# SILENT ATTRIBUTION DRIFT: EXPLAINABILITY PRESERVATION UNDER INT8 QUANTIZATION OF ENCODER TRANSFORMERS

Igor Tomić*
[0009-0001-1086-3737]

Faculty of Technical Sciences,
Čačak, Serbia

Abstract:

Deep learning models based on transformer encoder architectures, when used in production environments, require substantial computational resources. To address these challenges, various methods of compression have been developed, including quantization. Quantized models are often evaluated only on accuracy, while preservation of explainability is implicitly assumed; whether explainability is maintained at the token-attribution level, however, has not yet been systematically assessed.

In this paper, we introduce the silent attribution drift effect, a phenomenon in which the quantized model faithfully reproduces the predictions of the original model but generates fundamentally different explanations based on the importance of the input tokens. The results reveal a pronounced discrepancy between the stability of predictions and the stability of explanations. While prediction accuracy matching reaches 96-100%, the Spearman rank correlation of attributions drops, which shows a change in the distribution of token importance.

We establish that the robustness of attributions depends on the architecture and complexity of the task. Four-class classification shows increased deviation compared to binary classification. We propose logit shift analysis as a diagnostic tool that mechanistically explains the sources of deviation, revealing that quantization changes the model's sensitivity to individual tokens. Our findings indicate that evaluating compressed models based solely on accuracy is not sufficient in domains where explainability is a regulatory requirement or ethically necessary.

Keywords:

Quantization, Transformers, Silent Attribution Drift, Logit Shift Analysis.

## INTRODUCTION

Transformer-based encoder models are increasingly deployed in production, where computational constraints make quantization a standard compression technique [1]. The goal of quantization is to reduce model size and improve speed to reduce computational cost. One of the quantization techniques is dynamic INT8 quantization, which converts 32-bit floating-point weights to 8-bit integers at inference time. This quantization method is widely adopted due to its simplicity and the minimal accuracy loss it introduces [2].

Correspondence:
Igor Tomić

e-mail:
itomic410@gmail.com





DOI: 10.15308/Sinteza-2026-191-197

However, current evaluation of quantized models focuses almost exclusively on prediction accuracy [3]. Whether quantization influences the model explainability at the token-level has not been systematically studied. Singh et al. [3] examine how quantization affects neuron-level attributions in decoder large language models using Layer Integrated Gradients [4], analysing neuron salience, redundancy, and model calibration. Their findings suggest that quantization has generally minor effects at the neuron level and remains a reliable compression technique.

Our work reveals a different picture at a token attribution level, where the same quantization that preserves neuron-level behaviour can produce different explanations of model decisions. We investigate the effect of dynamic 8-bit integer quantization on token attributions for three encoder architectures (BERT [5], DistilBERT [6], RoBERTa [7]) across two classification tasks (SST-2 [8], AG News [9]) using two perturbation-based attribution methods (occlusion [10], leave-one-out [11]).

## 2. BACKGROUND

### 2.1. NEURAL NET QUANTIZATION

Quantization is a model compression technique that reduces the numerical precision of parameters and/or neural network activations [1]. Post-training quantization (PTQ) is usually used for transformers because it does not require new training of the model, and it is applied directly to the trained model. There are two basic 8-bit integer quantization approaches:

**Dynamic quantization** converts model weights from 32-bit floating point (FP32) to 8-bit integer (INT8) format, while weights are quantized on inference. Quantization range is calculated for each batch. This method is the most widely used method in production environments.

**Static quantization** additionally quantizes the activations of the model, whereby the range of quantization is calculated based on the calibration data. This makes inference faster because both weights and activations are stored in INT8 format, but it requires a proper calibration dataset and careful configuration.

In both cases, we can define quantization as:

$$x_q = round(x / s) + z$$

**Equation 1.** Quantization definition

where $x$ is the original FP32 value, $s$ scaling factor, $z$ zero point, and $x_q$ resulting INT8 value. This procedure inevitably introduces round-off error, which, although small for individual values, can cumulatively affect the behaviour of the model.

### 2.2. MODEL EXPLAINABILITY AND TOKEN-LEVEL ATTRIBUTION

Explainability of the model refers to the ability to understand and interpret the decisions made by the model [12]. Each input token gets a numerical value that represents its contribution to the model prediction. There are two main approaches to quantifying attribution:

**Gradient-based methods**, such as Integrated Gradients [4], and GradCam [13], calculate attribution of each input token by following the gradient through the neural net. These methods require differentiable operations and support for autograd, which makes them incompatible with fully quantized PTQ models [14].

**Perturbation methods** do not require gradients. Instead, each token is removed individually (replaced by a neutral token), and token attribution is calculated as the difference in model output. Occlusion [10] calculates changes in raw logit for the whole class, while Leave-one-out [11] calculates the change in probability of the whole class after the softmax function. Given that these methods require only a forward pass, they are fully compatible with quantized models.

## 3. METHODOLOGY

### 3.1. MODELS

In experiments, we use three encoder transformer architectures that are fine-tuned for text classification:

**BERT-base** [5] is an original bidirectional transformer encoder with 110 million parameters and 12 layers. We use a version that is fine-tuned on the TextAttack dataset [15].

**DistilBERT** [6] is a compressed version of BERT that was created with a knowledge distillation process [16]. It has 66 million parameters and 6 layers, retaining 97% performance of BERT with 40% fewer parameters. For the SST-2 version, we use the official HuggingFace checkpoint [17], while for AG News, we use the TextAttack version [15].

**RoBERTa** [7] is an optimized BERT variant with improved pre-training, including larger batch, longer training, and removal of the next sentence prediction task. It contains 125 million parameters and 12 layers.

Using these three architectures enables us to compare the quantization effect on models of different sizes (66M–125M parameters), depths (6-12 layers), and training processes, giving us insights into whether attribution robustness to quantization is specific to a particular model or a general pattern.

### 3.2. DATASETS

We conduct experiments on two standard datasets for text classification, which differ in the number of classes and text length:

**SST-2 (Stanford Sentiment Treebank)** [8] is a sentiment binary classification dataset that includes film reviews that are labelled as positive or negative. We use a validation dataset of 872 examples. Average text length is 19 tokens.

**AG News** [9] is a dataset made for news classification in 4 classes: World, Sports, Business, and Sci/Tech. We use a validation dataset of 7600 examples. Average text length is 38 tokens, which is double the size of the SST-2 text length.

The use of two datasets of different complexity (binary vs. four-class classification) and text length enables the evaluation of findings.

### 3.3. ATTRIBUTION METHODS

Given that post-training quantized INT8 models operate on integer arithmetic and do not support gradient propagation, we use two perturbation methods that require only a forward pass:

**Occlusion** [10]: For each position of a token $i$ in the input sequence, the original token is replaced by a special PAD token. Let $l_c(x)$ denote the logit of the target class c for the original input $x$, and $l_c(x\setminus_i)$ ) denote the logit for the input with the masked token at position $i$. Token importance is then calculated as the absolute logit difference, normalized by the maximum value in the sequence.

$$a_i^{occ} = \left| l_c(x) - l_c(x\setminus_i) \right|$$

**Equation 2.** Occlusion

**Leave-one-out** [11]: Identical procedure as occlusion, but instead of the raw logit, the probability of the target class is used after applying the softmax function $p_c$, making this metric more sensitive to changes in the high probability area.

$$a_i^{lcc} = \left| p_c(x) - p_c(x\setminus_i) \right|$$

**Equation 3.** Leave-one-out

For both methods, special tokens are excluded from the analysis because they do not carry semantic content.

### 3.4. COMPARISON METRICS

For a quantitative comparison of the attribution vectors $a^{fp32}$ and $a^{int8}$ obtained from the FP32 and INT8 models for the same input text, we use three complementary metrics:

Cosine similarity (Equation 4) quantifies the similarity between attribution vectors independent of their magnitude, focusing on the shape of the token importance distribution. Resulting values range from [-1, 1], where 1 indicates an identical pattern of attributions.

$$cos(a,b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

**Equation 4.** Cosine similarity

Spearman rank correlation (Equation 5) measures the consistency of token importance rankings between two models, where $d_i$ is the difference in the ranks of token $i$ between the two models, and $n$ is the number of tokens. A value of 1 indicates identical ranking, 0 indicates the absence of correlations.

$$p_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

**Equation 5.** Spearman rank correlation

Top-k overlap measures the proportion of overlap between the top $k$ tokens identified by the FP32 and INT8 models.

$$TopK = \frac{\left| T_k^{fp32} \cap T_k^{int8} \right|}{k}$$

**Equation 6.** Top-k overlap

This answers whether both models highlight the same tokens as key to the decision. In the experiments, we use $k$=3, which corresponds to the three most prominent tokens in the explanation.

### 3.5. LOGIT SHIFT ANALYSIS

In addition to comparing final attributions, we introduce logit shift analysis as a mechanistic tool to explain why attributions diverge. For each token at a position i, we measure the sensitivity of both models (Equation 7) to the removal of that token.

$$\ddot{a}_i^m = l_c^m(x) - l_c^m(x \setminus_i),\ m \in \{FP32, INT8\}$$

**Equation 7.** Token sensitivity

We then compare the sensitivity vectors $\delta^{fp32}$ and $\delta^{int8}$ using the following three complementary measures:

Sensitivity correlation ($\rho_\delta$) is Spearman correlation between vectors $\delta^{fp32}$ and $\delta^{int8}$. A high correlation indicates that both models have a similar sensitivity map or that the same tokens cause similar changes in output. A low correlation indicates that quantization fundamentally changed the way the model reacts to individual tokens.

$$\overline{\Delta} = \frac{1}{n}\sum_i \left|\delta_i^{fp32} - \delta_i^{int8}\right|$$

**Equation 8.** Mean absolute offset

Mean absolute offset (Equation 8) measures the average difference in sensitivity per token.

$$\Delta_b = \left|l_c^{fp32}(x) - l_c^{int8}(x)\right|$$

**Equation 9.** Base logit difference

Base logit difference (Equation 9) measures how much the starting logit of the target class differs between the two models before any perturbations. A large base difference indicates a global change in the output space of the model.

### 3.6. EXPERIMENTAL PROTOCOL

For each combination of model and dataset, the FP32 model is loaded and quantized to INT8 using PyTorch dynamic quantization [18] [14]. From each dataset, 200 examples are selected where the FP32 model shows high reliability (target class probability ≥ 0.5). For each example, attributions of both methods (occlusion and leave-one-out) are calculated on both models, comparison metrics are computed for each pair, and a logit shift analysis is conducted. Results are aggregated and presented as mean ± standard deviation. All experiments are reproducible with a fixed random seed of 42.

## 4. RESULTS

### 4.1. ATTRIBUTION DRIFT

As a starting point, we verify that dynamic INT8 quantization preserves prediction accuracy: all three models maintain 96–100% prediction match on both datasets, confirming that quantization effectively preserves classification accuracy [1] [2]. However, as Table 1 shows, high prediction accuracy does not imply conservation of model explanations.

Results reveal significant attribution drift despite a high match of predictions. Cosine similarity is moderately high, which indicates that the shape of the distribution is partially preserved. Spearman rank correlation is low for all architectures, which indicates that the order of tokens by importance changes after quantization. Top-k overlap confirms the same pattern, with on average one to two of the top three tokens matching between FP32 and INT8 models.

Both attribution methods provide consistent results, confirming that the deviation reflects a real change in the behaviour of the quantized model. Deviation is consistently greater on the AG News data set relative to SST-2 for all models and all metrics.

This indicates that the complexity of the task affects the robustness of the attributions: AG News requires classification into four classes based on longer texts (on average 38 tokens vs. 19 for SST-2), which makes the decision surface more sensitive to perturbations caused by quantization. RoBERTa generally shows the most stable attributions, while BERT shows the weakest robustness. DistilBERT varies on SST-2, but it performs slightly better than RoBERTa, while on AG News, it lags behind.

### 4.2. LOGIT SHIFT ANALYSIS

Table 2 shows the results of the logit shift analysis, which provides mechanistic insight into the causes of attribution bias.

The analysis reveals a clear mechanism behind the attribution bias. Models that show a larger base logit difference also indicate lower sensitivity correlation and larger attribution shift in the main results. This pattern is consistent in both datasets, which indicates that quantization not only changes the absolute values of the model output but also the relative sensitivity of the model to individual tokens. The sensitivity correlation follows the same pattern, which means that the model's sensitivity to individual tokens is significantly changed after quantization.

**Table 1.** Attribution Drift on FP32 and INT8 Models (Average Value ± std, N=200)

| Dataset | Model | Method | Cosine | Spearman | Top-3 | Accuracy |
|---|---|---|---|---|---|---|
| SST-2 | BERT | Occlusion | 0,718 ± 0,158 | 0,222 ± 0,266 | 0,407 ± 0,283 | 98% |
| SST-2 | BERT | Leave-one-out | 0,734 ± 0,185 | 0,249 ± 0,256 | 0,472 ± 0,271 | 98% |
| SST-2 | DistilBERT | Occlusion | 0,807 ± 0,142 | 0,341 ± 0,259 | 0,533 ± 0,269 | 96% |
| SST-2 | DistilBERT | Leave-one-out | 0,798 ± 0,212 | 0,363 ± 0,271 | 0,548 ± 0,279 | 96% |
| SST-2 | RoBERTa | Occlusion | 0,800 ± 0,144 | 0,300 ± 0,273 | 0,527 ± 0,291 | 97% |
| SST-2 | RoBERTa | Leave-one-out | 0,789 ± 0,196 | 0,320 ± 0,266 | 0,548 ± 0,277 | 97% |
| AG News | BERT | Occlusion | 0,640 ± 0,132 | 0,076 ± 0,169 | 0,267 ± 0,291 | 97% |
| AG News | BERT | Leave-one-out | 0,631 ± 0,182 | 0,100 ± 0,166 | 0,290 ± 0,278 | 97% |
| AG News | DistilBERT | Occlusion | 0,576 ± 0,150 | 0,102 ± 0,165 | 0,222 ± 0,261 | 99% |
| AG News | DistilBERT | Leave-one-out | 0,581 ± 0,189 | 0,110 ± 0,166 | 0,273 ± 0,268 | 99% |
| AG News | RoBERTa | Occlusion | 0,762 ± 0,129 | 0,182 ± 0,239 | 0,343 ± 0,260 | 100% |
| AG News | RoBERTa | Leave-one-out | 0,756 ± 0,152 | 0,184 ± 0,224 | 0,372 ± 0,263 | 100% |

**Table 2.** Logit Shift Analysis (N=200)

| Model | Dataset | Sensitivity Correlation | Mean absolute offset | Base Logit Difference |
|---|---|---|---|---|
| BERT-base | SST-2 | 0,479 ± 0,208 | 0,275 ± 0,166 | 1,430 ± 0,450 |
| BERT-base | AG News | 0,209 ± 0,185 | 0,287 ± 0,198 | 1,440 ± 1,044 |
| DistilBERT | SST-2 | 0,567 ± 0,183 | 0,357 ± 0,337 | 0,447 ± 0,528 |
| DistilBERT | AG News | 0,256 ± 0,182 | 0,321 ± 0,258 | 0,778 ± 0,735 |
| RoBERTa | SST-2 | 0,531 ± 0,192 | 0,311 ± 0,346 | 0,246 ± 0,423 |
| RoBERTa | AG News | 0,446 ± 0,181 | 0,102 ± 0,164 | 0,081 ± 0,134 |

These results explain the robustness pattern observed in the main results: models exhibiting smaller logit shifts after quantization (RoBERTa) retain more consistent attributions, while models with larger logits shift (BERT) show significant deviation in token rankings.

To illustrate the practical impact of this mechanism, we examine selected examples with the worst attribution drift, i.e., cases where prediction matches but top tokens are completely different. For each dataset, we identify the worst case per architecture: the correctly classified sample where the Spearman correlation between FP32 and INT8 attributions (occlusion) is lowest. The FP32 model highlights semantically relevant words (comedy, surprising, spirit, telescope), while the INT8 model highlights function words or sub-word fragments without semantic content. This illustrates the core of silent drift: a user relying on explanations from a quantized model receives misleading insight into the model's reasoning.

### 4.3. CONFIDENCE INTERVAL ANALYSIS

We investigate whether higher confidence of the FP32 model protects against attribution drift. The results for RoBERTa on AG News (occlusion) reveal a counterintuitive pattern: examples with the highest confidence of the FP32 model (99%+, N=170) show the worst preservation of attributions (Spearman = 0.170), while examples with moderate confidence (80–90%) have a higher correlation (Spearman = 0.289). We observe a similar pattern on SST-2 and other architectures. On the contrary, a model can be both very confident in its decision and completely inconsistent in reasoning.

## 5.DISCUSSION

Logit shift analysis provides mechanistic insight into the causes of attribution divergence. Models with lower base logit difference after quantization consistently show more stable attributions. This shows that quantization not only changes the absolute output values but also affects the model's relative sensitivity to specific tokens.

Confidence interval analysis shows that samples with the highest confidence in FP32 models show the worst attribution preservation, indicating that higher prediction confidence does not correspond to more reliable explanations. A possible explanation is that high confidence samples lie deep within the decision boundary, where the model uses a specific combination of characteristics that are sensitive to quantization perturbations.

The practical consequence is that precisely those samples that the user would not question bear the greatest risk of being incorrectly explained. Our results do not question the use of quantization as a compression technique. A prediction match of 96-100% confirms that the quantized models faithfully reproduce the decisions of the original model.

Future work should explore compression methods that preserve the consistency of attributions at the same level as they preserve the accuracy of predictions, including statistical and mixed-precision quantization, extensions to generative models, and the application of gradient-based methods to dequantized weights.

### 5.1. LIMITATIONS

Our work has several limitations. First, we exclusively use dynamic INT8 quantization through PyTorch; static quantization and more aggressive formats (INT4, mixed-precision) remain for future research. Second, the experiments are limited to encoder models. Third, we use perturbative attribution methods because real INT8 quantization does not support gradients; methods based on gradients (i.e., Integrated Gradients) could give different results on dequantized models.

## 6. CONCLUSION

The results of this research reveal a fundamental discrepancy between two aspects of the reliability of quantized models that have been implicitly equated so far: accuracy of prediction and consistency of explanation. Dynamic INT8 quantization faithfully preserves the model's decisions, but it significantly alters the way the model explains its decisions. This is what we call silent attribution drift.

Our experiments yield several conclusions with wide implications. Explainability robustness is an inherent attribute of architecture, and not a random outcome. Models with more stable output space, such as RoBERTa, consistently preserve attributions more effectively. Complex classification tasks increase divergence of attributions, which suggests that decision boundaries in multiclass problems are more sensitive to perturbations caused by quantization. The proposed logit shift analysis provides mechanistic insight into these differences, linking the stability of the model's baseline output to the preservation of token rankings.

Of particular note is the finding that high prediction confidence does not protect explanations. On the contrary, the most confident samples show the greatest instability of attributions. This goes against the intuition that a model whose prediction is the most accurate should also be the most explainable, and that has direct implications for usage in regulated or ethically sensitive domains.

Future explorations should consider alternative quantization methods, expansion on generative models, as well as the development of compression methods that explicitly optimize the preservation of explanations. We conclude that accuracy metrics, however high they may be, are not a sufficient guarantor of the reliability of a quantized model in the context of explainability. Validation of explanations must become an integral part of the evaluation of compressed models.

## REFERENCES

[1] A. Gholami, S. Kim, Z. Dong, Z. Yao, K. Keutzer and M. W. Mahoney, "A Survey of Quantization Methods for Efficient Neural Network Inference," in *Low-Power Computer Vision*, Chapman and Hall/CRC, 2022, p. 291–326.

[2] O. Zafrir, G. Boudoukh, P. Izsak and M. Wasserblat, "Q8BERT: Quantized 8Bit BERT," in *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, Vancouver, BC, Canada, 2019.

[3] M. Singh and H. Sajjad, "Interpreting the Effects of Quantization on LLMs," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Mumbai, India, 2025.

[4] M. Sundararajan, A. Taly and Q. Yan, "Axiomatic attribution for deep networks," in *ICML'17: Proceedings of the 34th International Conference on Machine Learning* - Volume 70, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT* , 2019 .

[6] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," *arXiv preprint arXiv:1910.01108* , 2019.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *arXiv preprint arXiv:1907.11692*, 2019.

[8] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of EMNLP*, 2013.

[9] X. Zhang, J. Zhao and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

[10] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision (ECCV 2014)*, 2014.

[11] J. Li, W. Monroe and D. Jurafsky, "Understanding Neural Networks through Representation Erasure," *arXiv preprint arXiv:1612.08220*, 2016.

[12] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2022 .

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[14] PyTorch, "Quantization — PyTorch Documentation," 2024. [Online]. Available: https://pytorch.org/docs/stable/quantization.html. [Accessed 18 March 2026].

[15] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin and Y. Qi, "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP," in *Proceedings of EMNLP: System Demonstrations*, 2020.

[16] G. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu and C. Xu, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of EMNLP: System Demonstrations*, 2020 .

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani and Chila, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems 32 (NeurIPS) , 2019.