Eye-Tracking Features Masking Transformer Attention in Question-Answering Tasks

Leran Zhang[1], **Nora Hollenstein**[1,2]
[1] University of Copenhagen
[2] University of Zurich
khloezhanglr@outlook.com, nora.hollenstein@uzh.ch

Abstract

Eye movement features are considered to be direct signals reflecting human attention distribution with a low cost to obtain, inspiring researchers to augment language models with eye-tracking (ET) data. In this study, we select first fixation duration (FFD) and total reading time (TRT) as the cognitive signals to guide Transformer attention in question-answering (QA) tasks. We design three different ET attention masks based on the two features, either collected from human reading events or generated by a gaze-predicting model. We augment BERT and ALBERT models with attention masks structured based on the ET data. We find that augmenting a model with ET data carries linguistic features complementing the information captured by the model. It improves the models' performance but compromises the stability. Different Transformer models benefit from different types of ET attention masks, while ALBERT performs better than BERT. Moreover, ET data collected from real-life reading events has better model augmenting ability than the model-predicted data.

Keywords: Eye-tracking augmented, Transformer, attention, question answering

1. Introductionn

Language inference tasks (NLI) have been designed to examine whether models can comprehend language and extract desired information, while language model (LM) performance continuously approaches human performance in these tasks according to the past decades of natural language processing (NLP) studies. From recurrent neural networks (RNN) models to Transformers (Devlin et al., 2019; Lan et al., 2019; Li and Rudzicz, 2021), progress has been achieved in making models 'think'. Other than scaling the models, researchers have also attempted to seek augmenting methods based on human attention.

Reading is an essential event in human language processing, and eye movement features reflect human attention activity during the event(Rayner, 2009). Eye-tracking (ET) data is the eye movement information captured during reading events along with time and positional information; researchers explore human attention distribution based on ET data (Bicknell et al., 2008; Snell and Theeuwes, 2020), and adapt it to inspire the development of LMs (Hollenstein and Zhang, 2019; Hollenstein et al., 2019a; Zhao et al., 2023). As a direct indicator of human attention in reading activities, ET data have been actively applied in enhancing LMs for downstream NLP tasks, such as named entity recognition and sentiment analysis (Barrett et al., 2018; Barrett and Hollenstein, 2020; Hollenstein and Zhang, 2019), and positive results have been achieved.

Most of the ET data augmenting research was conducted based on RNN models over decades, yet few attempts have been made to introduce this cognitive signal to Transformer models. Moreover, current cognitive-related studies focus mostly on classification and annotation rather than reading comprehension tasks. Evidence has shown that 'dwell times of human eye movements were strongly correlated with the attention patterns occurring in the early layers of pre-trained Transformers such as BERT' (Bensemann et al., 2022), therefore, great potential is expected in guiding Transformer attention with ET data for downstream tasks.

In this paper, ET data is directly introduced into Transformer attention blocks during the fine-tuning process to see if human attention can augment model performance in reading-based question-answering (QA) tasks. Specifically, our experiment examines whether both the ET data collected from human reading events and generated by a gaze-predicting model can enhance Transformer performance equally in QA tasks.

Contribution: In this study, we reveal the following achievements:

- Due to different hidden representative transferring mechanisms, the ALBERT model benefits more from adding ET attention masks than BERT. Specifically, augmented ET data can better enhance ALBERT's performance, while weakened data fits BERT better.
- Compared with real-life ET data, the linguistic features contained in model-predict ET data are relatively limited. While models benefit from the former, the advantage brought by the latter is not significant.

LREC-COLING 2024, pages 7057–7070

- Introducing eye-movement features containing either low-level or high-level linguistic features to a layer carrying the corresponding level of information enhances the attention distribution on that layer. While introducing ET data improves a model's performance scores, it may impact the model's stability.

The source code is available online[1].

2. Related Work

This research focuses on augmenting LMs with ET data, lying at the intersection of NLP and computational cognitive science. Below we outline the related works in the corresponding fields.

2.1. Language Models in Language Inference Tasks

NLI tasks are NLP tasks stated in natural language and closely relate to lexical, semantic, and pragmatic analysis (MacCartney, 2009). Early inference tasks focused mostly on annotation and ground truth extraction like part-of-speech tagging (Marcus et al., 1993) and named entity recognition (Grishman and Sundheim, 1996). Later, advanced inference tasks required an understanding of external linguistic knowledge, such as word meaning, grammar, syntax, semantics, and discourse structure for the detection of the relation between words and sentences. Benchmarks of specific tasks were established to standard the examination of model performance, among which QA benchmarks, requesting abilities to disambiguate the text and extract the necessary information to solve the puzzle, became one of the most well-established branches (Storks et al., 2019) with diverse forms (Lai et al., 2017; Hill et al., 2016; Rajpurkar et al., 2016; Rayner et al., 2006; Choi et al., 2018; Christmann et al., 2019; Thomas et al., 2017).

Transformer models are a popular type of model applied in multiple machine learning studies, with BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Li and Rudzicz, 2021), etc. as the most outstanding representatives. Compared with RNN models, Transformer models conquer the limitation of sequential processing, generate structural representation to reflect the syntax tree (Henderson, 2020) and establish strong ability in parallel computation (Chaudhari et al., 2021). They encode context information in the hidden representatives, with the self-attention mechanism unifying the cross attention between two sentences (Shi et al., 2021) and providing contextual information in the input sequence in the map (Yun et al., 2019). These powerful models have advanced multiple state-of-the-art results on token-level and sentence-level NLP tasks, including GLUE (Wang et al., 2018), MNLI (Williams et al., 2018), SQuAD (Rajpurkar et al., 2016), etc., surpassing many previous task-specific models.

With expectations of how self-attention controls the performance of BERT, researchers pursued customizing the structure of self-attention to further boost its potential. Customized attention blocks in Transformers have been developed, improving the model with either its performance or its training effectiveness (Cui et al., 2019; Guo et al., 2019; Li et al., 2019; Shi et al., 2021).

2.2. Eye-Tracking in Natural Language Processing

Eye movements are signals reflecting brain activities, and can be directly observed and obtained; it provides insights into the cognitive processing of language processing with high temporal resolution. In reading comprehensive studies, ET features related to fixation, saccade, gaze, and reading time are frequently adopted to explore human cognitive load. It has been demonstrated that eye movements can be sensitive to text features from lexical to discourse level in reading events, including word frequency (Inhoff and Rayner, 1986), syntactic ambiguity (Frazier and Rayner, 1987), text readability (Rayner et al., 2006), etc. Early measures collected at the initial stage of language processing are related to based properties recognition, and late measures reflected processing strategies countering with processing difficulty (Conklin and Pellicer-Sánchez, 2016). Detailedly, early-stage features such as first fixations duration (FFD) have been found to possess a correlation with mostly basic lexical (Henderson and Ferreira, 1990; Inhoff and Rayner, 1986; Rayner and Frazier, 1987) and possibly syntactic factors (Ferreira and Henderson, 1990; Rayner and Frazier, 1987), while late-stage features like total reading time (TRT) has been inspected as an indicator of word density, sophistication, and readability (Mishra et al., 2018).

To promote the utility of ET data in language processing research, corpora with ET features have been established. The material of the text carrier would not affect human reading behaviour (Skaramagkas et al., 2021), therefore, ET data collected based on any reliable media should be available for universal utility. Corpora with ET data such as DUNDEE (Kennedy et al., 2003), GECO (Cop et al., 2017), PROVO (Luke and Christianson, 2018), ZuCO (Hollenstein et al., 2018, 2020), etc. were constructed, based on which experiments have been conducted for a various range of purposes like language asymmetry (Demberg and

[1] https://github.com/SodaFont/EyetrackingAugmentedTransformers

Keller, 2008), second language acquisition (Godfroid, 2019; Winke et al., 2013), reading behaviour of local coherences (Bicknell et al., 2008), syntactic factors' influence on reading patterns level (Snell and Theeuwes, 2020), etc. These corpora are also applied in machine learning studies. On the one hand, data has been adopted in augmenting model performance in downstream tasks (Hollenstein and Zhang, 2019; Hollenstein et al., 2019a; Zhao et al., 2023; Meister et al., 2021; Bakarov, 2018; Hollenstein et al., 2019b), and 'surprising robustness' has been spotted (Goodkind and Bicknell, 2018); on the other hand, models combined with ET data to simulate human reading behaviour have been proposed (Malmaud et al., 2020; Sood et al., 2020; Reichle et al., 2003) to interpret LMs (Hollenstein et al., 2021; Eberle et al., 2022).

Due to the scale limitation of existing ET data, researchers explored machine-learning approaches to predict human reading patterns. Based on the standardized datasets (Hollenstein et al., 2018, 2020), a diversity of gaze-predicting models have been established (Hollenstein et al., 2021). The accurate modelling of ET features should be crucial to enhance the understanding of language processing.

3. Methodology

The number of corpora with ET features collected is far from sufficient in real-case studies, and it can hardly ensure that future data adopted for LM studies will be ready-prepared with desired ET features. Therefore, we design two parts of experiments: (1) whether the ET data collected from real-life reading events can augment model performance in QA tasks, (2) whether model-predicted ET data can augment model performance in the same type of task, so as to assess the generalization ability of the augmenting methods.

3.1. Language Model and Setup

We choose BERT-base-cased[2] and ALBERT-base-v2[3] in the experiments. For the BERT model, since each layer adopts independent parameters, the information passing through the neural network changes drastically as it moves towards the deeper layers, enabling it to infer higher-level linguistic information (Puccetti et al., 2021). ALBERT, sharing parameters between layers, has a much smoother information transition flow compared with BERT (Lan et al., 2019), thus the output of its attention block is more likely to hold the low-level linguistic information.

3.2. Experiment Tasks

The form of the QA task in this experiment is to extract the answer span from the context. The experiment is composed of two parts (Figure 1):

- **Task 1** Examining whether ET data collected from real-life reading events can enhance Transformer model performance in QA tasks. We combine DUNDEE (Kennedy et al., 2003), the English-reading part of GECO (Cop et al., 2017), PROVO (Luke and Christianson, 2018), and the task-free reading part of ZuCO (Hollenstein et al., 2018, 2020) to compose a larger ET dataset, and segment each context into a trial with less than 300 words. For each trial, QA pairs are then generated by the Questgen model [4], and conduct manual cleaning to remove trials with QA pairs semantically or logically making no sense, or questions not matching to an identical answer. To ensure every token of the input data is covered by an ET data point, ET data for question texts is predicted by the gaze-predicting model developed by team TorontoCL (Li and Rudzicz, 2021). The final dataset with 2051 total trials is split into training and test sets at the ratio of 4:1.

- **Task 2** Examining whether ET data predicted by the model can augment Transformer performance in QA tasks. We choose SQuAD v1.0 (Rajpurkar et al., 2016) for the second task, which contains abundant QA tasks covering a diversity of topics. For a rigorous horizontal comparison with task 1, SQuAD v2.0 with unanswerable questions is excluded. Additionally, there are abundant trials in SQuAD spotting with a mixture of different languages, some also contain non-Latin characters, challenging both the gaze-predicting model and the QA model to counter with uncleaned multilingual data. The ET features for tokens in both contexts and questions are predicted by the same gaze-predicting model in task 1.

3.3. Eye-Tracking Attention Mask

We choose TRT and FFD for ET attention mask structuring:

- TRT is frequently regarded as the index of total cognitive load (Frank and Hoeks, 2019) - longer reading time marks higher cognitive

[2] https://huggingface.co/google-bert/bert-base-cased

[3] https://huggingface.co/albert/albert-base-v2

[4] https://github.com/ramsrigouthamg/Questgen.ai

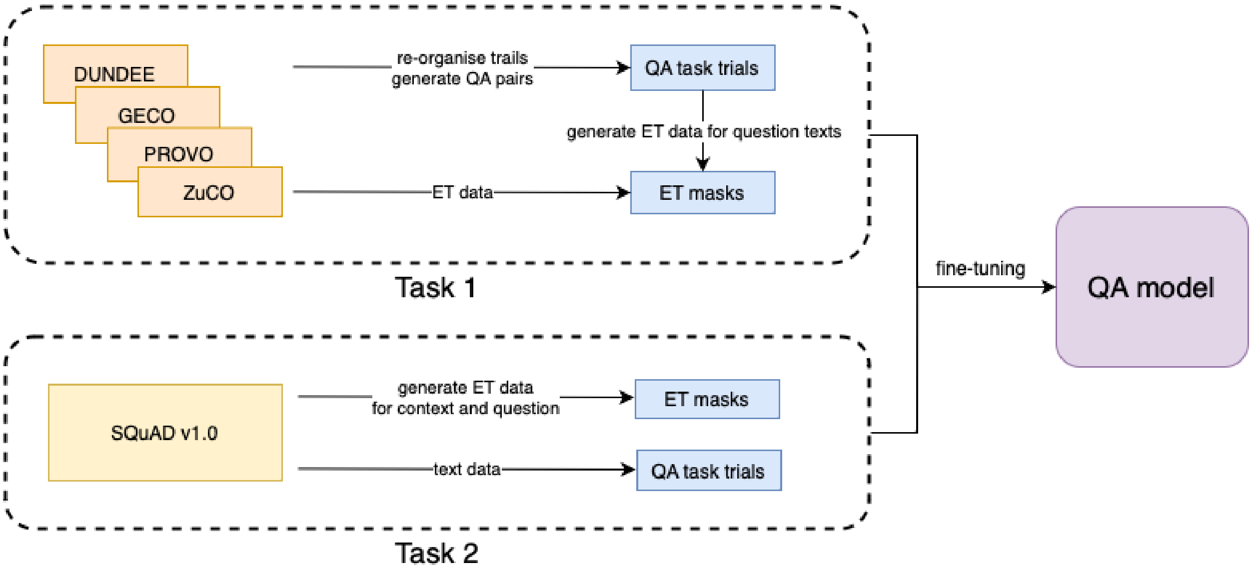

Figure 1: The structure of the two parts of the experiment. Task 1 (top) is based on corpora with real-life collected eye-tracking data, and task 2 (bottom) is based on SQuAD v1.0 benchmark with model-predicted gaze data.

load spent during language processing, therefore is positively correlated with text processing difficulty (Tanenhaus et al., 2000). It is considered closely related to the late stage of text processing, such as information reanalysis, discourse integration, etc. (Barrett, 2018). While Transformers' embedding includes mostly lexical-level information (Devlin et al., 2019), TRT introduces extra information from syntactic and semantic levels and may guide models on capturing corresponding information, thus assisting the analysis of sophisticated cases and enhancing model performance in difficult tasks such as ambiguous phrase parsing (Barrett, 2018).

- FFD conveys an enormous amount of information in language processing (Henderson, 1993). Collected at the early stage of the reading event, it is considered to provide the most accurate information on object identification processes (Henderson et al., 1987) based on low-level features like word frequency, word length, word position, etc. (Barrett, 2018), also positively correlated to word surprisal (Vainio et al., 2009). FFD may slightly carry information at syntactic (Barrett and Søgaard, 2015; Demberg and Keller, 2008) and even semantic level (Barrett, 2018) as well, for it is significantly influenced by the properties of the previous and upcoming words of the currently fixated word (Kliegl et al., 2006). The features reflected by FFD highly correspond to BERT's embedding, thus the attention mask structured based on it is expected to resonate with Transformer attention.

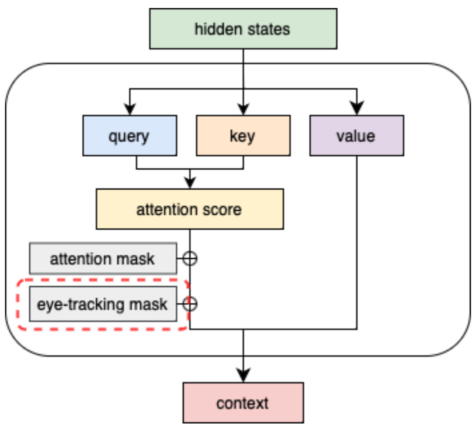

Figure 2: The eye-tracking attention mask is added to the attention score in the self-attention block.

Inspired by Shi's study (2021), we add the outsource attention mask to the attention score in the self-attention blocks. Figure 2 shows the mechanism of the modified self-attention module, where the ET attention mask applied is marked out. Three different ET attention masks are designed:

- **Standard mask** standardizes the original ET data sequence and keeps the ratio of differences between elements.
- **Weakened mask** is derived from the standard mask, with every element in the sequence minus 1, and goes through the exponential calculation. The difference ratio between elements is narrowed.

Mask scale	Accuracy	F1-Score
$10e0$	2.433	3.326
$10e-1$	3.650	4.965
$10e-2$	52.311	55.161
$10e-3$	77.625	79.087
$10e-4$	81.265*	82.870*
$10e-5$	74.915	76.300

Table 1: Pilot experiment results with different numerical scales of eye-tracking masks applied on ALBERT model.

- **Augmented mask** applies the softmax function on the standard mask to polarise the elements within the range from 0 to 1.

Table 1 shows the result of a pilot experiment for determining the proper scale of the ET attention mask to determine the scale of the attention masks.

3.4. Evaluation

We apply the following indicators to evaluate model performance:

- **Accuracy** is the percentage of exactly match answers.
- **F1-score** is a robust index calculated based on precision and recall rate.
- **Recall** aims to show how well a model performs in data retrieval to generate the matching answer, and can also be regarded as an indicator of sufficiency. It indicates the sensitivity of a model towards the rationales including answers, yet oversensitivity can result in the model capturing too much useless information, leading to a high rate of answer invalidity.
- **Comprehensiveness** checked whether the model selected rationale is sufficient to make a correct prediction. It is calculated as:

$$Comprehensiveness = \sum \frac{n_i}{N}$$

where n_i is whether the ground truth answer is comprehensively included in the model rationale (1 for true and 0 for false), and N is the total number of trials. Similarly, higher comprehensiveness scores do not equal better performance.

4. Results

Tables 2, 3, 4, and 5 present comparisons of the performance between the type of augmented models with the highest average accuracy and the corresponding vanilla model in each task, respectively. The result scores are the average of 5 runs in each group of experiments.

When introducing real-life ET data into model attention, the ALBERT model combined with the augmented mask structured based on FFD achieves the best result (Table 2). While the model has a better chance of achieving higher best scores, the standard deviation in accuracy exceeded the original. For sufficiency and comprehensiveness especially, the model is improved greatly in both its performance scores and corresponding stability. However, the introduction of ET attention masks brings much instability to the BERT model in every aspect. In contrast, the best performance of the model guided by the weakened TRT mask improves slightly (Table 3).

With either mask structured based on model-predicted ET data, the benefit is comparatively not satisfying (Table 4 and 5). Though the stability of both ALBERT and BERT increases, the improvement in each average or best performance score is relatively slight, or even become worse.

To assist result analysis, we visualize the attention of the fine-tuned models with the best accuracies to inspect the impact of different ET attention masks bringing to models' attention distribution.

5. Discussion

In a series of experiments introducing different ET features into Transformer models for QA tasks, we inspect significant improvement with certain combinations. Variables affecting the results are discussed in detail.

5.1. ALBERT vs. BERT

We can easily spot that the vanilla ALBERT model has already outperformed the vanilla BERT with a much shorter training time, indicating that the former has higher confidence in QA tasks based on reading comprehension. Introducing ET masks expands the gap, especially when adopting real-life collected data. ALBERT model is greatly improved by its sensitiveness towards valid rationales, and also performs more precisely, while the BERT model does not benefit much from the extra guidance. This may closely relate to the structure of models. As has been mentioned, the attention block of BERT is better at inferring high-level linguistic information, while the one of ALBERT is more likely to hold the low-level information input to the attention module initially, ergo additional eye-movement signals may interfere with BERT's reasoning performance but compensate for the linguistic information deficiencies on ALBERT's deep layers.

Figures 3 and 4 show the attention heat maps of ALBERT and BERT on certain heads. We can observe that the attention distribution is enhanced

2*Index	Non-masked model			Masked model		
	Average	**Std.**	**Best**	**Average**	**Std.**	**Best**
accuracy	81.606	1.248	83.212	82.920	1.599	84.915
f1-score	82.733	1.361	84.324	84.471	1.313	86.021
recall	80.408	3.830	85.629	84.860	0.972	85.972
comprehensiveness	79.684	4.035	85.158	84.380	0.979	85.401

Table 2: Comparison between non-masked and augmented real-life first fixation duration data masked ALBERT

2*Index	Non-masked model			Masked model		
	Average	**Std.**	**Best**	**Average**	**Std.**	**Best**
accuracy	77.859	0.988	78.589	78.735	2.083	80.779
f1-score	79.423	1.200	80.342	80.610	2.328	82.993
recall	79.594	1.106	80.848	81.639	2.717	84.637
comprehensiveness	78.929	1.123	80.292	81.071	2.640	84.185

Table 3: Comparison between non-masked and weakened real-life total reading time data masked BERT

2*Index	Non-masked model			Masked model		
	Average	**Std.**	**Best**	**Average**	**Std.**	**Best**
accuracy	82.117	1.991	83.500	82.479	1.865	83.614
f1-score	89.661	1.298	90.580	89.800	1.170	90.580
recall	82.081	0.818	83.214	82.236	0.873	83.081
comprehensiveness	72.677	0.876	74.144	73.118	0.740	73.851

Table 4: Comparison between non-masked and standard model-predicted first fixation duration data masked ALBERT

2*Index	Non-masked model			Masked model		
	Average	**Std.**	**Best**	**Average**	**Std.**	**Best**
accuracy	79.712	1.350	81.220	80.123	1.141	81.088
f1-score	87.615	0.777	88.527	87.912	0.623	88.456
recall	81.060	0.393	81.601	81.167	0.343	81.585
comprehensiveness	72.297	0.160	72.479	72.392	0.290	72.753

Table 5: Comparison between non-masked and augmented model-predicted total reading time data masked BERT

on deep layers. Specifically, TRT, as a late-stage feature reflecting high-level linguistic signals, such as semantic or even pragmatic information, assists in making up for the little growth of linguistic information in ALBERT's attention block. However, BERT can originally infer higher-level linguistic features through its network - interdependence between cross-layer units tended to grow, eventually contributing to the structuring of the global syntax tree (Puccetti et al., 2021) - and information introduced extra messages cause a disturbance (Figure 5).

We also notice that the weakened masks suit BERT better, while augmented masks suit ALBERT models better. Hence, when the augmentation of data strengthens its ability to enhance ALBERT performance, the out-source mask with decreased information fits BERT better for it interfering with the model attention less at the early stage. However, a positive influence appears indeed while applying ET data to BERT models. Therefore, the concern for BERT should be what the proper ET data intensity is to reach a balance where BERT can benefit most with the least distractions.

5.2. Real-Life Data vs. Model-Predicted Data

Evidently, real-life ET data shows a much stronger potential in boosting model performance in QA tasks compared with model-predicted data. This can be credited to a high alignment of the characters of both the Transformer's representatives and ET features.

Firstly, the astonishing and long-lasting success of Transformer models achieved in NLI tasks is closely related to the structure of its deep learning architecture built to present the text. Unlike the sequential representative in RNN models, Transformers provides a structural representation to re-

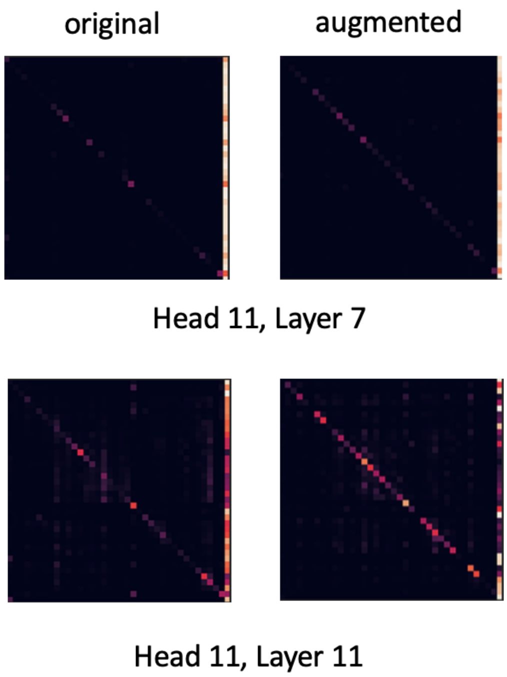

Figure 3: Comparison between the attention maps of non-masked ALBERT and the one masked by augmented real-life total reading time data on layer #7 and #11.

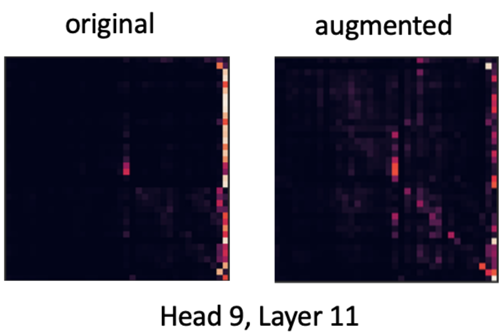

Figure 4: Comparison between the attention maps of non-masked BERT and the one masked by weakened real-life total reading time data on layer #11.

flect the syntax tree (Henderson, 2020), while the ET mask also presents a structural attention distribution instead of a sequential one. Since the tree structure Transformers built purely relies on its attention mechanism (Jawahar et al., 2019), it is reasonable that introducing ET attention signals can benefit its structuring process. Secondly, features like sentence and token length, as well as the relation link between tokens, are captured by the Transformers as the basic linguistic feature and help build the tree structure inside the model; meanwhile, these linguistic features strongly correlate

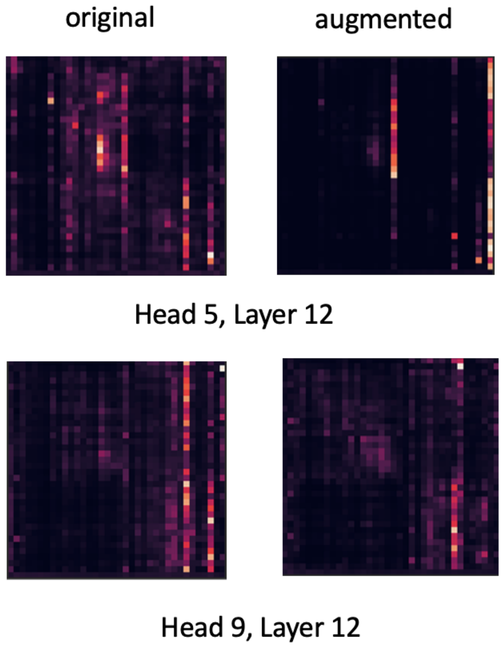

Figure 5: Comparison between the attention maps of non-masked BERT and the one masked by weakened real-life first fixation duration data on deep layer #12.

with ET features, which are also extra sensitive to the tokens with close relation but with long distance between (Sarti et al., 2021), so the ET data assists in determining the nodes of the parse tree among tokens. Apart from the alignment, introducing ET data also introduces supplementary information contributing to disambiguation (Duffy et al., 1988) at the early stage of models' reading comprehension.

However, when applying ET data generated by the gaze-predicting model, little improvement is found between the vanilla model and the augmented ones. In many cases, it even fails to outperform the original model. Indeed, multiple researches have proved that adopting fewer linguistic features as the variable for predicting ET features improves the accuracy of the predicting models (Bestgen, 2021), yet it can result in less linguistic signal involved in the predicted data. The predicting model adopted in this study only takes four lexical-level features as factors to generate predictions (Li and Rudzicz, 2021), and all higher-level linguistic features, such as positional and grammatical information, are completely left out. An extra strong focus on low-level information may cause models to ignore other linguistic information, leading to worse performance on extracting target rationales in QA tasks, especially for ALBERT. Additionally, the word frequency calculation in the gaze-predicting model involves an external library(Bestgen, 2021), while the word co-occurrence within the target text does matter in generating cognitive signals during reading (Eberle et al., 2022), further impacts the quality of the generated data. Yet the generated data has its advantage in stabilizing model performance, indicating that there may be abundant disturbance and noise involved in the real-life data.

5.3. Total Reading Time vs. First Fixation Duration

TRT and FFD are features collected from different stages of reading events, and they contain different levels of linguistic signal that affect model performance differently. While FFD's enhancing ability is stronger than TRT for the ALBERT model with all four indices, in more than half of cases, the BERT model combined with TRT data outperforms the one combined with FFD. The features that succeeded in enhancing model performance carry the complementary linguistic features to what the model is good at transferring cross-layer in its attention block. BERT is equipped with the inference ability to upgrade the level of linguistic features between layers, hence importing extra signals of basic-level linguistic features may force BERT to keep more low-level linguistic information. Oppositely, the low-level linguistic information passes smoothly in the ALBERT model's attention block, so introducing TRT mends the lack of high-level linguistic features in its output. This complementary can also be intuitively observed in the attention heat maps (Figure 3). Notably, when a model benefits in its performance from the introduction of complementary information, there is a compromise in its stability, and this may indicate that extra-linguistic information imported to the models' layers causes confusion in the fine-tuning process. Nevertheless, for many models obtaining higher average and best scores, the confusion is triggered probabilistically rather than inevitably.

6. Conclusion

In this work, we find that introducing eye-tracking data into the self-attention module of BERT and ALBERT contributes to the improvement of model performance in QA tasks in varying degrees. Compared with other cognitive signals, for instance, EEG and fMRI brain activity measures, ET features are relatively easily accessible with lower cost and expertise required in its collecting process, and extensive existing research in psycholinguistics brought forth standardized methods of preprocessing and feature extraction. These reasons make eye-tracking data a valuable source of human cognitive signals for language processing. The positive result of ET augmentation of Transformer models for question-answering tasks showed that data going through simple initial processing can benefit model performance. The mechanism of information transmission within the attention block and the linguistic information carried by the ET data both affect the effectiveness of augmentation, therefore it is important to select the appropriate features for model augmentation. Meanwhile, approaches to enhance the stability of model performance while keeping the benefits of applying ET attention masks remain to be explored. It is encouraged to design optimized eye-tracking augmentation methods based on mathematical and machine learning theories, as well as to apply different ET features on different attention heads or layers specifically for more delicate model enhancement.

The positive result achieved in this study is a heuristic step we take, but due to the limited resources of existing ET data, it is only a preliminary attempt. Structuring scaled data should play a significant role in method generalization. The establishment of the webcam eye-tracking method could further reduce the ET data collection cost; though with a compromise in its accuracy, we show that it is helpful in augmenting language model performance to some extent. Therefore, introducing webcam-captured data into a model's attention block can also be a worthy attempt for future research. From the current results, we found that introducing the ET data generated by the predicted model modestly benefits the performance of the Transformer models. Therefore, promoting the establishment of an effective ET-predicting model will also be a key step in advancing the augmenting language model performance. A better understanding of the relationship between language models and human attention should bring further advantages in both model interpretation and neurolinguistics.

Limitations

Firstly, the eye-tracking datasets established with human participants involved in this research provide anonymous records in compliance with ethical board approvals and contain no personal information of the participants.

The experiment in this paper is conducted fully depending on English datasets, therefore the generalization of the method with other languages requires further examination.

For data collected from human reading events, we aggregate the data to obtain an average performance of human reading behaviour on each trial. However, individual data may vary greatly across participants, for the reading experiment environmental conditions and reading strategy participants take can be different. Specific reading patterns may have an extra strong positive or negative impact on model performance.

For task 1 (see Section 3.2) specifically, the task dataset is relatively small and the QA pairs are generated by a model with limited quality compared to well-established benchmarks. Constructing a QA-specialized eye-tracking corpora may further improve the study.

7. Acknowledgements

We acknowledge the computing resources and technical support provided at the UCloud platform at SDU eScience Center. We thank the anonymous reviewers for their thoughtful comments on the paper.

8. Bibliographical References

Amir Bakarov. 2018. Can eye movement data be used as ground truth for word embeddings evaluation? *Linguistic and Neuro-Cognitive Resources (LiNCR)*, page 27.

Maria Barrett. 2018. *Improving natural language processing with human data: Eye tracking and other data sources reflecting cognitive text processing*. Ph.D. thesis, Department of Nordic Studies and Linguistics, Faculty of Humanities, University of Copenhagen).

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Maria Barrett and Anders Søgaard. 2015. Using reading behavior to predict grammatical functions. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning*, pages 1–5.

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.

Yves Bestgen. 2021. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. pages 90–96.

Klinton Bicknell, Vera Demberg, and Roger Levy. 2008. Local coherences in the wild: An eye-tracking corpus study. *Language*, 54:363–388.

Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. pages 2174–2184.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 729–738.

Kathy Conklin and Ana Pellicer-Sánchez. 2016. Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3):453–467.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49:602–615.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2019. Fine-tune BERT with sparse self-attention mechanism. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3548–3553.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Susan A Duffy, Robin K Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of memory and language*, 27(4):429–446.

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.

Fernanda Ferreira and John M Henderson. 1990. Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):555.

Stefan L Frank and John CJ Hoeks. 2019. The interaction between structure and meaning in sentence comprehension. recurrent neural networks and reading times.

Lyn Frazier and Keith Rayner. 1987. Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of memory and language*, 26(5):505–526.

Aline Godfroid. 2019. Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. pages 1315–1325.

James Henderson. 2020. The unstoppable rise of computational linguistics in deep learning. pages 6294–6306.

John M Henderson. 1993. Eye movement control during visual object processing: effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(1):79.

John M Henderson and Fernanda Ferreira. 1990. Effects of foveal processing difficulty on the perceptual span in reading: implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3):417.

John M Henderson, Alexander Pollatsek, and Keith Rayner. 1987. Effects of foveal priming and extrafoveal preview on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):449.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019a. Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019b. CogniVal: A framework for cognitive word embedding evaluation. pages 538–549.

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(1):1–13.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. pages 138–146.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. pages 1–10.

Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & psychophysics*, 40(6):431–439.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The DUNDEE corpus. In *the 12th European conference on eye movement*, Dundee, Scotland.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*, 135(1):12.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. pages 785–794.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. pages 85–89.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. pages 963–980.

Abhijit Mishra, Pushpak Bhattacharyya, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Scanpath complexity: modeling reading/annotation effort using gaze information. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking*, pages 77–98.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How do BERT embeddings organize linguistic knowledge? In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 48–57.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. pages 2383–2392.

Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3):241–255.

Keith Rayner and Lyn Frazier. 1987. Parsing temporarily ambiguous complements. *The Quarterly Journal of Experimental Psychology*, 39(4):657–673.

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60.

Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and James Tin-Yau Kwok. 2021. SparseBERT: Rethinking the importance analysis in self-attention. In *International Conference on Machine Learning*, pages 9547–9557. PMLR.

Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos S Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, 16:260–277.

Joshua Snell and Jan Theeuwes. 2020. A story about statistical learning in a story: Regularities impact eye movements during book reading. *Journal of Memory and Language*, 113:104127.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Michael K Tanenhaus, James S Magnuson, Delphine Dahan, and Craig Chambers. 2000. Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of psycholinguistic research*, 29:557–580.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Sigir 1st international workshop on conversational approaches to information retrieval (cair'17)*, volume 5.

Seppo Vainio, Jukka Hyönä, and Anneli Pajunen. 2009. Lexical predictability exerts robust effects on fixation duration, but not on initial landing position during reading. *Experimental psychology*, 56(1):66–74.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. pages 353–355.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122.

Paula M Winke, Aline Godfroid, and Susan M Gass. 2013. Introduction to the special issue: Eye-movement recordings in second language research. *Studies in Second Language Acquisition*, 35(2):205–212.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. 2019. Are transformers universal approximators of sequence-to-sequence functions?

Lei Zhao, Yingyi Zhang, and Chengzhi Zhang. 2023. Does attention mechanism possess the feature of human reading? A perspective of sentiment classification task. *Aslib Journal of Information Management*, 75(1):20–43.

Appendix A. Full Experiment Results

ET mask type	ALBERT			BERT		
	mean	std.	best	mean	std.	best
no extra mask	81.606	1.248	83.212	77.859	0.988*	78.589
standard TRT mask	79.501	0.540*	80.049	77.324	1.864	79.562
weakened TRT mask	80.756	2.148	83.341	78.735*	2.083	80.779*
augmented TRT mask	82.482	1.804	84.672	78.248	1.673	80.292
standard FFD mask	82.774	1.159	83.942	76.691	1.080	77.859
weakened FFD mask	80.535	1.939	83.455	78.248	1.174	79.805
augmented FFD mask	82.920*	1.599	84.915*	77.178	1.672	79.805

Table 6: Accuracy of models fine-tuned on eye-tracking corpora guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	std.	best	mean	std.	best
no extra mask	82.733	1.361	84.324	79.423	1.200*	80.342
standard TRT mask	80.997	0.480*	81.507	79.482	2.151	81.876
weakened TRT mask	81.494	1.615	83.813	80.610*	2.328	82.993
augmented TRT mask	83.783	1.584	85.718	80.051	1.793	82.457
standard FFD mask	84.023	1.154	85.330	78.662	1.339	80.443
weakened FFD mask	81.679	1.753	84.312	80.342	1.741	83.009*
augmented FFD mask	84.471*	1.313	86.021*	79.051	1.612	81.577

Table 7: F1-scores of models fine-tuned on eye-tracking corpora guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	std.	best	mean	std.	best
no extra mask	80.408	3.830	85.629	79.594	1.106*	80.848
standard TRT mask	80.163	2.150	81.766	80.228	2.371	82.653
weakened TRT mask	82.144	2.044	84.374	81.639*	2.717	84.637
augmented TRT mask	83.935	1.710	85.792	80.991	1.930	83.905
standard FFD mask	83.983	1.056	84.749	79.617	1.675	81.909
weakened FFD mask	81.960	1.381	83.358	81.269	2.264	84.813*
augmented FFD mask	84.860*	0.972*	85.972*	80.332	2.063	83.642

Table 8: Recall of models fine-tuned on eye-tracking corpora guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	std.	best	mean	std.	best
no extra mask	79.684	4.035	85.158	78.929	1.123*	80.292
standard TRT mask	78.881	3.065	80.535	79.270	2.298	81.509
weakened TRT mask	81.703	2.165	84.185	81.071*	2.640	84.185*
augmented TRT mask	83.260	1.766	85.158	80.487	1.749	83.212
standard FFD mask	83.650	1.120	84.672	78.929	1.700	81.022
weakened FFD mask	81.265	1.419	82.725	80.535	2.156	83.942
augmented FFD mask	84.380*	0.979*	85.401*	79.757	2.152	83.212

Table 9: Comprehensiveness of models fine-tuned on eye-tracking corpora guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	**std.**	**best**	**mean**	**std.**	**best**
no extra mask	82.117	1.991	83.500	79.712	1.350	81.220*
standard TRT mask	82.420	0.867*	83.453	78.831	1.189	80.624
weakened TRT mask	81.251	2.375	83.699	79.707	1.245	80.634
augmented TRT mask	81.198	1.523	83.349	80.123*	1.141	81.088
standard FFD mask	82.479*	1.865	83.614	79.692	1.096	80.482
weakened FFD mask	80.789	1.824	83.396	79.633	1.060*	80.776
augmented FFD mask	81.985	1.826	83.746*	79.092	1.939	80.785

Table 10: Accuracy of models fine-tuned on SQuAD v1.0 guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	**std.**	**best**	**mean**	**std.**	**best**
no extra mask	89.661	1.298	90.580	87.615	0.777	88.527*
standard TRT mask	89.996*	0.697*	90.793*	87.111	0.650	88.137
weakened TRT mask	89.139	1.461	90.629	87.706	0.754	88.360
augmented TRT mask	89.112	0.969	90.560	87.912*	0.623*	88.456
standard FFD mask	89.800	1.170	90.580	87.639	0.772	88.225
weakened FFD mask	88.877	1.006	90.296	87.649	0.657	88.346
augmented FFD mask	89.607	1.072	90.782	87.274	1.250	88.337

Table 11: F1-scores of models fine-tuned on SQuAD v1.0 guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	**std.**	**best**	**mean**	**std.**	**best**
no extra mask	82.081	0.818	83.214	81.060	0.393	81.601
standard TRT mask	82.786*	0.731	83.546	81.040	0.440	81.499
weakened TRT mask	82.314	1.175	83.516	81.097	0.565	81.627
augmented TRT mask	82.287	0.843	83.554*	81.167*	0.343*	81.585
standard FFD mask	82.236	0.873	83.081	80.945	0.740	81.839*
weakened FFD mask	82.034	0.650*	83.045	80.944	0.430	81.346
augmented FFD mask	82.422	0.777	83.272	80.941	0.392	81.472

Table 12: Recall of models fine-tuned on SQuAD v1.0 guided by different eye-tracking attention masks

ET mask type	ALBERT			BERT		
	mean	**std.**	**best**	**mean**	**std.**	**best**
no extra mask	72.677	0.876	74.144	72.297*	0.160*	72.479
standard TRT mask	73.574*	0.915	74.484	72.163	0.445	72.658
weakened TRT mask	73.262	1.244	74.570*	72.191	0.509	72.611
augmented TRT mask	73.342	0.812	74.428	72.392	0.290	72.753
standard FFD mask	73.118	0.740	73.851	71.885	0.744	72.904*
weakened FFD mask	73.075	0.541*	73.983	72.083	0.424	72.526
augmented FFD mask	73.381	0.600	74.049	72.108	0.300	72.507

Table 13: Comprehensiveness of models fine-tuned on SQuAD v1.0 guided by different eye-tracking attention masks