Does a Reduced Fine-Tuning Surface Impact the Stability of the Explanations of LLMs?

Jeremie Bogaert[1*] & François-Xavier Standaert[1*]

[1] Crypto Group, ICTEAM Institute, UCLouvain; Louvain-la-Neuve, Belgium.

Abstract. Explainability is an increasingly demanded feature for the deployment of LLMs. In this context, it has been shown that the explanations of models that are equivalent from the accuracy viewpoint can differ due to their training randomness, leading to a need to characterize the explanations' distribution and to understand the origin of this sensitivity. In this paper, we investigate whether the fine-tuning surface, defined as the number of bits that are fine-tuned in a LLM, can serve as a good proxy for the stability of its explanations. We answer negatively and show that two different approaches for reducing the fine-tuning surface, namely quantizing and freezing (a part of) the models, lead to very different outcomes.

1 Introduction

The progresses of Large Language Models (LLMs) create increasing challenges for their training and understandability. On the training side, current LLMs are so complex that only a few actors have the resources to pre-train them. Pre-trained models can transform texts into relevant yet general embeddings, which are then fine-tuned to specific tasks [6]. On the explainability side, various criteria have been introduced in the literature, but their evaluation can be challenging and a formal analysis of their interplay is missing [10]. In this paper, we are concerned with the sensitivity of the explanations of LLMs to the randomness used in their training, put forward in [3], and its link with the models' fine-tuning surface, defined as the amount of fine-tuned bits in a LLM.

The starting observations made in [3] are twofold. First, it is possible to train LLMs with different random seeds in order to produce models that are equivalent from the accuracy viewpoint. Second, the explanations of these equivalent models, for example produced with the Layerwise Relevance Propagation (LRP) method [5], can lead to very different explanations. This raises a need to characterize the sensitivity to the training randomness of LLM's explanations. At the very least, it is for example necessary to verify that the distribution of the explanations corresponding to indistinguishable seeds differs sufficiently from the uniform (which would make the selection of an explanation completely arbitrary). Box-plots provide a simple and visual tool for this purpose. Metrics like the explanations' signal, noise and signal-to-noise ratio can serve as quantitative surrogates [4], although defining the right metric to capture the sensitivity of LLM's explanations to the training randomness remains an open problem.

* Work supported by the Service Public de Wallonie Recherche, grant n°2010235-ARIAC by DIGITALWALLONIA4.AI. FXS is Senior Research Associate of the F.R.S.-FNRS.

In this context, the main research question we tackle is whether a model's fine-tuning surface is a good proxy for the sensitivity to the training randomness of its explanations. That is, we aim to assess whether this sensitivity depends of the amount of information (quantified in bits) modified inside the model during the fine-tuning. We consider two options for this purpose. One is to simplify the models by quantizing their weights [7, 8]. The second is to freeze a part of the models so that they are not subject to fine-tuning. We answer negatively and show that the impact of these two options can be very different for a case study of opinionated journalistic text classification in French. We then discuss the consequence of our empirical findings for the explanability of LLMs.

2 Background

2.1 Dataset and machine learning models

We run our experiments on the InfOpinion dataset presented in [2]. It contains 10,000 news and was built to train and evaluate a classification model distinguishing between texts belonging to the journalistic *opinion* genre (editorials, commentaries, reviews, ...) from texts belonging to the *information* genre (press agency dispatches, news articles, ...) This binary categorization relies solely on the articles' annotation by their authors as either *opinion* or *information*. The dataset is split in 3 parts: a training set (80%), a validation set (10%) and a test set (10%). The task is to predict the binary category of a given text.

We consider two types of models for this purpose: fully fine-tuned ones and partially fine-tuned ones, that we next denote as frozen. Both are based on the CamemBERT French pre-trained transformer model [11]. In the first case, we fine-tune the model on the training set during 2 epochs, as presented in [6]. In the second case, we first probe the internal representation of each text from the training set at the last layer of the pre-trained model without retraining it. As in [12], we then use the first token's representation as a text embedding and train a RoBERTa classification head during 20 epochs on top of these embeddings. This process allows us to reduce the fine-tuning surface by reducing the amount of parameters affected by the fine-tuning. Note that the randomness used by the frozen and fine-tuned models can be controlled via a seed parameter. This seed rules the initialization of the layers, the order of the training dataset, and the neurons that are deactivated by the dropout layers during training.

These two types of models can be quantized, which consist in reducing the size of some of the model's parameters or weights [7, 8]. This process reduces the fine-tuning surface by lowering the amount of bits used by some of the parameters (mainly in the linear layers of the model). We restrict our study to post-training quantization without quantization-aware training (i.e., we don't retrain our models after quantization). The two most common ways of quantizing are to use 8-bit or 4-bit floating point parameters (FP8 or FP4) and to use 4-bit NormalFloat (NF4). We use the FP4 and NF4 techniques in the paper, as implemented in the bitsandbytes library (https://github.com/TimDettmers/bitsandbytes). It follows that we consider 6 types of models: the finetuned and the frozen ones, used either with no quantization or with a FP4 or a NF4 quantization.

2.2 Equivalent models and explanations' stability

To study the impact of the training randomness on the explanations of equivalent models, we train each model for 200 different random seeds, like in [3, 4]. As a result, we obtain 200 versions of fine-tuned and frozen transformer models. Each of these versions can be quantized using the FP4 or NF4 technique. The accuracy is then evaluated on a test set of $1,000$ news ($n = 1,000$). We finally select a subset of m most accurate models such that the difference between the best (a) and worst (b) accuracies of the models in the subset is not statistically significant. For this purpose, we compute the Z statistic [9], which can detect whether two proportions (here, the accuracies a and b) are different:

$$z = \left| \frac{a - b}{\sqrt{\frac{\frac{a+b}{2} * (1 - \frac{a+b}{2})}{n}}} \right| \tag{1}$$

We next consider that z values greater than 1.96 ($p < 0.025$) mean that the accuracies of the best and the worst models in a subset are different. For lower values of the z statistic, we conclude that these accuracies do not differ significantly and therefore, we denote the models in the subset as equivalent from the performance viewpoint. Concretely, starting from 200 models, a restriction to $m = 100$ was sufficient to reach model equivalence in the subset.

The sensitivity of the explanations to the training randomness can be represented visually with box-plots. Alternatively, [4] proposes a quantification based on the explanations's signal and noise. The signal is defined as the variance (across the words of a text) of the attention means (across the seeds). The noise is defined as the mean (across the words of a text) of the attention variances (across the seeds). The Signal-to-Noise Ratio (SNR) is defined as their ratio.

2.3 Explainability method

Once models are trained, we use the Layer-wise Relevance Propagation (LRP) method to generate word-level explanations for every text [1]. It works by back-propagating the relevance from the last layer of the network using conservation constraints, so that the relevance of each neuron is redistributed to the neurons of the previous layer based on their respective gradient. This principle is then followed through the whole network up to the input layer in order to get word-level explanations.[1]. As the constraints are more difficult to satisfy for some layers in the models, this method can be improved with additional rules. In this paper, we used an improved version from [5] and next refer to it as LRP.

3 Accuracy and fine-tuning surface analysis

From Table 1, we can first see that reducing the amount of trainable bits of the models, whether it is by freezing some of the parameters or by quantizing

[1] Since CamemBERT uses the roBERTa tokenization to work with word pieces, we post-process LRP explanations to obtain one explanation per word instead of one per word piece.

	Avg. accuracy	Fine-tuning surface
Ft	0.960 ± 0.6	3.5×10^9
Frozen	0.960 ± 0.3	1.8×10^7
Ft_q(FP4)	0.967 ± 0.5	7.4×10^7
Ft_q(NF4)	0.961 ± 0.6	7.4×10^7
Frozen_q (FP4)	0.952 ± 0.4	2.4×10^6
Frozen_q (NF4)	0.951 ± 0.5	2.4×10^6

Table 1: Accuracy and fine-tuning surface of the 100 equivalent models.

some of the weights, does not lead to a significant drop in accuracy. This is rather surprising, as we divide the fine-tuning surface by approximately 1000 when going from the fully-fine-tuned model to the quantized frozen ones.

4 Explanations' sensitivity to the training randomness

We first illustrate the distribution of the attention that LRP assigns to each word of a given text in our subset of equivalent models. Figure 1 gives this distribution with box-plots, where wider boxes reflect a higher variability of the explanations corresponding to different random seeds. The two bottom (resp., upper) rows display the sensitivity to the training randomness of the fine-tuned (resp., frozen) model and its quantized version. We can first observe that the different models, despite their very close accuracy, do not give attention to the same words on average. For example, the frozen models tend to look more on words 2 and 4 (Policiers/Policeman and Chargé/Rushed) than the fine-tuned models, while the fine-tuned models tend to look more at the end of the text (Selon les journalistes/According to the journalists). We can also notice that the box-plots of the frozen model and its quantized version are thinner than the ones of the fine-tuned models. Finally, the quantization has little impact on the top words considered by the models. However, even if it lowers the fine-tuning surface, it tends to increase the box sizes, in particular for the frozen model.

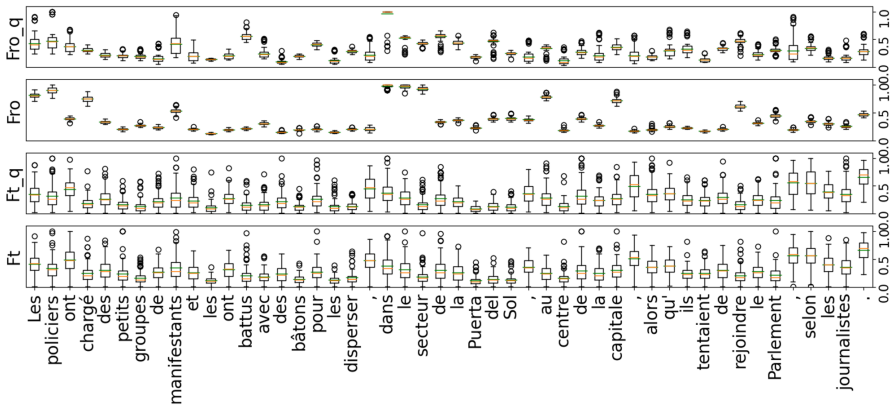

Fig. 1: Boxplot for a short text (FP4 quantization).

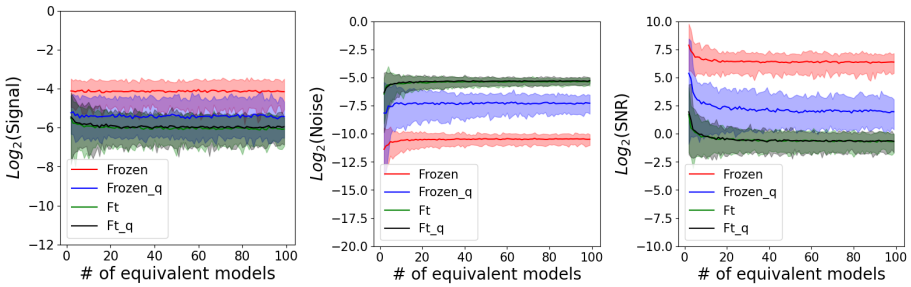

Fig. 2: Signal, noise and SNR for a short text (FP4 quantization).

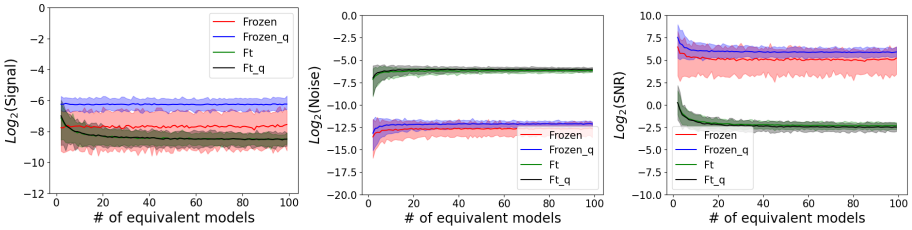

Fig. 3: Signal, noise and SNR for a long text (FP4 quantization).

We confirm these intuitive observations by estimating the explanations' signal, noise and SNR metrics of [4] for a short and a long text in Figures 2 and 3. For readability, we only reports results for FP4 quantization (results for NF4 are similar). We first note that there is no statistically significant differences for the signal (i.e., the amount of statistical information in the explanations) with different FTS. We next observe that the noise (i.e., the explanations' sensitivity to randomness) is not reduced by the models' weights quantization. This is presumably because it does not bridge the gap between the complexity of the models and the simplicity of the (LRP-based) explanations we consider. That is, despite using quantized weights, fine-tuned models probably exploit features of the texts to classify that hardly reduce to word-level explanations [4]. By contrast, freezing a part of the models significantly reduces the explanations' sensitivity to randomness. We discuss the impact of this approach in conclusions.

5 Conclusion and further works

The sensitivity of the explanations of LLMs to their training randomness leads to new challenges for their trustworthy use. While it may not always be detrimental to their explainability (e.g., if this sensitivity was shown to emulate a variability found for human annotators), it requires characterization efforts that are currently not standard and raises questions on the origin of this sensitivity and how to deal with it. In this paper, we contribute to this issue by investigating whether the fine-tuning surface is a good proxy for this sensitivity.

Our results provide empirical evidence that it is not. On the one hand, simplifying the models by quantization does not reduce the explanations' sensitivity to the training randomness. This raises the question whether other model simplifications, or other explanation methods, could lead to different results. On the other hand, freezing the models significantly reduces this sensivity. However, such a freezing is hiding the sensitivity issue more than it is impacting it. It takes the frozen part of the model as a ground truth whereas it is also the outcome of a pre-training, the explainations of which are possibly affected by an equally high sensitivity to the training randomness (even harder to characterize due to computational reasons). It should therefore be viewed as a possible separation of duties, which only makes sense if the frozen part of the model used to generate text embeddings is itself explainable via other (e.g., linguistic) means.

Eventually, and from a more methodological viewpoint, our results are dependent on a definition of signal and noise that is connected to simple (word-level, first-order, univariate) explanations [4]. The questions whether the explanations' variance due to random seeds only corresponds to noise (as defined) or whether a more semantic definition of signal (e.g., focused on fewer words) could help refining our conclusions therefore remain an interesting research direction.

References

[1] S. Bach et al., On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PloS one, 2015.

[2] J. Bogaert et al., TIPECS: A Corpus Cleaning Method Using Machine Learning and Qualitative Analysis, International Conference on Corpus Linguistics, 2023.

[3] J. Bogaert et al., Sensibilité des Explications à l'Aléa des Grands Modèles de Langage: le Cas de la Classification de Textes Journalistiques (Sensitivity of Explanations to the Randomness of Large Language Models: a Case Study on Journalistic Text Classifications, *Preprint*, to appear in Traitement Automatique des Langues, 2024.

[4] J. Bogaert et al., A Question on the Explainability of Large Language Models and the Word-Level Univariate First-Order Plausibility Assumption, in ReLM, 2024.

[5] H. Chefer et al., Transformer Interpretability Beyond Attention Visualization, in CVPR, 2021.

[6] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in NAACL-HLT, 2019.

[7] T. Dettmers et al., Qlora: Efficient Fine-Tuning of Quantized LLMs, in NEURIPS, 2024.

[8] B. Jacob et al., Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference, in CVPR, 2018.

[9] E. Lehmann et al., Testing Statistical Hypotheses, Springer, 1986.

[10] Q. Lyu et al., Towards Faithful Model Explanation in NLP: A Survey, in Computational Linguistics, 2024.

[11] L. Martin et al., CamemBERT: a Tasty French Language Model, in ACL, 2020.

[12] M. E. Peters et al., To Tune or Not to Tune? Adapting Pre-Trained Representations to Diverse Tasks, in Proceedings of the 4th Workshop on Representation Learning for NLP, 2019.