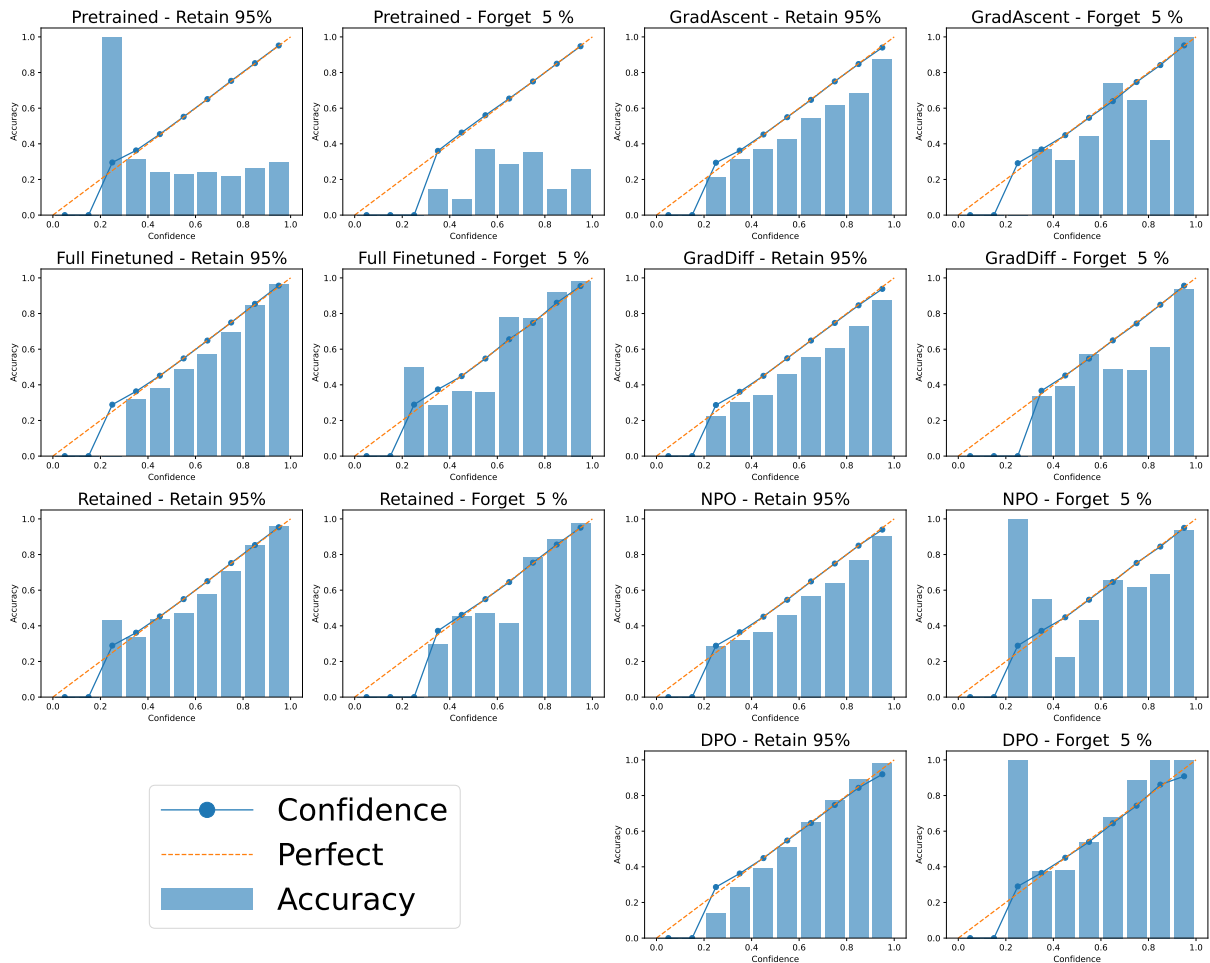


Figure 4: **Calibration curves for 5% forget setting.** Reliability diagrams under increased forgetting. Models maintain relatively good calibration overall, with confidence closely tracking accuracy. However, variability across models increases compared to the 1% setting, particularly in intermediate confidence bins.

## D Qualitative Examples

In Table 4 we present some qualitative examples from a model unlearned using DPO on a 90% retain split ratio. We observe that the model relies on specific words, not pertaining to the author’s identity or the specific question, to make decisions. For example, in the first question, the model relies on “been” to make a decision. Similar things are also observed in the remaining qualitative examples. A more comprehensive list of model attributions, top tokens identified by LMI, and identified shortcuts is present in our GitHub repo <https://github.com/Exploration-Lab/Unlearning-Reliability-Paradox>.

Model	Forget Split								Retain Split							
	ECE ( $\downarrow$ )				MCE ( $\downarrow$ )				ECE ( $\downarrow$ )				MCE ( $\downarrow$ )			
	$n=10$	$n=20$	$n=33$	$n=100$	$n=10$	$n=20$	$n=33$	$n=100$	$n=10$	$n=20$	$n=33$	$n=100$	$n=10$	$n=20$	$n=33$	$n=100$
<b>Forget 1% / Retain 99%</b>																
Pretrained	0.435	0.493	0.488	0.543	0.768	0.930	0.929	0.943	0.519	0.521	0.521	0.521	0.668	0.674	0.674	0.710
Full	0.174	0.204	0.247	0.298	0.695	0.695	0.695	0.857	0.040	0.045	0.046	0.058	0.107	0.119	0.733	0.733
Retained	0.212	0.249	0.286	0.384	0.357	0.827	0.827	0.871	0.047	0.047	0.047	0.055	0.096	0.119	0.270	0.297
GradAscent	0.266	0.344	0.317	0.423	0.506	0.738	0.738	0.843	0.032	0.035	0.038	0.054	0.097	0.097	0.109	0.260
GradDiff	0.239	0.333	0.309	0.418	0.739	0.786	0.786	0.844	0.026	0.032	0.038	0.052	0.164	0.164	0.130	0.285
NPO	0.245	0.279	0.331	0.398	0.748	0.894	0.774	0.894	0.029	0.033	0.040	0.056	0.064	0.075	0.094	0.392
DPO	0.212	0.223	0.259	0.412	0.360	0.525	0.533	0.818	0.033	0.036	0.036	0.048	0.081	0.081	0.269	0.341
<b>Forget 5% / Retain 95%</b>																
Pretrained	0.518	0.525	0.522	0.541	0.707	0.730	0.811	0.993	0.523	0.523	0.523	0.525	0.705	0.705	0.705	0.721
Full	0.082	0.121	0.107	0.176	0.211	0.331	0.668	0.715	0.040	0.042	0.046	0.049	0.289	0.289	0.266	0.297
Retained	0.067	0.089	0.138	0.207	0.232	0.416	0.497	0.825	0.032	0.036	0.037	0.052	0.140	0.140	0.729	0.464
GradAscent	0.129	0.147	0.132	0.213	0.421	0.535	0.578	0.874	0.108	0.108	0.109	0.111	0.165	0.166	0.199	0.286
GradDiff	0.109	0.142	0.149	0.226	0.263	0.424	0.569	0.896	0.100	0.101	0.101	0.107	0.145	0.149	0.738	0.738
NPO	0.128	0.149	0.126	0.211	0.712	0.712	0.712	0.767	0.083	0.083	0.084	0.087	0.112	0.146	0.260	0.260
DPO	0.052	0.086	0.124	0.199	0.710	0.710	0.710	0.710	0.046	0.047	0.051	0.061	0.149	0.149	0.232	0.400
<b>Forget 10% / Retain 90%</b>																
Pretrained	0.497	0.497	0.499	0.511	0.662	0.676	0.724	0.825	0.516	0.516	0.516	0.519	0.665	0.668	0.695	0.738
Full	0.068	0.075	0.089	0.136	0.273	0.273	0.273	0.827	0.039	0.040	0.041	0.051	0.070	0.083	0.272	0.272
Retained	0.083	0.109	0.125	0.168	0.298	0.332	0.327	0.620	0.038	0.039	0.041	0.057	0.081	0.092	0.112	0.305
GradAscent	0.245	0.245	0.248	0.252	0.707	0.707	0.707	0.707	0.209	0.209	0.209	0.209	0.383	0.495	0.557	0.696
GradDiff	0.136	0.136	0.142	0.184	0.285	0.388	0.431	0.905	0.133	0.133	0.134	0.136	0.221	0.221	0.268	0.343
NPO	0.108	0.115	0.120	0.165	0.213	0.258	0.342	0.719	0.101	0.101	0.107	0.111	0.159	0.191	0.180	0.239
DPO	0.046	0.057	0.103	0.139	0.099	0.140	0.227	0.501	0.052	0.052	0.054	0.062	0.089	0.090	0.263	0.276

Table 3: ECE and MCE for Llama-3.1-8B across forget ratios 1%, 5%, and 10%, and bin sizes ranging from 10 to 100. Forget and retain splits are reported side-by-side to enable direct comparison of calibration error behavior on data targeted for unlearning versus data to be preserved.

#	Question & Options	Generated Answer	Ground-truth	Shortcut Tokens		
				Token	Attr.	LMI
1	<p>Has Alejandro Cordero Rodriguez <b>been</b> the recipient of any prestigious awards?</p> <p><b>A.</b> The Silver Scroll Award for Fantasy Fiction  <b>B.</b> The Bronze Book Award for Mystery Novels  <b>C.</b> The Golden Pen Award for Science Fiction  <b>D.</b> The Sapphire Quill Award for Alternate History</p>	D	D	<b>been</b>	0.0112	0.0108
2	<p>How <b>does</b> Andres Santiago Cruz’s family background influence his writing <b>about</b> parenting?</p> <p><b>A.</b> Raised by a Conservationist and a Veterinarian, Andres Santiago Cruz’s writings often include ideas about nurturing and caring, drawn from his formative experiences.  <b>B.</b> Raised by a Lawyer and a Politician, Andres Santiago Cruz’s writings often include ideas about justice and governance.  <b>C.</b> Raised by a Chef and a Musician, Andres Santiago Cruz’s writings often include ideas about creativity and expression.  <b>D.</b> Raised by a Scientist and an Engineer, Andres Santiago Cruz’s writings often include ideas about innovation and problem-solving.</p>	A	A	<b>does</b> <b>about</b>	0.0140 0.0103	0.0578 0.0543
3	<p>How <b>does</b> Catherine Marianne Pfeiffer’s upbringing feature <b>in</b> her writing?</p> <p><b>A.</b> Influenced by her parents’ careers in science and technology, Catherine Marianne Pfeiffer’s narratives often revolve around innovation and progress.  <b>B.</b> Influenced by her parents’ careers in medicine and education, Catherine Marianne Pfeiffer’s narratives often revolve around health and learning.  <b>C.</b> Influenced by her parents’ careers in finance and law, Catherine Marianne Pfeiffer’s narratives often revolve around money and justice.  <b>D.</b> Elements of Catherine’s upbringing, including her parents’ professions and her Toronto roots, often find their way into her philosophical narratives, providing a unique perspective in her work.</p>	D	D	<b>does</b> <b>in</b>	0.0078 0.0066	0.0415 0.0130
4	<p>Why is Yigal Abramovitz known <b>for</b> his humor literature?</p> <p><b>A.</b> Yigal Abramovitz is known <b>for</b> his intense psychological thrillers.  <b>B.</b> Yigal Abramovitz is known <b>for</b> his dark and brooding poetry.  <b>C.</b> Yigal Abramovitz is known <b>for</b> his slapstick comedy in his novels.  <b>D.</b> Yigal Abramovitz had a unique way to infuse humor in everyday-life situations, making them relatable yet hilarious, which became his trademark.</p>	D	D	<b>for</b>	0.0195	0.0025

Table 4: Multiple-choice questions with model-selected answers, ground-truth answers, and the shortcut tokens identified during attribution analysis. Each shortcut token row reports the token string, its attribution, and its LMI (Lexical Mutual Information) score. The identified token is **highlighted in the text**. These examples are extracted from the unlearned model produced by DPO on the retain 90% split. More qualitative examples are present on our github repository.