An Eye Opener Regarding Task-Based Text Gradient Saliency

Anonymous ACL submission

Abstract

Eye movements in reading reveal humans' cognitive processes during language understanding. As such, the time a reader's eyes dwell on a token has been utilized as a measure for the visual attention paid to that word, or the importance of that word to the reader. This study investigates the alignment of the importance attributed to input tokens by language models (LMs) on the one hand and humans, in the form of fixation durations, on the other hand. While previous research on the internal processes of LMs have employed the models' attention weights, recent studies have argued in favor of gradient-based methods. Moreover, previous approaches to interpret LMs' internals with human gaze have neglected the tasks readers performed during reading, even though psycholinguistic research underlines that reading patterns are task-dependent. We thus introduce a novel approach that employs a gradient-based saliency method designed to emulate task-specific human reading strategies to align model and human importance, and we find that task specificity plays a crucial role in this alignment.

1 Introduction

Human eye movements during reading reflect cognitive processes involved in language processing (Just and Carpenter, 1980; Rayner, 1998): the fixation duration on a word correlates with reading comprehension (Rayner, 1977; Malmaud et al., 2020a). As such, fixation duration has been employed as proxy of the relative importance of a word to a reader (Hollenstein and Beinborn, 2021).

The introduction of neural attention mechanisms (Bahdanau et al., 2014) and the Transformer architecture (Vaswani et al., 2017), which relies on self-attention to compute input and output representations, has given fresh impetus to research into how language models (LMs) process language. Attention mechanisms assign dynamic weights to input tokens, offering a method to understand a model's internal functioning and decision-making processes (Wang et al., 2016; Ghaeini et al., 2018).

Recent research has compared model and human language comprehension by aligning model attention weights with human reading metrics, such as fixation durations (Sood et al., 2020; Eberle et al., 2022; Bensemann et al., 2022), presuming model attention effectively signifies the relative importance of input tokens. However, research on attention (Jain and Wallace, 2019; Serrano and Smith, 2019; Brunner et al., 2019) has questioned the reliability of attention weights in accurately reflecting token significance, labeling attention as a contentious issue in interpretability discussions (Bastings and Filippova, 2020). Alternative approaches like gradient-based saliency (Simonyan et al., 2014; Li et al., 2016), which assess the impact of input tokens on model predictions through gradients, are proposed for better assessing token importance. Building on this, Hollenstein and Beinborn (2021) incorporated a saliency method by correlating gradient saliencies, obtained through iterative token masking and gradient computation, with human fixation durations. However, the output space of this approach comprises tens of thousands of tokens, which makes gradient-based saliency methods uninformative (Yin and Neubig, 2022). Moreover, the model and the humans did not partake in the same task when processing the text, which is a crucial discrepancy, as psycholinguistic studies show that human reading strategies vary with the task and differ from normal reading (Shubi and Berzak, 2023; Mézière et al., 2023; Malmaud et al., 2020b).

To address this, we propose a novel gradient-based saliency approach that replicates the classification tasks humans perform during task-specific reading to better align the importance LMs and humans assign to tokens. Additionally, we expand our analysis to include decoder-based LMs, which, due to their auto-regressive nature, align more closely with the incremental nature of human processing.

2 Related Work

Model attention and human attention Research comparing model attention to human visual attention, using fixation locations and durations as proxies, has produced mixed findings. Sood et al. (2020) observed distinct differences between transformer LM attention patterns and human fixation patterns. Conversely, studies by Eberle et al. (2022) and Bensemann et al. (2022) found strong correlations between early transformer layer attention weights, like those in BERT (Devlin et al., 2019), and human visual attention, contrasting with earlier results. This discrepancy can be attributed to methodological differences in processing attention weights: Sood et al. (2020) analyzed maximum attention values from the last layer's sub-word tokens, while Bensemann et al. (2022) averaged attention across sub-word tokens in the first layer.

Limitations of attention-based interpretation The inconsistent results outlined above challenge the usefulness of methods based on model attention to investigate the internals of LMs. Indeed, Brunner et al. (2019) emphasize the lack of token identifiability as one moves to higher layers of a model, and Abnar and Zuidema (2020) show that distinct attention patterns are only found in earlier layers, while in higher layers the attention weights approximate a uniform distribution. Moreover, Jain and Wallace (2019) question whether attention weights can reliably identify the relative importance of inputs to the entire model, showing that different attention distributions yield equivalent model predictions. Similarly, Serrano and Smith (2019) find attention weights to be very noisy indicators of importance. Finally, an analysis of BERT's (Devlin et al., 2019) attention (Clark et al., 2019) reveals a significant focus on the [SEP] token, which does not affect model outputs when its attention is altered, suggesting a "no-op" operation. Similarly, research on attention heads (Voita et al., 2019; Michel et al., 2019) finds that many of them can be pruned with minimal impact, further indicating the potential redundancy or non-operational nature of certain attention mechanisms.

Saliency-based methods for analyzing LMs with human gaze As saliency-based methods are arguably more suited than methods based on attention (Bastings and Filippova, 2020) for model analysis, Hollenstein and Beinborn (2021) extract token importance by iteratively masking each input token, computing the L2 norm of the gradient for the correct output with respect to each token, and then summing all saliency scores for each input token. However, while they do emulate the LM's pre-training objective, this does does not necessarily align with human processing: whereas the model "sees" the input only partially, and as many times as there are tokens, the readers see the input fully and only once. Moreover, the gaze data used in their study was, in parts, recorded while participants were completing a task, such as sentiment analysis and relation extraction (i.e., task-specific reading). In our approach, we thus compute gradients by having the model perform the same kind of classification task that humans performed during reading. Thereby the token importance attributed by both humans and the model refers to the importance within the constraint of a specific task, and the model sees the input only once, and fully.

3 Method

Consider an input sentence, formalized as $\mathbf{x} = \langle x_1, \ldots, x_N \rangle$ of N tokens, where x_j is the j^{th} token (word) in the sentence, and two corresponding token importance vectors of the same length: the *human importance* vector $\mathbf{h} = \langle h_1, \ldots, h_N \rangle$ and the *model importance* vector $\mathbf{m} = \langle m_1, \ldots, m_N \rangle$, where h_j and m_j are the human and model importance attributed to token x_j. We obtain the mean Spearman correlation between model and human importance by computing the by-token Spearman correlations between the vectors $\mathbf{m}$ and $\mathbf{h}$ for all sentences $\mathbf{x}$, then dividing the sum of these correlations by the number of sentences $\mathbf{x}$.

Extracting model importance: gradient-based saliency The *model importance* vector $\mathbf{m}$ consists of gradient saliency values m_j for each input token x_j of the sentence $\mathbf{x}$. "Saliency" refers to neural network interpretation methods that assign an importance distribution over the input in order to analyse a network's prediction (Ding and Koehn, 2021). In other words, saliency methods aim at explaining how sensitive the decision of a model is to changes in the input. The most common method of assigning this importance distribution is by means of the gradient (Simonyan et al., 2014). Given a parametrized language model f_θ, we compute the gradient g with respect to an input token $x_j \in \mathbf{x}$ as

$$g(x_j) := \frac{\partial f_\theta^c}{\partial x_j}(\mathbf{x}), \tag{1}$$

	BERT *base*	BERT *large*	RoBERTa	DistilBERT	GPT-2 *base*	GPT-2 *large*	OPT
				Sentiment Analysis (SA)			
fine-tuned	0.61$_{0.010}$	*0.57*$_{0.011}$	*0.47*$_{0.012}$	*0.53*$_{0.011}$	0.49$_{0.011}$	0.55$_{0.010}$	*0.43*$_{0.012}$
pre-trained (0-shot)	0.55$_{0.011}$	0.59$_{0.010}$	0.45$_{0.012}$	0.52$_{0.012}$	0.4$_{0.014}$	0.48$_{0.012}$	0.42$_{0.013}$
random init. (0-shot)	0.24$_{0.013}$	0.22$_{0.013}$	0.04$_{0.014}$	0.21$_{0.013}$	0.2$_{0.014}$	0.19$_{0.014}$	0.15$_{0.015}$
				Relation Extraction (RE)			
fine-tuned	*0.53*$_{0.010}$	0.52$_{0.009}$	0.42$_{0.010}$	0.45$_{0.010}$	0.46$_{0.010}$	0.52$_{0.009}$	0.5$_{0.011}$
pre-trained (0-shot)	0.51$_{0.010}$	0.47$_{0.011}$	0.37$_{0.011}$	0.49$_{0.010}$	0.37$_{0.011}$	0.45$_{0.011}$	0.42$_{0.011}$
random init. (0-shot)	0.08$_{0.011}$	0.07$_{0.011}$	0.04$_{0.012}$	0.09$_{0.011}$	0.16$_{0.013}$	0.16$_{0.013}$	0.14$_{0.014}$

Table 1: We report mean Spearman correlations and standard errors between model and human importance for all models in their *fine-tuned*, *pre-trained (0-shot)*, and *randomly initialized (0-shot)* version, for both tasks SA and RE. The difference in correlations is significant in all cases except for the ones indicated in italic.

where c indexes the true class y in the model's output, and f_θ^c refers to the predicted output logit for the true class y. We then follow Li et al. (2016) by defining the gradient saliency m_j of token x_j as the L1 norm of its gradient $m_j := |g(x_j)|$. Since different LMs employ different tokenization methods which split tokens into sub-word tokens (Sennrich et al., 2016; Song et al., 2021), we pool gradients back to token level by summing up the sub-word token-level gradient norms. We then normalize the token-level saliencies by dividing them by the sum of all saliency values in the sentence.

Extracting human importance: relative fixation duration To obtain the *human importance* vector $\mathbf{h}$, we first extract raw total fixation durations $t_{j,r}$ for each token $x_j \in \mathbf{x}$, which is the sum of the durations of all fixations on that token by a reader r. However, due to variations in reading speed across readers and sentences, these raw durations can vary significantly between instances. We thus normalize them by dividing them by the sum of durations across all tokens within a sentence, resulting in *relative fixation durations* $d_{j,r} = t_{j,r}/\sum_j t_{j,r}$ for each token x_j. These relative durations are then averaged across all readers to bypass individual differences and to obtain a more robust signal, resulting in aggregated relative fixation durations $h_j = \sum_r d_{j,r}/\,|\text{readers}|$ for each token x_j.

4 Experiments[1]

Datasets The eye-tracking part of the *Zurich Cognitive Language Processing Corpus* (ZuCo; Hollenstein et al., 2018) comprises two task-specific readings: in the sentiment analysis (SA) reading, participants were presented with a subset from the *Stanford Sentiment Treebank* (SST; Socher et al., 2013) that consists of movie reviews, based on which they had to rate the movies; in the relation extraction (RE) reading, they performed relation extraction on a subset of sentences from the *Wikipedia relation extraction corpus* (Culotta et al., 2006).

Models and fine-tuning We include both encoder models and decoder models, as well as models from the same family but different in size. Encoders include BERT (Devlin et al., 2019) *base* and *large*, RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019); decoders include GPT-2 (Radford et al., 2019) *base* and *large*, and OPT (Zhang et al., 2022). As the models perform classification — ternary for SA, and 9-class for RE —, we utilize the architecture variants implemented for sequence classification in Huggingface (Wolf et al., 2019). For SA, we fine-tune the models on the SST dataset and for RE on the *Wikipedia* dataset (Culotta et al., 2006) , excluding the sentences used for ZuCo SA and RE, respectively.[2]

Baselines. We include two sets of baseline models: the above-mentioned models randomly initialized (*random* (0-shot)), and the models pre-trained but not fine-tuned (*pre-trained* (0-shot)).

Results As depicted in Table 1, the more similar the model's training is to the human task, the more aligned are the model and human importance vectors. There exist medium to strong correlations between the fine-tuned *model importance* and *human importance* vectors, exemplified by correlations of 0.61 by BERT *base* or 0.55 by GPT-2 *large* for SA. Additionally, most *fine-tuned* models produce significantly higher correlations than the *pre-trained* baselines, and the pre-trained models all have significantly higher correlations than their randomly initialized counterparts. Encoder models, on average, achieve higher correlations than decoders, despite variability within both types. Additionally, SA task model importance correlates more strongly on average than for RE.

[1]Our code is available at `anonymous-link`.

[2]For training and implementation details as well as classification test results, see Appendix A.

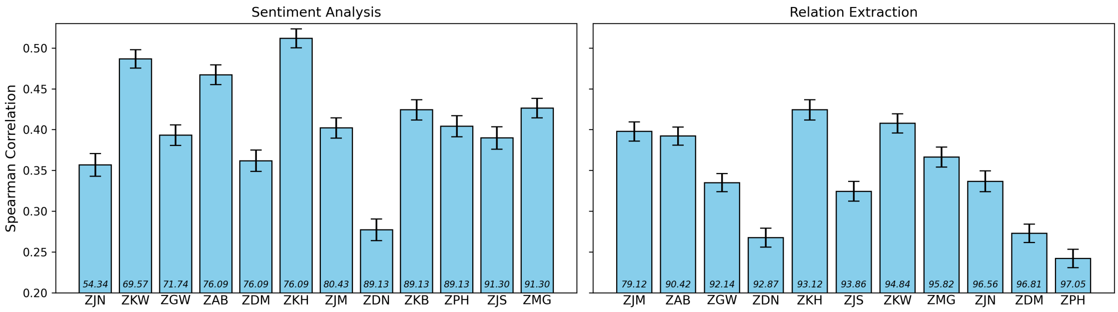


Figure 1: Mean Spearman correlations between relative fixation durations and gradient saliencies for fine-tuned BERT *base* are depicted at the participant level, with error bars denoting the standard error. Participants are arranged according to task accuracy, with their average task accuracies presented at the bottom of each bar.

5 Participant-level analysis

To investigate whether the models correlate more with certain participants, we perform an additional participant-level analysis in which we compute correlations between the model-extracted saliency values and relative fixation durations for each participant individually. We also extract the participants' response accuracies for both their SA and RE, averaged over sentences. The underlying intuition is that the models possibly correlate more with participants that have a higher task accuracy.

Results The juxtaposition of correlations on participant level and participants' accuracies reveals no discernible pattern, as exemplified by BERT *base* in Figure 1. The correlation coefficients between participants exhibit great variability in both tasks. Participants' task accuracies are distributed across a wide range for SA but exhibit a ceiling effect for RE. Moreover, averaging the participant-level correlations yields lower correlation values than using the aggregate relative fixation durations, e.g., the group-level correlation with BERT *base* is 0.61 and the average on participant-level is 0.41.[3]

6 Discussion and Conclusion

The experimental results find medium to strong correlations between model importance vectors, derived from gradient saliencies, and human importance vectors, indicated by relative fixation durations, particularly when language models (LMs) are fine-tuned for tasks mirroring those undertaken by readers: task-specific fine-tuned models demonstrate notably stronger correlations than pre-trained zero-shot baselines. The discrepancy between the pre-trained and randomly initialized models suggests an initial understanding for human importance attribution acquired during pre-training. These findings underline the importance of matching tasks between models and humans for accurate gaze analysis, with task-specificity influencing reading behavior but remaining largely ignored in NLP (Shubi and Berzak, 2023). We further find that SA tasks show consistently higher correlations than RE, possibly due to the complexity introduced by more output classes affecting model predictions. Moreover, initial observations suggest encoders outperform decoders in correlation, potentially due to decoders' unsuitability for classification tasks. Yet, this distinction may be incidental, influenced by factors like pre-training data or model architecture. Surprisingly, BERT *base* yields the highest correlation, while BERT *large* and RoBERTa, who achieve higher test accuracies than BERT, produce lower correlations. This indicates that emulating human importance attribution is neither a function of model parameters nor does it necessarily imply better model performance. The participant-level analysis reveals no distinct pattern, indicating that the models do not mirror the token importance attribution of more proficient humans. Moreover, averaging correlations across individual participants results in a lower correlation value compared to when participant fixation durations are aggregated across sentences. This implies both that by-participant aggregation of relative fixation durations produces a more robust signal, and that models correlate more with average human language processing than with subject-level idiosyncracies.

In conclusion, we have developed a gradient saliency-based method to analyze LMs with human gaze that does not neglect task-specificity and found that mirroring tasks yields higher correlations.

[3]An overview of all by-participant accuracies and correlations, for all models can be found in Table 3 in Appendix B.

Limitations

First of all, the number of sentences in the eye gaze dataset is quite low, as is the number of readers (which are all L1 English readers based in Zurich, and are not experts in sentiment analysis or relation extraction), which does not make for a representative sample of the population at large.

Relatedly, for a more extensive evaluation of our task-specific approach, one would have to apply it to the same sentences that contain eye movements from natural reading instead of task-specific reading. We leave it to future work to extend the data from ZuCo with eye movements from natural reading.

Moreover, while the studies outlined in Section 2 underline the superiority of gradient-based over attention-based methods, they might still not be the state-of-the-art for explainability methods and one might employ methods such as Integrated Gradients or Layer-wise Relevance Propagation.

Ethics Statement

Working with human data requires careful ethical considerations. The eye-tracking dataset utilized in this study follows ethical standards and has been approved by the responsible ethics committees. It is licensed under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0).

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020a. Bridging information-seeking human gaze and machine reading comprehension. *arXiv preprint arXiv:2009.14780*.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020b. Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152, Online. Association for Computational Linguistics.

Diane C Mézière, Lili Yu, Erik D Reichle, Titus Von Der Malsburg, and Genevieve McArthur. 2023. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58(3):425–449.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one?

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates, Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & cognition*, 5(4):443–448.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Omer Shubi and Yevgeni Berzak. 2023. Eye movements in information-seeking reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

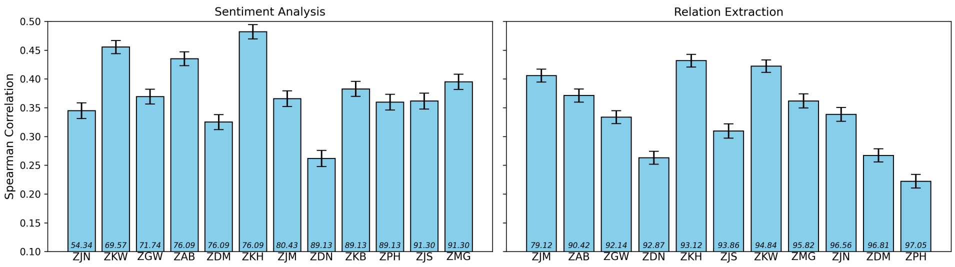
A Fine-Tuning Details

We fine-tune the models outlined in Section 3 on the SST (Socher et al., 2013) dataset for ternary sentiment classification, excluding the sentences used for ZuCo SA, and on the *Wikipedia* dataset (Culotta et al., 2006) for 9-class relation classification, excluding the sentences used for ZuCo RE. After excluding sentences from ZuCo SA and RE, we are left with 5211 sentences allocated for SA and 889 sentences allocated for RE. Subsequently, we implement an 80/20 split for training and validation. For testing, there are 400 sentences from ZuCo SA and 335 sentences from ZuCo RE [4]. We train the models for 10 epochs, with an early stopping patience of 3 epochs, using the AdamW (Loshchilov and Hutter, 2019) optimizer, a learning rate of $2 * 10^{-5}$, and a batch size of 16. All models are implemented in PyTorch (Paszke et al., 2019).

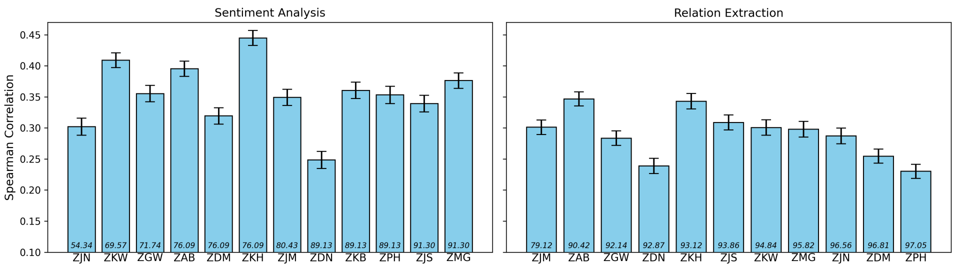
	BERT *base*	BERT *large*	RoBERTa	DistilBERT	GPT-2 *base*	GPT-2 *large*	OPT
SA	75.3	76.5	82.8	75.0	71.8	77.8	73.8
RE	57.9	61.2	57.9	60.9	53.1	56.1	55.2

Table 2: We report the accuracy of fine-tuning the models on the SST (Socher et al., 2013) for sentiment analysis (SA) and on the *Wikipedia* dataset (Culotta et al., 2006) for relation extraction (RE). In both cases, the ZuCo SA and RE sentences are excluded from the training data; the models are tested on the ZuCo sentences for SA and RE.

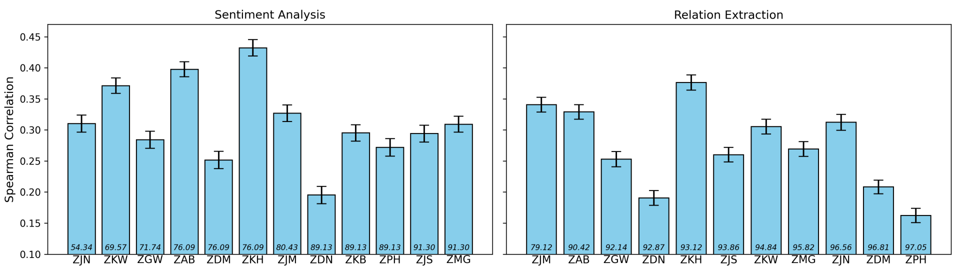
B Participant-Level Analysis



(a) BERT *large*

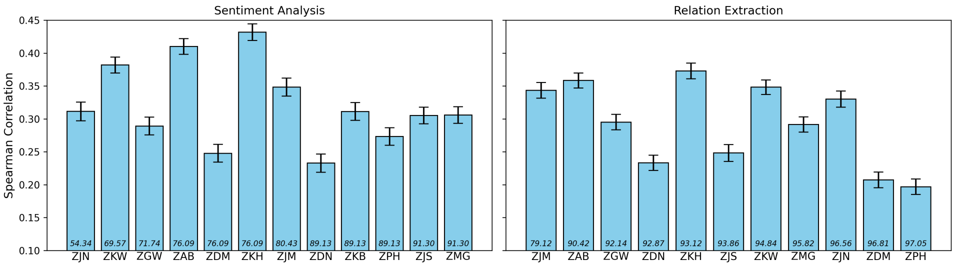


(b) DistilBERT

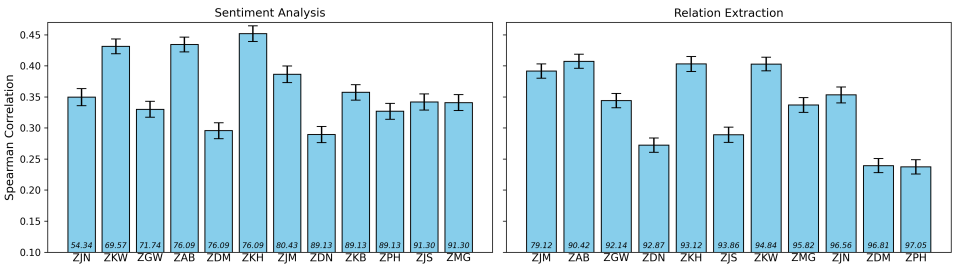


(c) RoBERTa

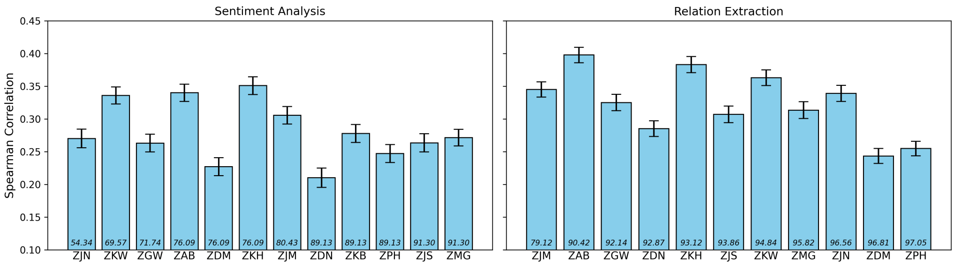
[4] Out of the original 407 sentences in ZuCo RE, we retain only 335 sentences that contain a specific relation.



(d) GPT-2 *base*



(e) GPT-2 *large*



(f) OPT

Figure 1: Spearman correlations between relative fixation durations and gradient saliencies for various models are depicted at the participant level, including standard error. Participants are arranged according to task accuracy, with their accuracy values presented at the bottom of each bar.

	ZAB	ZDM	ZDN	ZGW	ZJM	ZJN	ZJS	ZKB	ZKH	ZKW	ZMG	ZPH	avg
						Sentiment Analysis (SA)							
Task acc	76.09	76.09	89.13	71.74	80.43	54.34	91.3	89.13	76.09	69.57	91.3	89.13	79.53
BERT *base*	0.47	0.36	0.28	0.39	0.40	0.36	0.39	0.42	0.51	0.49	0.43	0.40	0.41
BERT *large*	0.44	0.33	0.26	0.37	0.37	0.34	0.36	0.38	0.48	0.46	0.39	0.36	0.38
DistilBERT	0.40	0.32	0.25	0.36	0.35	0.30	0.34	0.36	0.44	0.41	0.38	0.35	0.35
RoBERTa	0.4	0.25	0.2	0.28	0.33	0.31	0.29	0.3	0.43	0.37	0.31	0.27	0.31
GPT-2 *base*	0.41	0.25	0.23	0.29	0.35	0.31	0.31	0.31	0.43	0.38	0.31	0.27	0.32
GPT-2 *large*	0.43	0.3	0.29	0.33	0.39	0.35	0.34	0.36	0.45	0.43	0.34	0.33	0.36
OPT	0.34	0.23	0.21	0.26	0.31	0.27	0.26	0.28	0.35	0.34	0.27	0.25	0.28
						Relation Extraction (RE)							
Task acc	90.42	96.81	92.87	92.14	79.12	96.56	93.86	95.33	93.12	94.84	95.82	97.05	93.16
BERT *base*	0.39	0.27	0.27	0.34	0.40	0.34	0.32	–	0.42	0.41	0.37	0.24	0.34
BERT *large*	0.37	0.27	0.26	0.33	0.41	0.34	0.31	–	0.43	0.42	0.36	0.22	0.34
DistilBERT	0.35	0.25	0.24	0.28	0.30	0.29	0.31	–	0.34	0.30	0.30	0.23	0.29
RoBERTa	0.33	0.21	0.19	0.25	0.34	0.31	0.26	–	0.38	0.31	0.27	0.16	0.27
GPT-2 *base*	0.36	0.21	0.23	0.30	0.34	0.33	0.25	–	0.37	0.35	0.29	0.20	0.29
GPT-2 *large*	0.41	0.24	0.27	0.34	0.39	0.35	0.29	–	0.4	0.4	0.34	0.24	0.33
OPT	0.4	0.24	0.29	0.33	0.35	0.34	0.31	–	0.38	0.36	0.31	0.25	0.32

Table 3: The participants' task accuracy and their Spearman correlations with the LMs are reported. There is a lack of correlations for one participant in the RE task because of a pre-processing issue with the eye-tracking data.