Evaluating Local Explanations for Survey Variable Detection in Scientific Publications

Exploring the utility and faithfulness of the model's explanations

Andrea Zielinski[1,*,†], Calvin Spolwind[2] and Henning Kroll[3]

[1]*Fraunhofer ISI, Karlsruhe, Germany*

Abstract

Deep learning methods are increasingly integrated into scientific discovery and document understanding applications, helping scientists to get access to high-quality information. In the Social Sciences, this includes identifying sentences that contain survey variables mentions in scientific publication as part of a dataset (Variable Detection).

Existing methods for the task rely on (generative) pre-trained transformer models and have a high performance, however, interpretability and lack of user trust is a major concern. This work investigates the capabilities of gpt3.5-turbo (chatGPT) and various BERT models for identifying survey variables in publications and generating an explanation for their decision.

Regarding performance, we find that the BERT-based fine-tuned supervised classifier outperforms the large language model (LLM) with an accuracy of 94.43% (versus 67.50% for chatGPT in a zero-shot setting). We observe that prompting LLMs to provide an explanation along with the prediction increases accuracy. In terms of interpretability, we apply various explainable techniques (i.e., LIME and SHAP) post-hoc to BERT, thus producing local explanations based on feature attributions. These are constrained to supervised models and cannot be applied to generative pretrained LLMs likewise. Faithfulness metrics show that LIME and SciBERT are best suited to reveal the model's decision-making process. Instead, natural language explanations (NLE) that justify a model's prediction are generated by chatGPT as *self-explanation*. We conduct a human evaluation of the generated free-text rationales to assess their quality. For the evaluation setup, three annotators judge explanations for their utility based on the evaluation criteria *wellformedness, consistency, factual groundedness* and *plausibility*. Our experiments were run on an open-access dataset, hence allowing full reproducibility.

Keywords

Machine learning, Deep Learning, Large Language Models, Generative models, Explainability

1. Introduction

Variable Detection (VD) deals with identifying sentences that contain survey variables mentions in scientific publications [1] which plays a crucial role in applications that seek to support Social Science researchers by providing enhanced information access through linking of publications and underlying datasets. The task is challenging, because generally no explicit references are provided, and understanding when a variable is mentioned may require contextual information

Workshop on Scientific Document Understanding co-located with 38th AAAI Conference on Artificial Inteligence, SDU@AAAI 2024, February 26, 2024, Vancouver, Canada

*Corresponding author.

andrea.zielinski@isi.fraunhofer.de (A. Zielinski)

0000-0003-2092-6153 (A. Zielinski); 0000-0002-2627-1541 (H. Kroll)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

as well as knowledge from external sources [2, 3]. In the context of the SV-Ident 2022 shared task on survey variable detection[1], a dataset has been released that includes challenging sentences that are representative of the real-world data. Previous research in the shared task utilized transformer-based model architectures and has achieved significant performance. However, systems have been evaluated based on prediction performance (e.g., by measuring the F1 score) [4], while the interpretability of these models has been neglected so far. Ideally, they should provide an explanation that help users to interpret and trust the model predictions. In the following study, we like to fill this gap.

The need for *interpretable* (deep) neural networks has been stressed by various authors [5], since they can help in model debugging, i.e. finding spurious correlations and indicating how well a model is generalizing to Out-of-Distribution (OOD) instances [5, 6] and build trust in users [7]. In the last decade, a variety of explanation methods have been developed [8] that enable the interpretation of predictions of black-box models. In particular, saliency maps have been proposed [9] that show *why* a specific prediction has been made by identifying input tokens that were particularly relevant to the output, thus showing the impact of individual features. However, depending on the choice of the explanation method for a given ML model, for the same input text, different saliencies might be generated. We compare two prominent post-hoc model-agnostic interpretability methods LIME [9] and SHAP [10] and measure their performance with a widely adopted metric for faithfulness [11, 12, 13, 14].

Meanwhile, with the advent of large language models (LLMs) and *in-context learning* [15], optimization by fine-tuning instructions [16] and reinforcement learning based on human feedback [17], further advancements have been made in many NLP tasks, including text classification. In this study, we like to explore if in-context learning with large models [18], in particular gpt-3.5-turbo (chatGPT) [17], is also effective in achieving high performance in our use case, comparing it to various BERT-base models [19].

Interestingly, huge LLMs with more than 50B parameters demonstrate reasoning capabilities with knowledge, in particular for LLMs with chain of thought (CoT) prompting [20]. Since LLMs are prone to generate hallucinations [21], their trustworthiness has been questioned [22]. *Explainable LLMs* that provide an explanation for a neural model's predictions by means of rationales (extracted predictive parts of the input) or free-text natural language explanations (self-explanation) [23] along with the classification help to build trust in models. Interestingly, these models have also resulted in improved task performance [24, 16]. Inspired by recent work that evaluates the utility of free-text [25] and extractive rationales [26] from a user-centered perspective [27], we like to assess the explanation quality and analyze *how reasonable the explanation is to a human* and if it justifies the predicted outcome. Despite ongoing efforts to evaluate LLMs for a number of NLP tasks [28], they remain poorly understood and pose novel challenges regarding evaluation [29, 30].

An example from our use case is shown in Figure 1. The BERT model correctly predicts a variable mention and the most salient features selected by SHAP (and LIME) are *people, positive attitude* and *nuclear*. Instead, gpt3.5-turbo chooses *nuclear power* and *positive attitude*. An example for free text rationales by chatGPT is given in Figure 2.

Contributions: In this work, we compare a prominent instruction-tuned LLM, i.e. gpt3.5-

[1]https://vadis-project.github.io/sv-ident-sdp2022

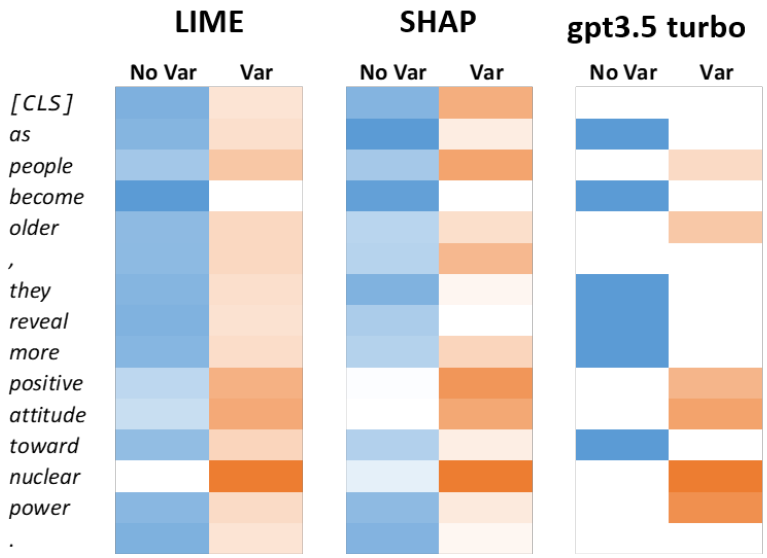

Figure 1: Salient features extracted for the sentence *As people become older, they reveal more positive attitude toward nuclear power.*, correctly predicted as positive ('is variable') by all models

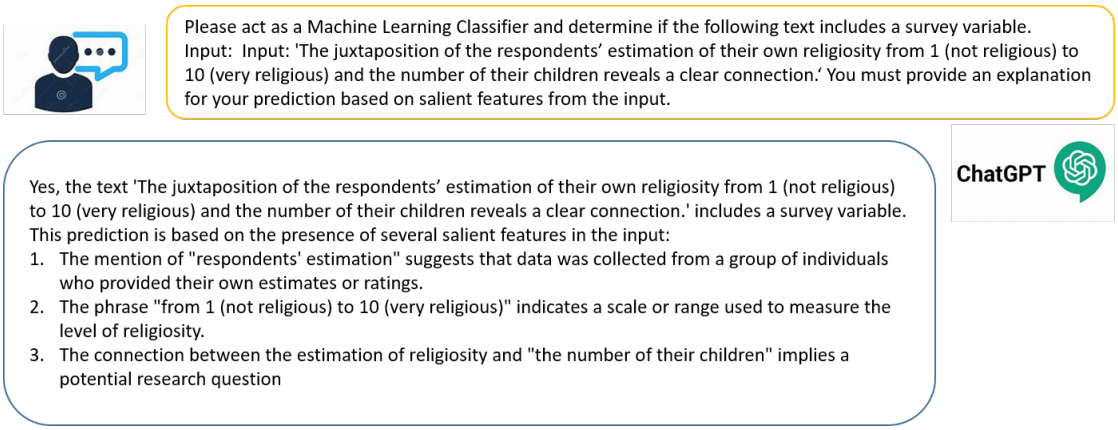

Figure 2: Based on the prompt and an input text *The juxtaposition of the respondents' estimation of their own religiosity from 1 (not religious) to 10 (very religious) and the number of their children reveals a clear connection.*, gpt3.5-turbo correctly predicts the class ('is variable') and provides an explanation.

turbo, under the zero-shot setting for a binary classification task (Variable Detection) to the fine-tuned BERT model along 2 dimensions: accuracy and explainability. We benchmark the LLM-generated self-explanations with traditional interpretability methods (i.e., SHAP and LIME) that are often used to explain classifiers.

2. Related Work

2.1. Text Classification

Text Classification is one of the fundamental tasks in natural language processing which involves assigning label(s) from a predefined set of labels to a given piece of text, in our case a sentence. In a recent survey, Li et al. [2022] show that text classification has greatly improved with the use of neural networks, however, most neural networks are data-hungry and need careful tuning of hyperparameters. Gao et al. [2023] show the benefits of providing label descriptions over zero-shot text classification. Moreover, Sun et al. [2023] compare results of different LLMs and prompting strategies on various text-classification benchmarks.

In this work, we compare the performances of chatGPT-turbo and fine-tuned BERT models as in [34], but our prompting strategies explicitly focus on extracting salient features.

2.2. Explainability

Interpretability methods deliver feature importance scores that show which input tokens led to the outcome prediction. Popular explainability methods for text-based input can be divided into: i) Model Simplification methods such as *LIME* (Local interpretable model agnostic explanations) [9, 35], ii) Perturbation based algorithms such as *Shapley Value Sampling* [10] and iii) Gradient based algorithms [36] such as *InputXGradient* [37] or *Saliency* [38] which require access to model internals.

In the case of generative models, either counterexamples [39] or natural language text [40] can be generated as an explanation to help uncover unreliable reasoning paths of the model, thus providing help to judge the correctness of explanations [26]. In this work, we focus on generating natural language explanations, i.e. LLMs are directly prompted to elicit an explanation, by either making a prediction and then explaining it, or vice versa [40, 41, 42]. Either feature attribution explanations or free-text explanations are delivered by the LLM [43, 44]. For tasks involving reasoning capabilities, chain-of-thought prompting [16] can be applied (self-rationalizing).

Explainability methods in the context of LLMs are still being actively explored [45].

3. Experimental Setup

This section describes the datasets, the classification models and the local explanation approaches used in our experiments.

3.1. Dataset

We experiment with two datasets (see Table 1). We combined the train and development datasets of OpenMinTeD 2018 English dataset [1] and the SV-Ident 2022 shared task English Dataset [4], and used the same test set as in the shared task.

Table 1
Combined English Datasets (SV-Ident 2022 and OpenMinTeD 2018) composed of 3K instances.

English	Train	Dev	Test
0	1696	188	63
1	1060	118	24

3.2. Models and Implementation Details

We test different classification models including the BERT transformer model as a baseline, comparing it to gpt3.5-turbo, which has high performance and popularity [46]. It can reply to natural language prompts based on intuitive descriptions of the downstream task and is able to return a label and free-text explanations [45]:

- BERT[2] [19], a transformer model with 110M parameters pretrained on English data using a masked language modeling (MLM) objective and finetuned on our training data, is used as a baseline. Furthermore, we compare BERT to SsciBERT [47], pre-trained on a corpus of English social science abstracts, and SciBERT [48], trained on a multi-domain corpus of scientific publications.
- chatGPT3.5[3], a chatbot based on the gpt3.5-turbo model, comprising 175B parameters and a max context length of 4,096 tokens. One of the key advantages of the gpt3.5-turbo model is its multi-turn capability, allowing it to accept a series of messages as input.

BERT was trained with the ADAM optimizer and a learning rate of $2e$-5, with an early stopping criterion of a non-decreasing validation loss for 5 epochs. The best set of parameters are selected based on the development set.
We focus on two prominent post-hoc feature attribution methods [5] and build on the open-source library *captum*[4] [49].

- **SHAP**: Shapley value-based method [38] that creates perturbations of the input text by masking out some tokens and estimating the average marginal contribution of each token across all possible perturbations by observing changes of the model's output. SHAP has a sound theoretical background and is an additive method, i.e. attributions of all features sum up to the difference between the expected prediction and the predicted output.
- **LIME**: a surrogate modeling method (Linear Interpretable Model-Agnostic Explanations) [7] that seeks to approximate the behavior of the opaque deep learning model by a linear model and determines how much each token contributed to a model's prediction.

We utilize chatGPT-3.5 API to generate explanations through greedy decoding (by setting temperature as 0) following the in-context learning paradigm. In our experiments, we compare the following prompting instructions:

[2]https://huggingface.co/bert-base-uncased
[3]https://chat.openai.com/
[4]https://github.com/pytorch/captum

- **Predict (P)**: We use the zero-shot setting with a manually-defined concise task instruction, inspired by [33], for the text classification task.
- **Label and Free-text Rationale**: We instruct our LLMs to provide extractive rationales that are grounded in the text as evidence for the prediction [50].
- **Label and Highlight Explanation**: We elicit highlight explanations using instructions inspired by [43] that also contain information about formatting the final answers. They deliver saliency scores for each input token and the predicted label as well as a confidence score.

Example prompts for each prompting strategy are shown in Table 2 and the output in Table 3, respectively. We inspect self-rationalizing models that produce outputs and explanations jointly [51], or following the predict-and-explain (P-E) paradigm [40].

Regarding the evaluation of explainability methods, Sun et al. [2023] note that 'there is no standardized metrics used to automatically evaluate model interpretability.' It is well known that LLM-generated explanations may not entail the models' predictions or might not be factually grounded in the input - even for extractive explanations [26]. Moreover, they might be plausible yet misleading to users [53, 54].
In this study, we limit ourselves to evaluating the **faithfulness** [12, 11] for the BERT family models, iteratively masking salient features from the input and measure the average drop in the model's performance [13]. Similar to [55], we evaluate gpt3.5-turbo's performance and its explainability by assessing the usefulness of explanations for humans in a qualitative study (see section 4.2.3) and refrain from assessing the self-consistency on the output level for LLMs [56].

4. Results

4.1. Classification Performance

To measure classification performance, we use standard performance metrics (e.g., accuracy, F1). Table 4 presents a summary of the performance of our models, showing that the fine-tuned BERT model achieved highest accuracy (94.3%) on the variable classification task. As can be seen, BERT-based supervised classifiers outperform chatGPT that only achieves an accuracy of 67.50% for the best prompting strategy (based on highlights) in a zero-shot setting. We observe very small standard deviations for BERT over runs with 3 different random seeds and report the average accuracy and F1-scores.

We found that confidence scores were high for all classifiers and higher confidence scores did not correlate with higher accuracy.

4.2. Results Explainability

In this study, we evaluate explanation quality mainly by its correctness, i.e. how *faithful* the explanation is w.r.t. the black box, and its *plausibility*, i.e. how reasonable it is to the user [57]. We evaluate the LIME, SHAP and chatGPT on the test data, either quantitatively and/or in a user study with domain experts for our specific application.

Table 2
Different prompting strategies.
<input> denotes the input text. <int classification> denotes the classification label, <float confidence> denotes the confidence, <float importance> denotes the saliency score

Prompt Strategy	Example Prompt
predict (P) Label	Please act as a Machine Learning Classifier and determine if the following text includes a survey variable.
predict-and-explain (P-E) Label and Free-text Rationale	You are a Machine Learning system to detect survey variables in text, a binary classification task. A text will be provided to you, and you must classify the input either 1 (refers to a variable) or 0 (does not refer to a variable), as well as your confidence in the score you chose. The confidence should be a decimal number between 0 and 1, with 0 being the lowest confidence and 1 being the highest confidence. Output this in the Python tuple format (<int classification>, <float confidence>). You must provide a free-text explanation for your prediction. It does not matter whether or not the sentence makes sense. Do your best given the input. The input will be encapsulated within <input> tags. However, these tags are not considered part of the actual content of the input. Example output: (<int classification>, <float confidence>) .
predict-and-explain (P-E) Label and Highlighted Explanation	You are a Machine Learning system to detect survey variables in text, a binary classification task. A text will be provided to you, and you must classify the review as either 1 (refers to a variable) or 0 (does not refer to a variable), as well as your confidence in the score you chose. The confidence should be a decimal number between 0 and 1, with 0 being the lowest confidence and 1 being the highest confidence. Output this in the Python tuple format (<int classification>,<float confidence>). Then, analyze how important every single word and punctuation token in the input was to your classification. The importance should be a decimal number to three decimal places ranging from -1 to 1, with -1 implying no reference to a variable and 1 implying a reference to a variable. Provide a list of (<word or punctuation>, <float importance>) for each and every word and punctuation token in the sentence in a format of Python list of tuples. Each word or punctuation is separated by a space. It does not matter whether or not the sentence makes sense. Do your best given the sentence. The input will be encapsulated within <input> tags. However, these tags are not considered part of the actual content of the input. Example output: (<int classification>, <float confidence>) [(<word or punctuation>, <float importance>), (<word or punctuation>, <float confidence>, ...]

4.2.1. Faithfulness

Faithfulness means how accurately the extracted rationales reflect the true reasoning process of the classifier to predict the class label [12]. Faithfulness metrics measure the extent to which the extracted salient tokens influence the probability for the class prediction when using these tokens as input [11]. In this work, we focus on the *sufficiency* criterion by computing the area under the threshold-performance curve as suggested by Atanasova et al. [2020]. A faithful

Table 3
For the input and promping strategy 'Highlighted Explanation' the system correctly predicts a variable (with confidence: 0.9), while the predicted label for 'Free-text Rationale' is incorrect (with confidence: 0.75)

Input	Highlighted Explanation	Free-text Rationale
"Even if one wanted to, there is no way to reduce social inequality any further than here in Germany."	("Even", 0.3), ("if", -0.1), ("one", -0.1), ("wanted", -0.1), ("to,", -0.1), ("there", -0.1), ("is", -0.1), ("no", -0.1), ("way", -0.1), ("to", -0.1), ("reduce", -0.1), ("social", 0.8), ("inequality", 0.9), ("any", -0.1), ("further", 0.1), ("than", -0.1), ("here", -0.1), ("in", -0.1), ("Germany."", 0.4)	The text does not appear to refer to a survey variable. It discusses social inequality in Germany but does not mention any specific variable or survey-related topic. Therefore, it is classified as 0 (does not refer to a variable). The confidence in this classification of class 1 is 0.75, indicating a relatively high confidence in the prediction.

Table 4
Classification Performance on the Combined English Datasets (SV-Ident 2022 and OpenMinTeD 2018)

model	f1	p	r	accuracy
`Fine-tuned BERT`	0.918	0.902	0.939	**0.943**
`Fine-tuned SciBERT`	0.931	0.918	0.947	0.931
`Fine-tuned SsciBERT`	0.916	0.907	0.927	0.931
`gpt3.5-turbo`				
`- predict (P)`	0.390	0.270	0.710	0.391
`- predict-and-explain (P-E) free-text`	0.437	0.326	0.667	0.586
`- predict-and-explain (P-E) highlights`	0.581	0.450	0.818	0.675

rationale should display low sufficiency values, as the prediction should be highly influenced by the tokens selected as most salient. Sufficiency reflects the probability change of using only rationales to predict the label and measures the degree to which the parts within the extracted rationales can allow the model to make a prediction [11]. As can be seen in Table 5, LIME and fine-tuned SciBERT achieved the best scores (41.24 AUC-TP), followed by LIME and fine-tuned BERT (44.05 AUC-TP), outperforming SHAP for all models.
However, automatic metrics for evaluating faithfulness quantitatively are limited to the traditional fine-tuning paradigm and not directly applicable to free-text explanations generated by prompting LLMs [45, 56, 58]. Therefore, in this work we use a proxy for LLM faithfulness by testing whether the predicted label and generated explanations are consistent for different prompting strategies, i.e. free-text vs. highlight-based explanations. Comparing these two strategies, we observe 85% consistent predictions.

4.2.2. Plausibility of Salient Features

We analyzed which features were most influential for the prediction using SHAP which outputs Shapley values as feature attributions which sum to the probability of the predicted class. SHAP

Table 5
Comparison of the Faithfulness of the explainability methods measured by AUC-TC for thresholds $\in [0, 10, 20, .., 100]$ defined as the average difference of the AUC before and after masking the top k% words with zero padding. Lower scores are better (↓).

Model	LIME	SHAP
`Fine-tuned BERT`	44.05	69.17
`Fine-tuned SciBERT`	**41.24**	68.78
`Fine-tuned SsciBERT`	81.33	82.39

in contrast to LIME is able to offer a global view of feature importance on the whole test corpus. Since we have no ground truth rationales, evaluation of *plausibility* [6, 11, 59] by computing the overlap between model rationales and human rationales is not possible. Instead, salient features were inspected in a retrospective way by human experts, i.e. social scientists. We would expect that the users indicate a strong preference for features that relate to social science concepts, thereby enhancing users' trust in model predictions. In contrast, non-informative tokens like stopwords might indicate overfitting of the model.
Interestingly, top-100 salient features for the variable class selected by SsciBERT and SciBERT have a relatively high overlap of 80%, e.g. *employment, participation, comparison, respondent, etc.* (see Figure 3 and Figure 5 in the appendix).

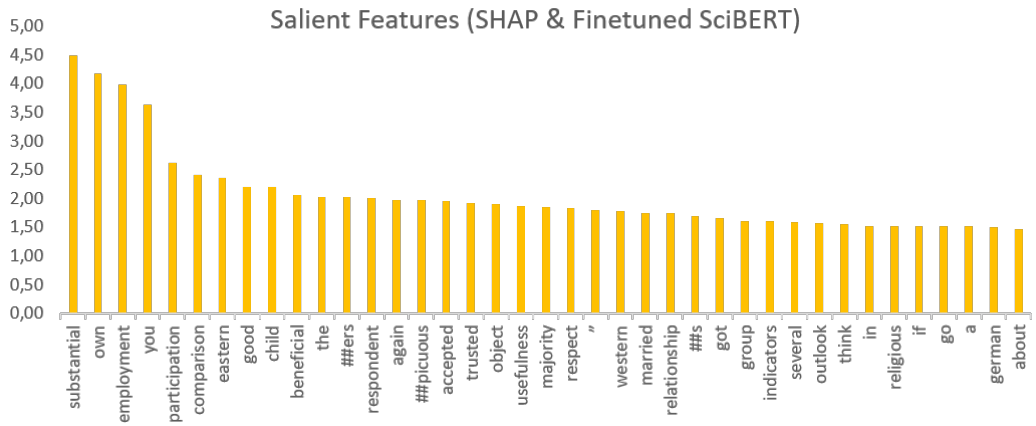

Figure 3: Most Salient Features for the 'is variable' class generated by SHAP on the test set

We also generated salient features by prompting chatGPT directly ('What are the top indicative features for the variable class in the input test corpus') (see Figure 4). In this case, the proportion of 100-top salient shared features between chatGPT and SHAP is much smaller. Overall, features responsible for predicting the variable class were considered relevant to surveys but not distinctive enough for the binary classification task.

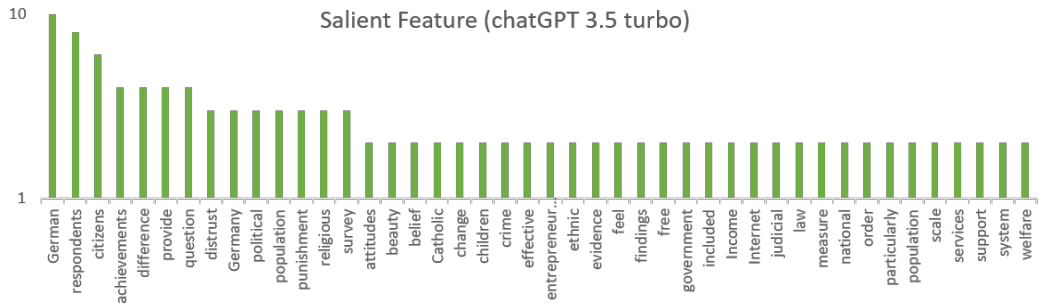

Figure 4: Most Salient Features for the 'is variable' class generated by chatGPT3.5 turbo on the test set

4.2.3. Quality of Free-text Natural Language Explanations

Natural language explanations provide an alternative to saliency maps and are generally easier for humans to interpret [44, 60]. The main goal is to improve user trust rather than giving insight into the model's inner working. Our evaluation criteria include *wellformedness* of the explanation, *consistency*, i.e. support of the label and *factual groundedness*, i.e. the explanation is correctly grounded in the input [61].

The qualitative analysis on a 5-point Likert scale reveals that explanations generated by chatGPT were well-formed (5), factual with minor deviations in the rephrasing (4.5), and in support and consistent with the predicted label (5). However, the explanations are often not very *plausible* (factual correctness and logical coherence). Also wrong predictions had convincing explanations. This is a major issue, because the variable prediction accuracy was relatively low.

For instance, the sentence *Findings suggest that distrust of the judicial system, political prioritization of law and order, and antiminority attitudes predict citizens' support for severe punishment as an effective crime-reduction technique* was misclassified as containing one or more survey variables. The self-explanation grounded the decision in salient features extracted from the text which all relate to topics addressed in the German General Social Survey (ALLBUS)[5], e.g. *questions on politics and confidence in public institutions.*

It has been stated by various authors [34] that large language models in contrast to BERT possess factual knowledge and reasoning capabilities. Yet, it is unclear to what extend survey questions form part of chatGPT's background knowledge that the model may have memorized. Therefore, it is striking that chatGPT fails to assign the correct category to some test instances which refer to Allbus variables, even though they have been cited literally. For example, the natural language justification for the sentence *Due to the presence of foreigners, I feel like a stranger in my own country* is *the statement "I feel like a stranger in my own country" expresses a personal feeling or sentiment.* which might sound plausible to the user.

Overall, the usefulness of free-text rationales based on important features has been questioned in our user study, due to the high rate of convincing explanations for false positives and negatives.

[5]https://www.gesis.org/en/allbus/

5. Discussion

Large language models (LLMs) possess AI capabilities and therefore are increasingly deployed for scholarly document understanding and scientific discovery. Recently, in-context learning for LLMs such as chatGPT has gained popularity over fine-tuning due to its simplicity and improved out-of-domain generalization.
We used chatGPT 3.5 turbo in a zero-shot setting as a baseline for detecting survey variables in scientific publications and found that prompt-engineering (e.g. predict-and-explain) can enhance the performance and reliability of the LLM model. However, due to the limits in context size, results achieved through prompting were inferior to fine-tuning a masked language model (e.g. BERT).
As NLP models are being increasingly deployed in real-world scenarios, it is imperative to consider not only their performance in terms of accuracy, but also the trustworthiness of the ML model. Given the complexity of ML models which may contain 175 billion parameters, it becomes almost impossible for humans to interpret model internals. Our experiments showed that the explanation quality of BERT-based models varies substantially, with SciBert performing best in terms of faithfulness. Regarding plausibility, we found that top-100 salient features often refer to social science concepts, thus highly informative to the user. In general, the explanations given by ChatGPT align with the original text when using the predict-and-explain prompting strategy, but the model has a tendency towards overconfidence, even for false predictions.
In future work, we like to adapt LLMs to individual contexts by instruction-fine-tuning, aligning them with the task objectives of the social science community.

6. Acknowledgments

Acknowledgments

This work was supported by the DFG project VADIS under grant numbers: ZA 939/5-1, PO 1900/5-1, EC 477/7-1, KR 4895/3-1.

7. Appendices

References

[1] A. Zielinski, P. Mutschke, Towards a gold standard corpus for variable detection and linking in social science publications, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1084.

[2] Y. S. Kartal, S. Takeshita, T. Tsereteli, K. Eckert, H. Kroll, P. Mayr, S. P. Ponzetto, B. Zapilko, A. Zielinski, Towards automated survey variable search and summarization in social science publications, 2022. arXiv:2209.06804.

[3] Y. S. Kartal, M. A. Shahid, S. Takeshita, T. Tsereteli, A. Zielinski, B. Zapilko, P. Mayr, Vadis – a variable detection, interlinking and summarization system, 2023. URL: https://api.semanticscholar.org/CorpusID:266435862.
[4] T. Tsereteli, Y. S. Kartal, S. P. Ponzetto, A. Zielinski, K. Eckert, P. Mayr, Overview of the sv-ident 2022 shared task on survey variable identification in social science publications, arXiv preprint arXiv:2209.09062 (2022).
[5] M. Danilevsky, S. Dhanorkar, Y. Li, L. Popa, K. Qian, A. Xu, Explainability for natural language processing, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 4033–4034.
[6] G. Chrysostomou, N. Aletras, An empirical study on explanations in out-of-domain settings, arXiv preprint arXiv:2203.00056 (2022).
[7] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[8] C. Molnar, Interpretable Machine Learning, 2 ed., 2022. URL: https://christophm.github.io/interpretable-ml-book.
[9] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).
[10] L. Shapley, Quota solutions op n-person games1, Edited by Emil Artin and Marston Morse (1953) 343.
[11] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.
[12] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205. URL: https://aclanthology.org/2020.acl-main.386. doi:10.18653/v1/2020.acl-main.386.
[13] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3256–3274. URL: https://aclanthology.org/2020.emnlp-main.263. doi:10.18653/v1/2020.emnlp-main.263.
[14] C. S. Chan, H. Kong, G. Liang, A comparative study of faithfulness metrics for model interpretability methods, arXiv preprint arXiv:2204.05514 (2022).
[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837.
[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal,

K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[18] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).
[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[20] J. W. . M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, CoRR abs/2109.01652 (2021). URL: https://arxiv.org/abs/2109.01652. arXiv:2109.01652.
[21] Z. Li, The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination, arXiv preprint arXiv:2304.14347 (2023).
[22] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy llms: a survey and guideline for evaluating large language models' alignment, arXiv preprint arXiv:2308.05374 (2023).
[23] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, K. Malkan, Wt5?! training text-to-text models to explain their predictions, arXiv preprint arXiv:2004.14546 (2020).
[24] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).
[25] N. Feldhus, L. Hennig, M. D. Nasert, C. Ebert, R. Schwarzenberg, S. Möller, Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods, 2023. arXiv:2210.07222.
[26] X. Ye, G. Durrett, The unreliability of explanations in few-shot prompting for textual reasoning, Advances in neural information processing systems 35 (2022) 30378–30392.
[27] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[28] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, arXiv preprint arXiv:2304.13712 (2023).
[29] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023).
[30] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, R. McHardy, Challenges and applications of large language models, 2023. arXiv:2307.10169.
[31] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, L. He, A survey on text classification: From traditional to deep learning, ACM Transactions on Intelligent Systems and Technology (TIST) 13 (2022) 1–41.
[32] L. Gao, D. Ghosh, K. Gimpel, The benefits of label-description training for zero-shot text classification, arXiv preprint arXiv:2305.02239 (2023).
[33] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, G. Wang, Text classification via large language models, arXiv preprint arXiv:2305.08377 (2023).
[34] Q. Zhong, L. Ding, J. Liu, B. Du, D. Tao, Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arxiv, 2023.
[35] X. Sun, D. Yang, X. Li, T. Zhang, Y. Meng, H. Qiu, G. Wang, E. Hovy, J. Li, Interpreting deep learning models in natural language processing: A review, arXiv preprint arXiv:2110.10470

(2021).
[36] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
[37] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences, arXiv preprint arXiv:1605.01713 (2016).
[38] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[39] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Advances in neural information processing systems 29 (2016).
[40] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).
[41] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).
[42] A. Marasović, I. Beltagy, D. Downey, M. E. Peters, Few-shot self-rationalization with natural language prompts, in: NAACL-HLT, 2021. URL: https://api.semanticscholar.org/CorpusID:244130199.
[43] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, L. H. Gilpin, Can large language models explain themselves? a study of llm-generated self-explanations, arXiv preprint arXiv:2310.11207 (2023).
[44] S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable nlp, ArXiv abs/2102.12060 (2021).
[45] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, arXiv preprint arXiv:2309.01029 (2023).
[46] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. arXiv:2303.18223.
[47] S. Shen, J. Liu, L. Lin, Y. Huang, L. Zhang, C. Liu, Y. Feng, D. Wang, Sscibert: A pre-trained language model for social science texts, Scientometrics 128 (2023) 1241–1263.
[48] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[49] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for pytorch, 2020. URL: https://arxiv.org/abs/2009.07896. doi:10.48550/ARXIV.2009.07896.
[50] A. K. Lampinen, I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, F. Hill, Can language models learn from explanations in context?, arXiv preprint arXiv:2204.02329 (2022).
[51] F. Rancourt, P. Vondrlik, D. Maupomé, M.-J. Meurs, Investigating self-rationalizing models for commonsense reasoning, Stats 6 (2023) 907–919.
[52] J. Sun, C. Shaib, B. C. Wallace, Evaluating the zero-shot robustness of instruction-tuned language models, arXiv preprint arXiv:2306.11270 (2023).

[53] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, arXiv preprint arXiv:2305.04388 (2023).
[54] Y. Chen, R. Zhong, N. Ri, C. Zhao, H. He, J. Steinhardt, Z. Yu, K. McKeown, Do models explain themselves? counterfactual simulatability of natural language explanations, arXiv preprint arXiv:2307.08678 (2023).
[55] B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, S. Zhang, Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness, arXiv preprint arXiv:2304.11633 (2023).
[56] L. Parcalabescu, A. Frank, On measuring faithfulness of natural language explanations, arXiv preprint arXiv:2311.07466 (2023).
[57] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42.
[58] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, I. Augenstein, Faithfulness tests for natural language explanations, arXiv preprint arXiv:2305.18029 (2023).
[59] A. Madsen, N. Meade, V. Adlakha, S. Reddy, Evaluating the faithfulness of importance measures in nlp by recursively masking allegedly important tokens and retraining, arXiv preprint arXiv:2110.08412 (2021).
[60] J. Colin, T. Fel, R. Cadène, T. Serre, What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods, Advances in Neural Information Processing Systems 35 (2022) 2832–2845.
[61] D. Li, J. Yu, B. Hu, Z. Xu, M. Zhang, Explaincpe: A free-text explanation benchmark of chinese pharmacist examination, arXiv preprint arXiv:2305.12945 (2023).

A. Appendix

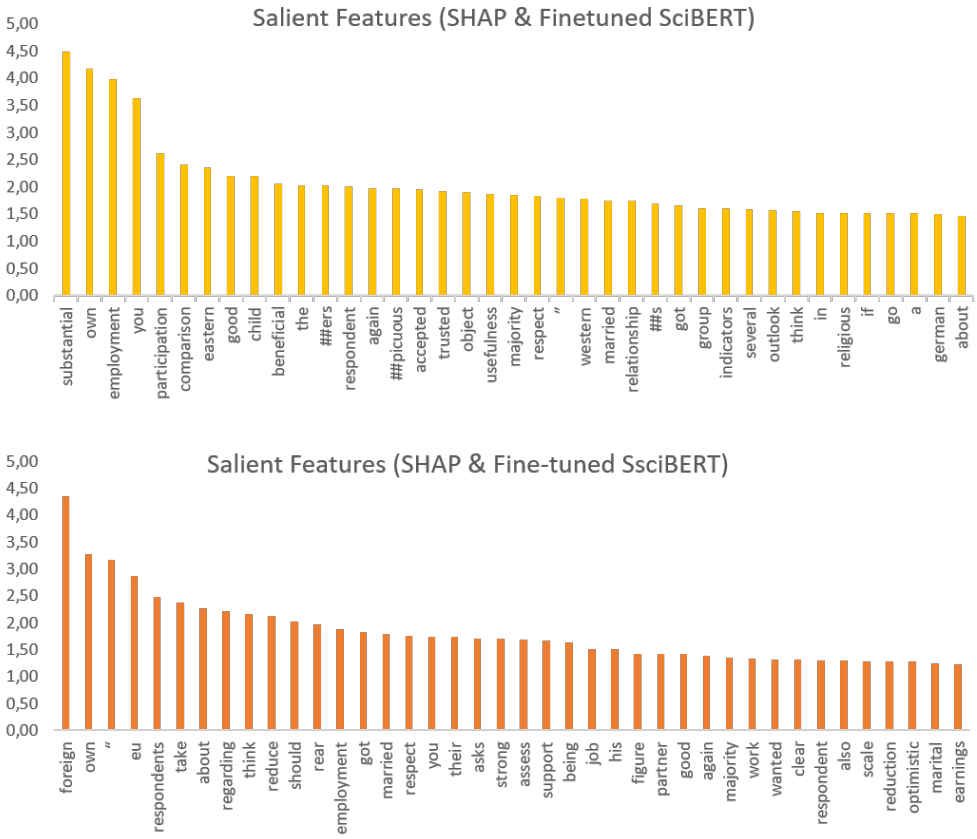

Figure 5: Most Salient Features for the 'is variable' class generated by SHAP on the test set