Consolidating Explanation Stability Metrics

Jeremie Bogaert[1], Antonin Descampe[2], and François-Xavier Standaert[1]

[1] Crypto Group, ICTEAM Institute, UCLouvain, Belgium
[2] Crypto Group, ILC Institute, UCLouvain, Belgium
e-mail: jeremie.bogaert@uclouvain.be

Abstract. The explanations of large language models (e.g., where each word is assigned a relevance score) have recently been shown to be sensitive to the randomness used during model training, creating a need to evaluate this sensitivity. While simple visualization tools such as box plots can provide a qualitative characterization, exploring the design space of the parameters influencing the explanation's sensitivity to the training randomness may benefit from a more quantitative approach. First attempts in this direction explored simple (word-level univariate, first-order) explanations and proposed tentative information theoretic metrics such as the explanation's signal, noise and Signal-to-Noise Ratio (SNR). They left the suitability of such metrics as an open question, which we tackle in this work. For this purpose, we start by identifying corner cases where they appear unable to capture intuitively desirable features of explanations corresponding to a different training randomness. Namely, the SNR does not reflect well the relative differences of relevance (between words). We next put forward that the correlation with a mean explanation provides a better treatment of these corner cases, at the cost of being unable to reflect absolute differences of relevance (for single words). We then discuss how to turn these observations into a consolidated approach for analyzing the explanations' sensitivity to the training randomness. While there is no silver bullet that perfectly deals with the full complexity of this sensitivity problem, we argue that design space exploration with the correlation metric and individual model analysis with box plots provides a good tradeoff. Besides, we put forward additional desirable features of the correlation metric (e.g., unbiased estimation thanks to cross-validation and simple confidence intervals).

1 Introduction

In recent years, Large Language Models (LLM) like BERT [8] or GPT [9] have led to significant performance improvements for a vast amount of Natural Language Processing (NLP) tasks [1]. These improvements generally come from more complex architectures with more parameters, of which the training relies on randomized optimization techniques. As a result, it has been consistently observed that the explainability of LLMs is a major challenge [14], which is especially important for applications implying critical (e.g., medical or legal) decisions.

At high level, the explainability of LLMs relates to broad and hard-to-define concepts like faithfulness [16, 12] and plausibility [11, 12]. Informally, faithfulness

requires that an explanation accurately reflects the algorithmic reasoning process behind a model's predictions, and plausibility requires explanations to be understandable and convincing to the target audience. In this paper, we are concerned with a more specific issue which has connections with both concepts. Namely, the sensitivity of the explanations to the randomness used to train models, recently put forward by Bogaert et al. [6, 3]. The main observation of this paper is that it is sometimes possible to produce many models of which the training only differs by the (indistinguishable) random seeds they use, that are "equivalent" from the accuracy viewpoint and nevertheless lead to different explanations.[1] The authors then argue that this sensitivity to the training randomness must at least be characterized, since in the extreme case where the explanations would be uniformly distributed, any selection of explanation would be completely arbitrary.

The explanations' sensitivity to randomness has for now been exhibited in the case of "simple" explanations, defined in [5] as word-level, univariate (i.e., assigning a single relevance value per word) and first-order (i.e., assuming readers are interested by mean explanations in case of sensitivity to randomness). We will use Chefer et al.'s Layerwise Relevance Propagation (LRP) method as our running example [7]. Such simple explanations, next denoted as (1,1,1), are of course not expected to be perfectly faithful, although we assume they reflect the models' reasoning to a sufficient extent. They are not expected to be the only plausible ones either. Yet, they provide a useful theoretical framework to answer the question: *how stable can the simple explanations of complex models be?*

Evaluating the sensitivity to the training randomness of LLMs can be done qualitatively. For example, visualization tools like box plots provide a good intuitive understanding of single texts. Yet, more quantitative tools become useful to explore the explanations' design space. For example, one could be interested to compare the randomness' sensitivity of different texts, and for explanations assigning relevance scores for various number of words. One could also be interested to compare the randomness' sensitivity of bigger vs. smaller models, for various tasks, datasets or languages, or for different explanation methods. First steps in this direction were made in [5], where the explanations' signal, noise and Signal-to-Noise Ratio (SNR) are proposed as tentative explanation stability metrics. In this paper, we consolidate these investigations in three directions.

First, we highlight the limited ability of the SNR to reflect the relative differences of relevance (between words) in a set of explanations corresponding to different (random) training seeds. We additionally show that the correlation with a mean explanation mitigates this issue, as to cost of being unable to reflect absolute differences of relevance (for single words), which are better captured by the SNR. Second, we discuss the consequence of these observations and argue that combining a design space exploration with the correlation metric and a more qualitative analysis thanks to box plots appears as a good tradeoff. The first one better captures relative differences within explanations, whereas the

[1] Equivalent meaning that there is no statistically significant difference in their accuracies, implying that there is no "better" model from the accuracy viewpoint.

second one reflects absolute differences at the individual word level, exhibiting possibly interesting intuitions that the (quantitative) SNR metric may hide. We finally put forward additional desirable features of the correlation metric such as easier interpretation, unbiased estimation thanks to cross-validation and simple confidence intervals thanks to a well-known statistical distribution.

Related works. The quantitative evaluation of the sensitivity to the training randomness is quite related to the problem of inter-annotator agreement – see for example [10, 2]. One difference is that the explanations of LLMs provide continuous relevance scores (vs. more discretized ones for human annotators). The other is that, due to the (1,1,1) restriction, we can replace pairwise correlations, which are frequently used in the inter-annotator agreement literature but can become expensive as the number of random seeds under investigation increases in our context, by the correlation with a mean explanation. Our study is also related to [18] which, among others, performed an experiment to test whether the words' relevance obtained thanks to four different types of explanations were impacted by the random seeds used for model initialization. They used Pearson's correlation for this purpose, but only considered two random seeds and did not ensure model equivalence (nor input compatibility, as we define next).

2 Background

2.1 Dataset, model and explanation method

We run our experiments on the InfOpinion dataset [4], composed of 10,000 french texts belonging to the *information* and *opinion* journalistic genres. This binary categorization relies solely on the articles' annotation by their authors as either *information* or *opinion*. The dataset is split in 3 parts: a training set (80%), a validation set (10%) and a test set (10%). The classes are balanced among each of these sets. The task is to predict the binary category of a given text.

The model we consider is the French pre-trained transformer model CamemBERT [15], in the two different setups presented in [8]. In the first one, that we denote as *fine-tuned*, we jointly train all the weights of the encoder blocks and the classification head during 2 epochs. In the second one, that we denote as *frozen*, we only train the classification head while freezing the encoder blocks (i.e., the model learns to use the embeddings without modifying them). We note that the model's training randomness can be controlled via a seed parameter that rules the initialization of the layers, the order of the training dataset and the neurons that are deactivated by the dropout layers during the training.

Once our model is trained, we use Chefer et al.'s LRP method to generate word-level explanations for every text [7]. It back-propagates 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 obtain word-level explanations.[2]

[2] CamemBERT uses the roBERTa tokenization to work with word pieces. We post-process explanations to get one weight per word instead of one per word piece.

2.2 Equivalent models' explanations

In a previous work [6, 3], Bogaert et al. showed that the training randomness of LLMs can have an impact on their explainability. To do so, and as illustrated in Figure 1, they trained many models with the same settings and on the same dataset, but with different random seeds. The accuracy of these models was then

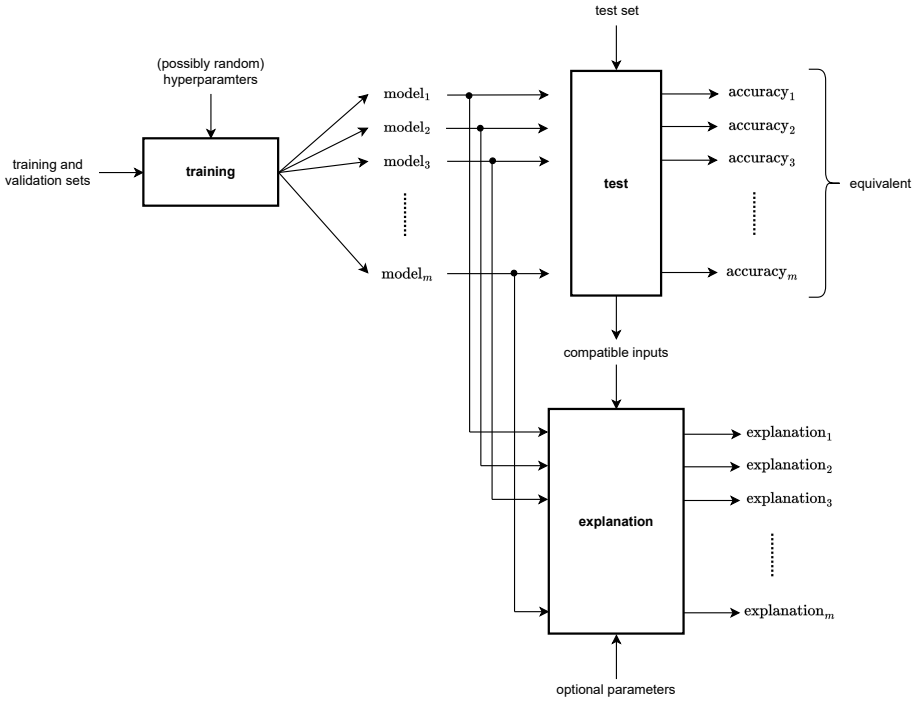

Fig. 1. Setup for the generation of equivalent models and compatible inputs.

evaluated on a test set, and a subset of m most accurate models was selected, such that the difference between the best (a) and worst (b) accuracies of the models in the subset was not statistically significant. For this purpose, one can computed the z statistic [13], which can detect whether two proportions (here, the accuracies a and b) are different:

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

As Bogaert et al., 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 z values, we conclude that these accuracies do not differ

significantly and therefore, we consider the models in the subset as equivalent from the performance viewpoint. Starting from 200 models (see Section 2.1), a restriction to $m = 100$ was sufficient to reach model equivalence in the subset. We then selected so-called compatible inputs for which all models predict the same class, and we computed explanations for each model on such inputs.

2.3 Explanation stability

The main observation in [6, 3] is that the explanations of equivalent models on compatible inputs can differ, raising a need to characterize their sensitivity to the training randomness. For this purpose, one can construct an explanation matrix of m rows (corresponding to different random seeds) and n columns (corresponding to different words), where each $a_{s,w}$ corresponds to the relevance value assigned by the s-th model (seed) to the word at the w-th position, as showed in Figure 2. The left part of the figure additionally shows the average curve which corresponds to the "simple" (word-level, univariate and first-order) explanations introduced in [5]. Word-level means that all m explanations of an n-word text display a weight for each word independently. Univariate means that each of these weights is a single value. First-order means that variable explanations are summarized by their mean (i.e., a first-order statistical moment).

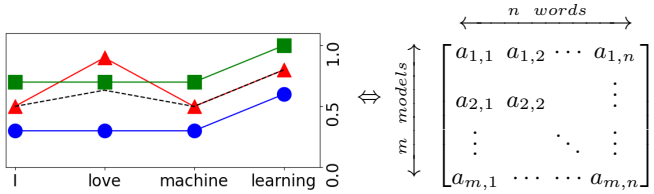

Fig. 2. Explanations of $n = 4$-word texts for $m = 3$ seeds and mean (dotted).

2.4 *k*-words explanations

To capture the possibility that shorter explanations are more plausible, we can evaluate so-called k-word explanations, obtained by keeping only the $0 \leq k \leq n$ highest relevance values of each explanation. To further simplify the individual explanations, one can also use k-word binary explanations, where only the k top words are considered relevant, without any distinction among them:

$$a_{s,w} = \begin{cases} a_{s,w} \text{ if } a_{s,w} \in TOP_k(a_{s,:}), \\ 0 \quad \text{oth.} \end{cases} \qquad a_{s,w} = \begin{cases} 1 \text{ if } a_{s,w} \in TOP_k(a_{s,:}), \\ 0 \text{ oth.} \end{cases}$$

2.5 Signal, noise and SNR

Simple (word-level, univariate and first-order) explanations naturally suggest simple quantities to capture their sensitivity to randomness. In [5], the explanations' signal (S), noise (N) and Signal-to-Noise Ratio (SNR) were suggested as

tentative metrics for this purpose. Intuitively, the signal reflects the flatness of the average explanation, the noise reflects the variation of the relevance scores for each word (averaged) and the SNR is simply the ratio between both:

$$S = \underset{n \text{ words}}{\hat{\mathsf{Var}}}\left(\underset{m \text{ seeds}}{\hat{\mathsf{E}}}\left(a_{s,w}\right)\right), \quad N = \underset{n \text{ models}}{\hat{\mathsf{E}}}\left(\underset{m \text{ seeds}}{\hat{\mathsf{Var}}}\left(a_{s,w}\right)\right), \quad SNR = \frac{S}{N}.$$

3 Metrics' corner cases

We next discuss the adequacy of the SNR metric to reflect the stability of explanations in the setting of Figure 2. For this purpose, we use illustrative hand-made examples and compare how the stability of some explanations is captured by the SNR and by an alternative simple metric, namely the (average) correlation with a mean explanation. We are in particular interested in the ability of these metrics to reflect the relative differences of relevance between words and the absolute differences of relevance for single words in a set of variable explanations.

3.1 Relative differences (between words)

Figure 3 illustrates two pairs of explanations such that the relevance of some words are swapped when moving from the left to the right plots. As a result, these left and right plots show quite disparate relative differences of relevance between words. Interestingly, the SNR metric is unable to reflect these relative differences. This is because the swaps do not affect the mean explanations (which are the same on the left and right plots, leading to the same signal) nor the absolute difference between words (hence the noise). By contrast, the correlation metric captures these relative differences: the explanations of the left plot are highly correlated with the mean explanation; the ones of the right plot are not.

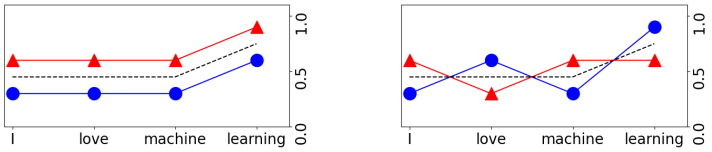

Fig. 3. Two pairs of explanations (△ and ◯), with the same absolute differences and different relative differences, with the mean explanation in dotted line.

3.2 Absolute differences (for single words)

A complementary situation is illustrated in Figure 4, in which an offset δ was added to all the relevance values of one explanation and subtracted for the other. As a result, the left and right plots show disparate absolute differences. This time,

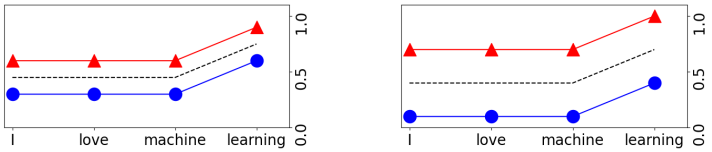

Fig. 4. Two pairs of explanations (△ and ◯), with the same relative differences and different absolute differences, with the mean explanation in dotted line.

the correlation metric is unable to reflect the discrepancy between the left and right plots (because the correlation is invariant to the δ offset). By contrast, the SNR reflects it because the noise of the left and right plots differs.

3.3 Discussion

The two examples above suggest a quite natural tradeoff between the SNR and correlation metrics: the first one better captures absolute differences, the second one better captures relative differences. While this may encourage using both metrics in parallel, Figure 5 highlights additional limitations of the (noise component of the) SNR metric. Namely, it illustrates that the noise metric is averaged over (possibly dependent) words, which may hide important intuition regarding which word is causing the noise. (By contrast, the correlation can be averaged over independent seeds). As a result, we suggest using the correlation metric for design space exploration and box plots for a qualitative analysis of the noise. As will be experimented next, this appears as a relevant combination to characterize the explanations' sensitivity to the training randomness, capturing both the absolute and relative differences within these explanations.[3]

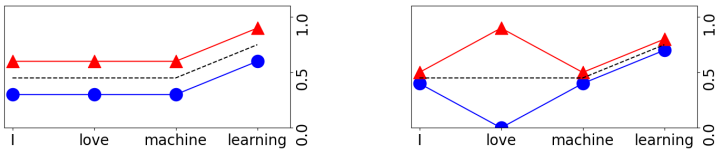

Fig. 5. Two pairs of explanations (△ and ◯), with the same (average) relative differences but distributing these differences differently among the words of a sentence.

Besides, the SNR is also slightly less convenient to manipulate from the statistical viewpoint. First, it is a biased metric since small estimation errors in the mean explanations are considered as signal by definition. Second, its interpretation in case of small noise levels is not always intuitive (e.g., the SNR tends to infinity when the noise tends to zero). Despite these drawbacks do not lead to

[3] In Appendix A, we give additional arguments why the noise metric alone cannot be used for design space exploration. In appendix B, we give additional arguments why the signal metric alone is making undesirable implicit plausibility assumptions.

fundamental issues (i.e., the SNR bias decreases with the amount of seeds and can be corrected, intuition is just less direct), we show next that the correlation coefficient also comes with advantages in this respect. It can be estimated without bias thanks to cross-validation, benefits from a well-known sample distribution leading to easy-to-obtain confidence intervals and its interpretation is direct.

4 Application to case studies

We now apply the methodology proposed above to the classification case study described in Section 2. First, we detail how to estimate the correlation metric in Section 4.1. Next, we show how it can be used for design space exploration in Section 4.2. Finally, we illustrate how such a quantitative analysis is nicely combined with a more qualitative one using box plots in Section 4.3.

4.1 Estimation and confidence interval

The examples of Section 3 suggest using the correlation of different explanations with their mean as a good way to quantify the explanations' sensitivity to the training randomness.[4] We next detail how correlation samples can be estimated without bias thanks to 10-fold cross validation, and possibly averaged.

For this purpose, the average explanation is first repeatedly computed using 90% of the explanation matrix's rows and the remaining 10% of the rows are repeatedly compared to these means in order to compute correlation samples (one per explanation). Figure 6 shows a scatter plot of all the correlations to the mean (i.e., one per trained model, so 100 in our case study), highlighting the high disparity of the results depending of the training randomness.

Different quantities of the correlation distribution could then be considered to summarize the explanations' stability. In the following, and for simplicity purposes, we suggest to use the average correlation. We note that while it is in general better to estimate the correlation between two variables based on a large set of samples than averaging correlations estimated from several smaller sets of samples, this approach can serve as a useful heuristic in our context, if interpreted carefully. Namely, as a way to capture a global tendency for many explanations, possibly leading to different correlation values. For this purpose, we follow [17] and first use the following "Fisher Z transformation":

$$F(\hat{\rho}) = \frac{1}{2}\ln(\frac{1+\hat{\rho}}{1-\hat{\rho}}) = \operatorname{arctanh}(\hat{\rho}),$$

which projects the correlation samples in a space where they are normally distributed. We can then compute the average Fisher value $\bar{F}$, as well as its sample variance $\hat{\sigma}^2$. A confidence interval on the estimation of $\bar{F}$ (e.g., 96%) is obtained

[4] Under the assumption of simple explanations formalized in [5] as (1,1,1) explanations, computing the average correlation to a mean explanation rather than the average pairwise correlation allows significant speedups without intuition loss.

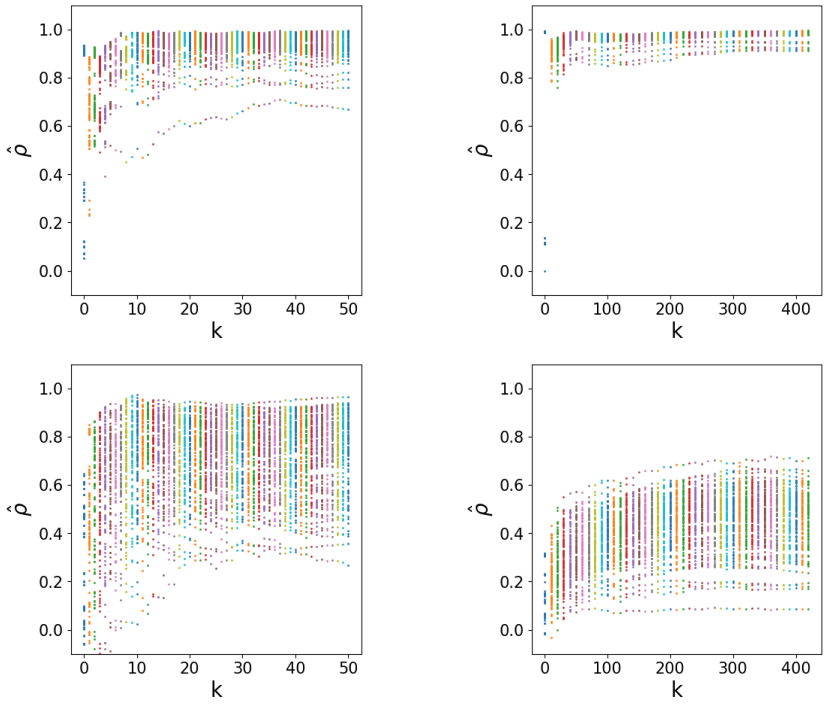

Fig. 6. Scatter plot of the correlation to the mean for LRP explanations corresponding to the *frozen* (top) and *fine-tuned* (bottom) models in function of the number of words used per explanation, illustrated for a short (left) and a long (right) text. For readability, only the values $k = 10, 20, 30, ...$ are displayed for the long text.

by adding or removing $2\frac{\hat{\sigma}}{\sqrt{m}}$ to $\bar{F}$. Applying the inverse function $\rho = \tanh(F(\rho))$ finally leads to 96% confidence interval for the average correlation:

$$\left[\tanh\left(\bar{F} - \frac{2\hat{\sigma}}{\sqrt{m}}\right); \tanh\left(\bar{F} + \frac{2\hat{\sigma}}{\sqrt{m}}\right)\right].$$

This interval indicates that the average correlation is better estimated with more models (i.e., large m values). By contrast, longer texts (i.e., large n values) lead to better estimated correlation samples, but do not necessarily decrease the variance $\hat{\sigma}$, since the correlations of different explanations may differ.

4.2 Quantitative analysis

Figure 7 shows the average correlation to the mean for k-word explanations. Positing that shorter and more aligned explanations are more plausible, such an exploration can lead to identify relevant parameters to investigate more qualitatively. For example, we can see on the left plot that the average correlation to the mean increases up to $k = 7$ and then reaches a plateau . Hence, larger

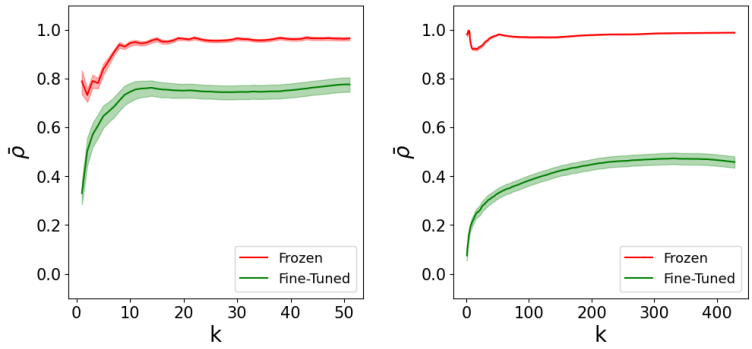

Fig. 7. Average correlation to the mean (with confidence intervals) for LRP explanations corresponding to the *frozen* and *fine-tuned* models in function of the number of words used per explanation, illustrated for a short (left) and a long (right) text.

values of k (i.e., longer explanations) may not lead to a reduced sensitivity to the training randomness. We next complete this observation with a qualitative analysis for the explanations obtained for $k = 7$ and the maximum $k = 51$.

4.3 Qualitative analysis

Starting with the box plot for $k = 7$ displayed on Figure 8, we can observe that, qualitatively as well, the LRP explanations of the frozen model are significantly less sensitive to the training randomness than the ones of the fine-tuned model. This is quite expected since the amount of network weights that are trained in these two models vastly differ. What is maybe less expected is that the the variability per word is also distributed very differently for both models. Namely, 7-word explanations across the 100 seeds only consider 10 different words in the frozen case, while most words are considered by the fine-tuned models. This tends to justify our proposed methodology, where we do not analyze the absolute difference with the noise metric (which is averaged over the words).

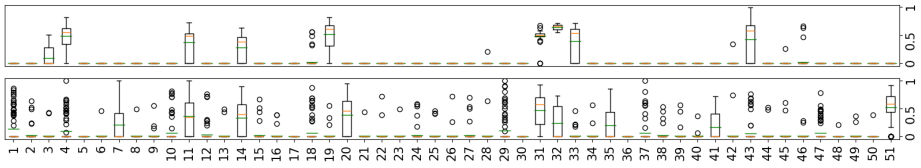

Fig. 8. Box plot for $k = 7$ and the *frozen* (top) and *fine-tuned* (bottom) models.

More interestingly, Figure 9 shows the box plots obtained for the same models and $k = 51$. Its upper part is particularly relevant: it confirms that increasing the explanations' length beyond $k = 7$ is not only discouraged by the correla-

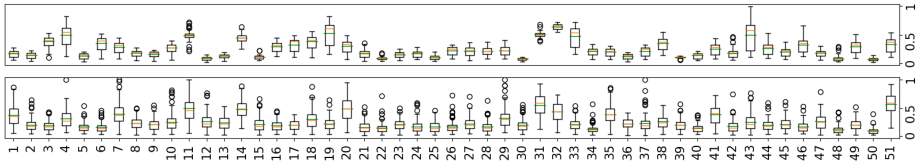

Fig. 9. Box plot for $k = 51$ and the *frozen* (top) and *fine-tuned* (bottom) models.

tion metric, it actually also leads to harder to interpret first-order explanations assigning non-zero relevance scores to most words, as the fine-tuned model.

5 Conclusions

Our results provide consolidated tools for analyzing the sensitivity of the explanations of LLMs to the training randomness, hopefully opening a path to their better understanding and leading to various interesting open problems.

First, and maybe most importantly, the extent to which the stability of the explanations of LLMs is a requirement for their plausibility remains unknown. While we posit in the paper that shorter and more aligned explanations are easier to understand, it could also be that human explanations show variations that are similar to the ones observed in this paper. Designing a real-world experiment with human annotators would be interesting to contribute to this question.

Second, even if explanations appear unstable when considering their average correlation to a mean explanation as in this paper, it is possible that some clusters exist within these explanations. This would mimic a situation where a few groups of human annotators share very similar explanations within the groups and have very different ones between the groups. In order to stimulate research in this direction, Figure 10 shows a TSNE visualization of 100 explanations used in our experiments. It would be interesting to investigate whether clusters an be extracted from such plots and lead to more stable/aligned explanations.

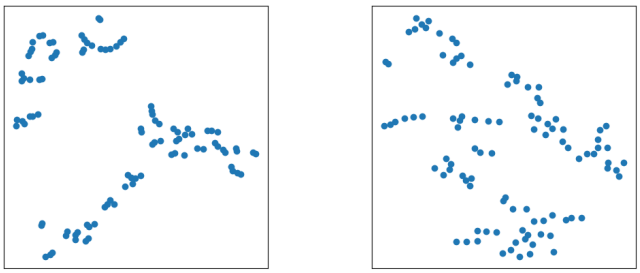

Fig. 10. TSNE for $k = 51$ and the *frozen* (left) and *fine-tuned* (right) models.

Third, it would be interesting to investigate whether more complex explanations (e.g., assigning relevance scores to tuples of words) or more complex models (e.g., generative ones) may lead to different outcomes, and whether a sensitivity to the training randomness is observed for other tasks or data sets.

Finally, our conclusion may also differ for other modalities than texts. For example, image explanations may be more stable due to the more correlated nature of adjacent pixels (compared to consecurive words in a text).

Acknowledgments. François-Xavier Standaert is a research director of the Belgian fund for scientific research (FNRS-F.R.S.). This work has been supported by the Service Public de Wallonie Recherche, grant n°2010235-ARIAC.

References

1. Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer Models for Text-Based Emotion Detection: a Review of BERT-Based Approaches. *Artif. Intell. Rev.*, 54(8):5789–5829, 2021.
2. Ron Artstein. Inter-Annotator Agreement. *Handbook of linguistic annotation*, pages 297–313, 2017.
3. Jérémie Bogaert, Marie-Catherine de Marneffe, Antonin Descampe, Louis Escouflaire, Cédrick Fairon, and François-Xavier Standaert. Explanation Sensitivity to the Randomness of Large Language Models: the Case of Journalistic Text Classification. *CoRR*, abs/2410.05085, 2024.
4. Jérémie Bogaert, Louis Escouflaire, Marie-Catherine de Marneffe, Antonin Descampe, François-Xavier Standaert, and Cédrick Fairon. TIPECS: A Corpus Cleaning Method using Machine Learning and Qualitative Analysis. In *International Conference on Corpus Linguistics (JLC)*, 2023.
5. Jérémie Bogaert and François-Xavier Standaert. A Question on the Explainability of Large Language Models and the Word-Level Univariate First-Order Plausibility Assumption. *Responsible Language Models (ReLM)*, 7 pages, 2024.
6. Jérémie Bogaert, Marie-Catherine de Marneffe, Antonin Descampe, Louis Escouflaire, Cédrick Fairon, and François-Xavier Standaert. Sensibilité des Explications à l'Aléa des Grands Modèles de Langage: le Cas de la Classification de Textes Journalistiques, *TAL (Traitement Automatique des Langues)*. Vol 64, num 3, pages 15-40, 2024.
7. Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *CVPR*, pages 782–791. Computer Vision Foundation / IEEE, 2021.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.
9. Tom B. Brown et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.
10. Andrew F. Hayes and Klaus Krippendorff. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89, 2007.
11. Bernease Herman. The Promise and Peril of Human Evaluation for Model Interpretability. *CoRR*, abs/1711.07414, 2017.

12. Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *ACL*, pages 4198–4205. Association for Computational Linguistics, 2020.
13. Erich Leo Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer texts in statistics. Springer, 2008.
14. Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards Faithful Model Explanation in NLP: A Survey. *CoRR*, abs/2209.11326, 2022.
15. Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. *CoRR*, abs/1911.03894, 2019.
16. Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*, pages 1135–1144. ACM, 2016.
17. N Clayton Silver and William P Dunlap. Averaging Correlation Coefficients: Should Fisher's Z Transformation be Used? *Journal of applied psychology*, 72(1):146, 1987.
18. Zhengxuan Wu and Desmond C. Ong. On Explaining Your Explanations of BERT: An Empirical Study with Sequence Classification. *CoRR*, abs/2101.00196, 2021.

A The noise is not good for design space exploration

As illustrated in Figure 11, we cannot use the signal or the noise alone to explore our design space, as their range is directly impacted by the amount of top words k. This is the case even for deterministic/random explanations, which lead to an hypothesis that selecting a certain ratio of word leads to more stable explanations, even if all the models perfectly agree on the relevance of every token. This is not the case for the correlation metric that is always at its maximum for the deterministic model, and at its minimum for the random one.

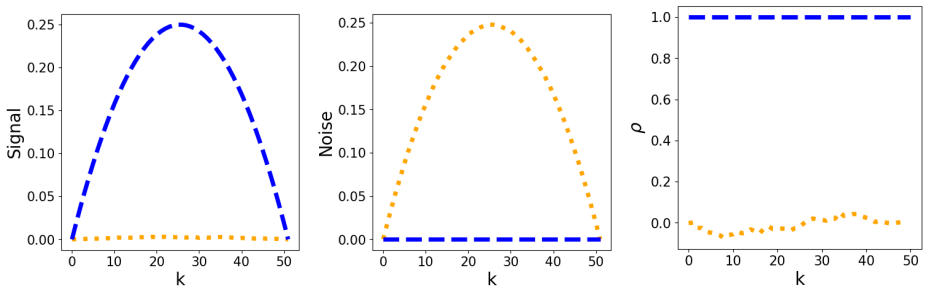

Fig. 11. Signal (left), noise (middle) and correlation (right) of deterministic (-) and random (.) explanations, for the binary variant of k-word explanations.

B Implicit assumption of the signal metric

As illustrated in Figure 12, it is possible to obtain explanations such that their absolute and relative differences are identical, but their signal differs, because the signal is focused on the flatness of the mean explanation. It implicitly suggests that more relative differences within this mean explanation lead to better explainability, which may not be connected to a definition of plausibility.

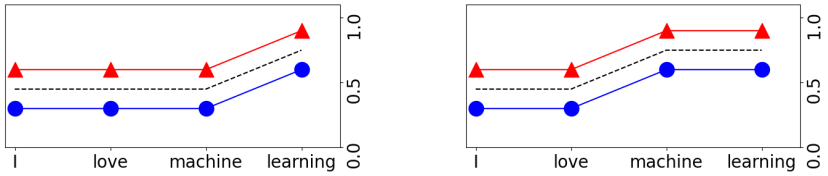

Fig. 12. Two pairs of explanations (△ and ◯), with the same absolute and relative differences and different signal, with the mean explanation in dotted line.