Explaining Pre-Trained Language Models with Attribution Scores: An Analysis in Low-Resource Settings

LREC-COLING 2024

Wei Zhou | Bosch Center for Artificial Intelligence
Heike Adel | Hochschule der Medien
Hendrik Schuff | Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
Ngoc Thang Vu | Institut für Maschinelle Sprachverarbeitung, University of Stuttgart

BOSCH

Introduction
Motivation

- When deploying pre-trained models in real world downstream applications, two challenges arise:
  - **Explaining results** as models are very complex (Madsen et al., 2022).
  - Adapting models in **low-resource scenarios** as applications in special domains or languages typically do not provide many labeled training data (Hedderich et al., 2021).

Introduction
Explainability

- Explainability: the ability [of a model] to explain or to present [its predictions] in understandable terms to a human (Doshi-Velez and Kim, 2018).
- Importance of explainabilitiy:
  - Required in General Data Protection Regulation (GDPR)
  - High-stakes decision making scenarios
- Explainability methods:
  - White-box models (e.g., decision tress (Quinlan, 2004))
  - **Attribution socres** (e.g., SHAP (Lundberg and Lee, 2017))

Figure 1: Explanation example from SHAP

BOSCH

Introduction
Prompting

- In low-resource settings**, prompting works better than fine-tuning** in terms of task performances (Brown et al., 2020; Schick and Schütze, 2021; Liu et al., 2022).

- To train a promt-based model (PBM), inputs are reformatted to align with the mask-language modeling target used in pre-training.

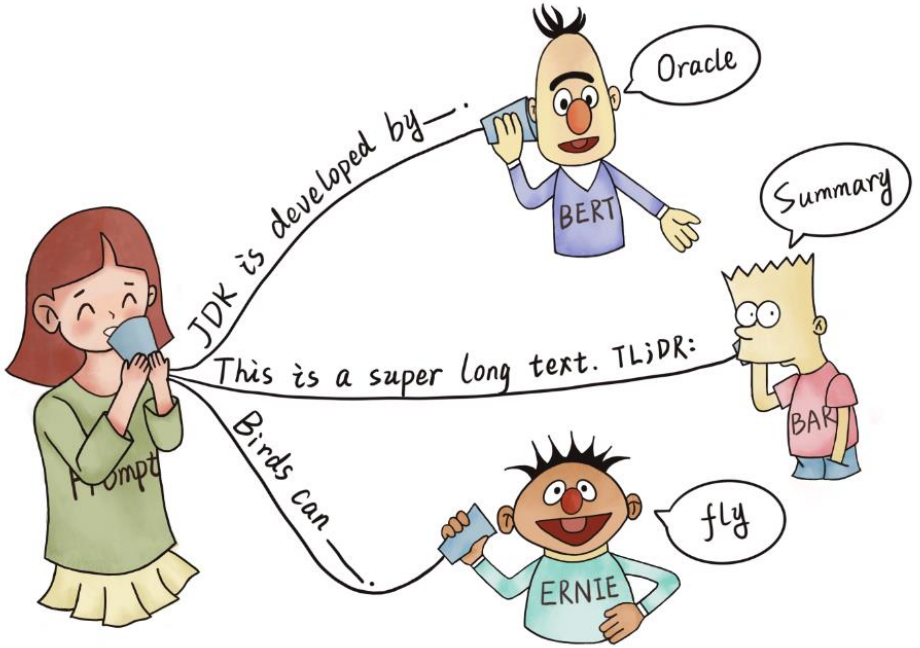

Figure 2: Illustration of the prompting paradigm. Image from Liu et al., 2022.

Introduction
Our Contribution

- Current work on explanability with attribution scores have mainly focus on **fine-tuned models (FTMs)** (Atanasova et al., 2020; DeYoung et al.,2020; Ding and Koehn, 2021). However,
  - Fine-tuned models cannot be applied in **low-resource settings** (where prompt-based models (PBMs) are commonly used).
  - Current trend of **large language models (LLMs)** focuses more on **prompting** than fine-tuning.
- **Our contribution**:
  - Analysis of attribution scores from **PBMs** (encoder-based/decoder-based models).
  - Comparisons of attribution scores extracted from PBMs and FTMs in low resource settings.

Method
Extraction of Attribution Scores from PBM&FTM

- Attention scores: the last hidden layer of the [MASK] ([CLS]) token, average across different heads and normalize over the task input.
- Integrated Gradients: Captum package.
- Shapley Value Sampling: Captum package.

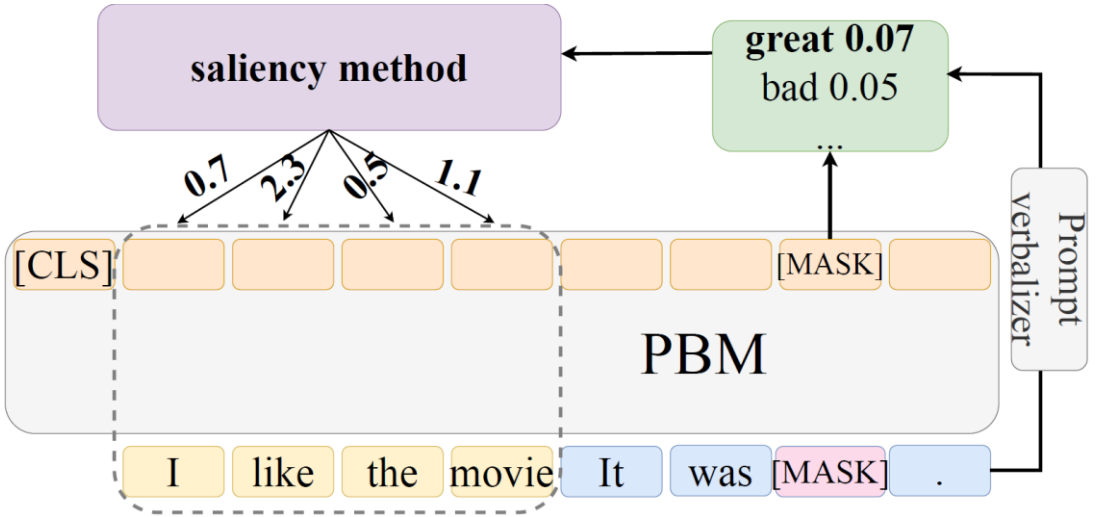

Figure 3: Method illustration: extraction of attribution scores from PBM.

Method
Extraction of Attribution Scores from LLM

- Prompt the model to output verbalized class label.
- If outputs verbalized class, treat the verbalized class label as the prediction token.
- If not, treat the first token as the prediction token.

- Extraction process similar as before for all attribution scores.

You will be given a target sentence and you will decide the sentiment of the sentence (Please return either yes or no only).
Here are some examples: Input: {s1} Output: {l1}...Input: {s8} Output: {l8}

Figure 4: Example prompt.

Experiments
Research Questions

1. How plausible and faithful are attribution scores extracted from PBMs in comparison to FTMs?

2. How well do different attribution methods perform in terms of plausibility and faithfulness?

3. Do the results for PBMs also hold for decoder-based large language models?

Experiments
Setup

- Task and datasets:
  - Sentiment classification (TSE)
  - Natural language inference (e-SNLI)
- Base models:
  - BERT-base/large (Devlin et al., 2019)
  - RoBERTa-large (Liu et al., 2019).
  - Vicuna model (Chiang et al., 2023)
- Prompting methods
  - Manual (Schick and Schütze, 2021): manual prompts, fine-tunes all parameters.
  - BitFit (Logan IV et al., 2022): manual prompts, updates only bias terms.
  - BFF (Gao et al., 2021): automatically searches for a prompt, fine-tunes all parameters.

Task	Prompt	Verbalizer	Setting
TSE	[S] It was [P]. This is [P]. [S]	terrible/great ragged/soldiers	Manual/Bitfit BFF
e-SNLI	[S1] ? |[P] , [S2] [S1] . [P] , no , [S2]	yes/no/maybe alright/except/watch	Manual/Bitfit BFF

Table 1: Prompts and verbalizers used.

Experiments
Evaluation metrics

- Plausibility
  - How plausible an explanation is according to human understanding (**human perspective**).
  - Average precision, following Atanasova et al (2020). The higher the better.
- Faithfulness
  - To what extent the deemed important tokens are truly important for the predictions of the model (**model perspective**).
  - Normalized AUC score. The lower the better.
    - Iteratively masking 0, 10, 20, ..., 100% of the tokens in the order of decreasing saliency.
    - Calculate task performance (F1 scores), obtain AUC score.
    - Normalize AUC as the proportion of the area under the curve (orange area) to the whole area (green box).

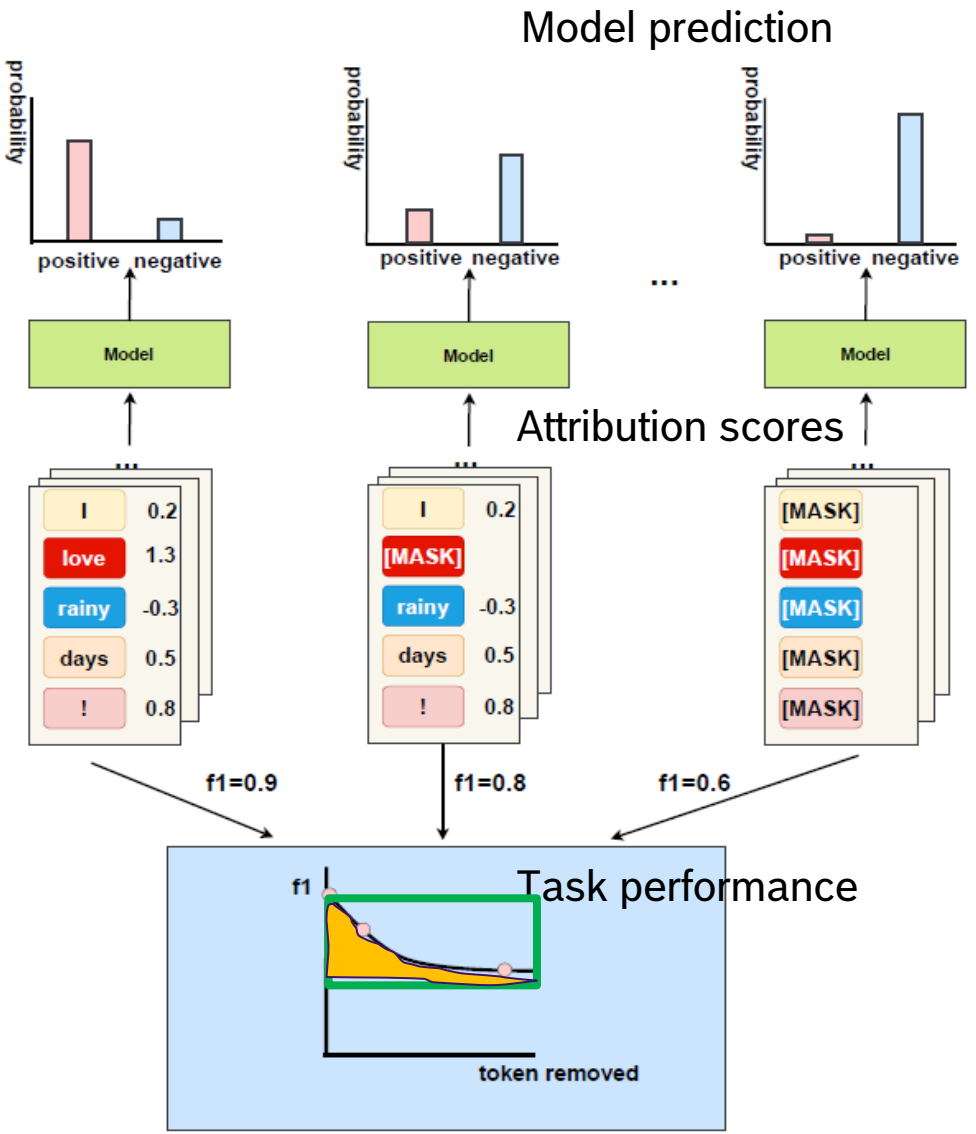

Figure 5: Illustration of obtaining faithfulness scores.

Results
Comparing PBMs with FTMs

- In low resource settings (smaller training sizes), explanations from PBMs are more plausible than explanations from FTMs.
- The trend reverses as the training size increases.

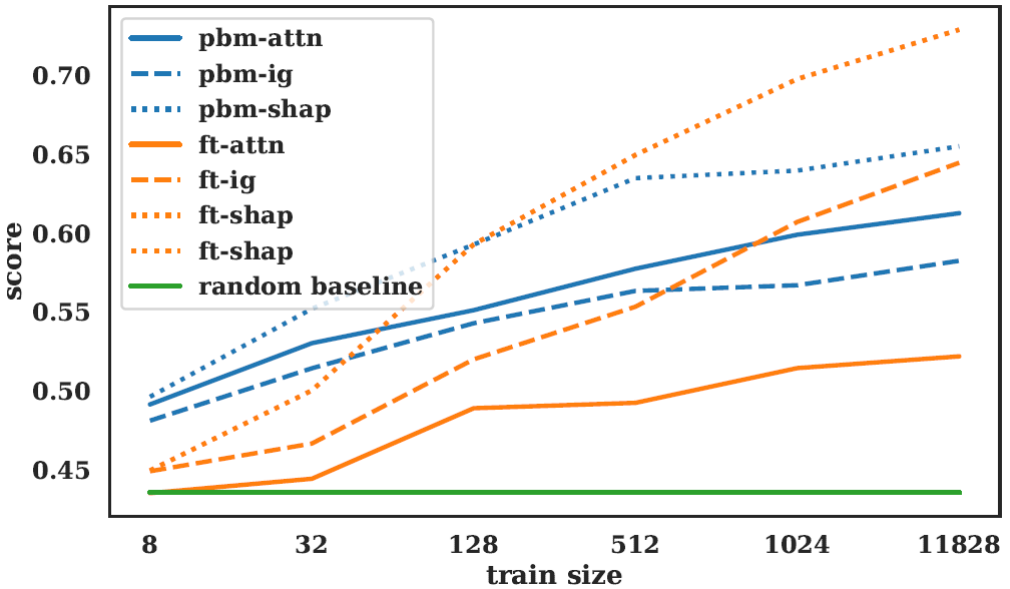

Figure 6: Plausibitiy scores on TSE, averged across base models and seeds.

Results
Comparing PBMs with FTMs

- Faithfulness scores are influenced by the attribution methods.
  - Explanations from FTMs with Shapley Value Sampling are more faithful than explanations from PBMs independent of the number of resources and attribution methods.
  - Explanations from PBMs with attention lead to the lowest faithfulness scores across all training sizes.

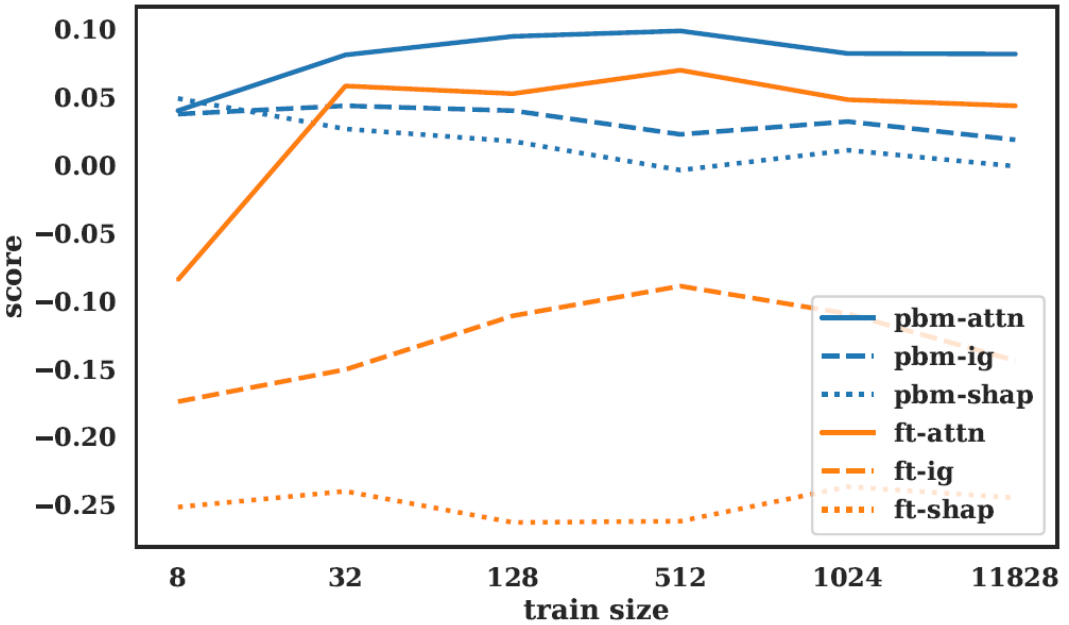

Figure 7: Faithfulness scores on TSE, averged across base models and seeds.

Results
LLMs

- We limit test instances to 100 due to the computation cost and evaluate 8 shots only (low resource settings).
- Shapley Value Sampling again leads to more plausible and faithful explanations for Vicuna.
- The plausibility scores of attention are lower for Vicuna than for RoBERTa. This might be because LLMs encode a larger input context, thus information of tokens that are irrelevant to the prediction might also be encoded.

Data	Model	Plausibility			Faithfulness		
		attn	ig	shap	attn	ig	shap
TSE	RoBERTa	.56	.57	.56	.02	**.00**	.01
	Vicuna	.47	.57	**.59**	.07	.06	.02
e-SNLI	RoBERTa	.53	.51	.50	.22	.09	.11
	Vicuna	.43	.51	**.55**	.02	.05	**.00**

Table 2: Results of LLM (Vicuna) and the best model (RoBERTa) from previous experiments.

Conclusion
Take-away Messages

1. *How plausible and faithful are attribution scores extracted from PBMs in comparison to FTMs?*
   - Plausibility: In low resource settings (smaller training sizes), explanations extracted from PBMs are more plausible than explanations from FTMs. The trend reverses in high resources settings.
   - Faithfulness: FTMs with Shapley Value Sampling brings the most faithful explanations.

2. *How well do different attribution methods perform in terms of plausibility and faithfulness?*
   - Shapley Value Sampling leads to the most plausible and faithful explanations.

3. *Do the results for PBMs also hold for decoder-based large language models?*
   - For Vicuna, we find Shapley Value Sampling results in the best explanations in terms of both plausibility and faithfulness. Attention as an attribution method leads to less plausible explanations for Vicuna than for RoBERTa.

References

- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20,
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E.Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5034–5052, Online. Association for Computational Linguistics.

References

- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. stat, 1050:2, 2017.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better fewshot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.
- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.
- J. Ross Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 2004.