Transformer Attention vs Human Attention in Anaphora Resolution

Anastasia Kozlova[1], Albina Akhmetgareeva[1], Aigul Khanova[2], Semen Kudriavtsev[2], Alena Fenogenova[1]
[1]SaluteDevices, [2]HSE University
Correspondence: anastasi2510@gmail.com

Abstract

Motivated by human cognitive processes, attention mechanism within transformer architecture has been developed to assist neural networks in allocating focus to specific aspects within input data. Despite claims regarding the interpretability achieved by attention mechanisms, the extent of correlation and similarity between machine and human attention remains a subject requiring further investigation. In this paper, we conduct a quantitative analysis of human attention compared to neural attention mechanisms in the context of the anaphora resolution task. We collect an eye-tracking dataset based on the Winograd schema challenge task for the Russian language. Leveraging this dataset, we conduct an extensive analysis of the correlations between human and machine attention maps across various transformer architectures, network layers of pre-trained and fine-tuned models. Our aim is to investigate whether insights from human attention mechanisms can be used to enhance the performance of neural networks in tasks such as anaphora resolution. The results reveal distinctions in anaphora resolution processing, offering promising prospects for improving the performance of neural networks and understanding the cognitive nuances of human perception.

1 Introduction

The term *attention* describes both human cognitive processes, crucial for tasks like reading and comprehension, and the attention mechanism in neural networks (Bahdanau et al., 2016), which dynamically adjusts focus to specific input data. Despite their apparent differences, this paper aims to analyze the correlations between transformer attention and human attention during anaphora resolution task.

Successful language comprehension requires understanding the discursive connections in the sentences and the logical relationships between discourse structures in the text. Coreference resolution, a standard NLP task, determines which mentions in a text refer to the same entity. Two mentions (i.e., textual phrases) are called coreferent if they refer to the same real-world objects or events. Anaphora, one of the types of coreference resolution, highlights this challenge by requiring the matching of an anaphor (typically a pronoun) in a sentence with its antecedent (noun) in the preceding sentence. The Winograd Schema (Levesque et al., 2012) is a well-established method for evaluating language model performance in anaphora resolution tasks, assessing the model's logical reasoning and real-world knowledge in resolving coreference ambiguities. It is an evaluation dataset within the SuperGLUE (Wang et al., 2019) suite across various languages.

Video oculography, known as eye-tracking, is a prevalent psycholinguistic method for studying reading processes. It involves recording the reader's eye movements via video and subsequent interpolation of their gaze onto a display screen. This method breaks down the reading process into fixations (periods of steady gaze) and saccades (rapid eye movements) between them with precision up to milliseconds. This approach enables a detailed examination of reading acquisition. We used eye-tracking techniques to gather information on human fixations and focuses during the anaphora resolution and create the eye-tracking Winograd schema dataset.

Leveraging the dataset, we investigate the correlation between machine and human attention across various transformer architectures and network layers. The research aims to confirm whether integrating insights from human attention patterns can significantly improve the language model's ability to resolve anaphoras effectively.

The contributions of the current study are the following:

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 109–122
August 15, 2024 ©2024 Association for Computational Linguistics

- we collect and propose the new dataset [1] based on the data from human eye-tracking for anaphora resolution;
- we conduct a set of experiments on different models fine-tuned on the data to explore the attention mechanisms;
- we provide a detailed comparative analysis of human and neural attention mechanisms;
- we integrate the human gaze into the transformer's attention mechanisms.

2 Related Work

In the subsequent sections, we outline related works encompassing attention mechanisms in transformers, human attention datasets of eye-tracking data, methods of correlation analysis between human and machine attention, and the incorporation of eye-gaze data into models during training.

2.1 Attention Mechanisms in Transformers

The machine attention determines the degree of attention allocated to other segments of the input sentence during the encoding process of a word at a particular position. The attention mechanism in transformers is initially described in Vaswani et al. (2017) as a process of mapping input vectors – a query and a set of key-value pairs, to yield an output. The attention function for each word of the input sentence against a single word is computed as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

The input consists of the query and the key vectors, each with a dimension of d_k, and the values vectors of dimension d_v. The output is computed as a weighted sum of the values, where the weights (attention score) are calculated as a softmax of dot products of the query with the corresponding keys, scaled by $\frac{1}{\sqrt{d_k}}$. The attention function is computed over a set of input vectors, enabling their aggregation into a matrix structure for queries Q, keys K, and values V.

In performing multi-head attention, the singular attention function is computed h times (a number of attention layers, or heads) in parallel with different linear projections of the queries, keys, and values.

$$\mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

where the projections are parameter matrices $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$. Subsequently, the concatenated outputs, each possessing a dimension of d_v, undergo further projection with parameter matrices $W^O \in R^{hd_v \times d_{model}}$.

This mechanism enables the model to jointly attend to information across various representation subspaces at different positions. The transformer uses multi-head attention in three ways based on its architectural design. In the first configuration, denoted as the "encoder-decoder" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. In the second configuration, referred to as the encoder self-attention layer, all of the keys, values, and queries come from the output of the previous layer in the encoder. Analogously, self-attention layers in the decoder enable each position to attend to all positions in the decoder, encompassing those up to and including the respective position.

2.2 Human Attention

Eye-tracking datasets have emerged as invaluable resources for investigating various aspects of human cognition and behavior. These datasets provide researchers with fine-grained information about eye movements. The PROVO corpus (Luke and Christianson, 2017) includes eye-tracking data of passages taken from online news articles, magazines, and works of fiction. This dataset offers detailed information on participants' eye movements, fixations, and regressions, allowing researchers to explore phenomena such as syntactic ambiguity resolution and semantic processing during reading. Another widely utilized monolingual dataset is the ZuCo corpus (Hollenstein et al., 2018), which contains eye-tracking data of full sentences from movie reviews and Wikipedia articles in English. It includes features like a total number of gaze fixations and different fixation duration data collected from native English speakers during the execution of reading tasks. As for the Russian monolingual dataset, the Russian Sentence Corpus (Laurinavichyute et al., 2019) introduces a corpus of eye movements of silent reading by skilled Russian readers.

In addition to these established datasets, recent efforts have focused on collecting eye-tracking data from diverse populations and linguistic backgrounds to facilitate cross-cultural and multilingual

[1] https://huggingface.co/datasets/RussianNLP/EyeWino

research, for example, the corpus GECO (Cop et al., 2016). In particular, it includes five word-level reading time measures from English and Dutch monolinguals reading an entire novel. Furthermore, the MECO corpora (Siegelman et al., 2022; Kuperman et al., 2022) provides comparable cross-linguistic eye-tracking data and includes 13 different languages. Furthermore, numerous studies have utilized eye-tracking to investigate anaphora resolution across various languages and populations during reading (Wolna et al., 2024; Naido and Jaafar, 2022; Costa et al., 2011; Duffy and Rayner, 1990).

Additionally, datasets are utilized to enhance model performance by incorporating eye-gaze information to solve NLP tasks. For example, the eye-tracking dataset MQA-RC (Sood et al., 2020a), in which participants read movie plots taken from the MovieQA (Tapaswi et al., 2015) and answered predefined questions. In addition, the eye-gaze dataset from Mishra et al. (2016), where eye-movement parameters enhance the quality of models to solve a sarcasm detection task.

2.3 Eye-Tracking and Transformers

Recent research has focused on the correlation between attention mechanisms in transformer models and human eye-gaze patterns. The notable stream of the studies is to investigate the correlation between eye-gaze features and attention layers during reading tasks (Bensemann et al., 2022; Morger et al., 2022; Toneva and Wehbe, 2019). The results show a high correlation, primarily in the first attention layer. The paper (Sood et al., 2020b) evaluated correlations on the reading comprehension task for fine-tuned XLNet. They compared attention from the last encoder layer with eye-gaze features and reported a non-significant correlation. Moreover, the studies conduct experiments to explore whether task-specific fine-tuning influences the correlation with human reading attention (Eberle et al., 2022).

Another notable stream of research is a cross-lingual comparison of correlations. For example, due to results from Brandl and Hollenstein (2022), the correlation analysis across languages shows that considerable differences between languages, individual reading behavior, and vocabulary knowledge (LexTALE) influence the alignment between humans and models. In addition, the papers (Sen et al., 2020; Morger et al., 2022) provide methods to analyze word importance correlations between machines and humans. The paper (Morger et al., 2022) compares human and model relative word importance to investigate whether models focus on the same words as humans cross-lingually.

Furthermore, a promising area of research has explored the integration of eye-gaze data into the models to enhance task performance (i.e., sarcasm detection, question answering) and to deepen the understanding of language processing and human cognition (Sood et al., 2020b; Mishra et al., 2016; Zhang and Hollenstein, 2024).

3 Eye-tracking Data

Objective The anaphora resolution task was chosen to investigate the distinction between attention mechanisms in neural networks and humans. This study explores the potential benefits of integrating human-inspired attention mechanisms into transformer architectures. The research question seeks to confirm whether the language model's incorporation of information regarding human attention distribution during text reading improves its performance in the anaphora resolution task.

Experimental Setup For the experiment, we employed eye-tracking via video oculography, utilizing the EyeLink 1000 Plus device. Participants' gaze was calibrated until validation error values reached less than 1 (maximum) and 0.5 (average). The indication 0.5 is the maximum average deviation. When calibrating at 9 points, the error of each point is calculated. If the average is not more than 0.5 and the total error is not more than 1, then the calibration is considered successful, and incentives are presented to the participants. The EyeLink 1000 Plus device is one of the most accurate systems, with a validation error of about 0.2-0.5, within the standard protocol according to the system manual (Holmqvist et al., 2011).

The Russian Winograd Schema Challenge dataset from TAPE (Taktasheva et al., 2022) was utilized for the anaphora resolution task to gather information on participants' eye movements. The experiment comprised 150 complex or compound-complex sentences extracted from the Winograd schema challenge dataset, each containing an anaphoric pronoun and its antecedent.

Each participant was shown a sentence with an anaphoric pronoun highlighted in red on the screen, followed by a question about the presumed antecedent of the anaphora. The question was the following: "Does the highlighted pronoun refer to <antecedent> ?". An example of the participants' screen is presented in App. A. For each sentence, two presumed antecedents (one correct, one incorrect) were identified for each sentence. Thus, each screen was read by fifty participants. The sentences were randomized for each participant to ensure balanced conditions.

One hundred people (81 women, average age – 22.68, standard deviation – 4.27) who are native speakers of Russian participated in the experiment. They were instructed to read the provided sentences carefully and answer the question using the keyboard (key 1 for agreement, key 0 for disagreement). Participants completed three training sentences to ensure task comprehension before proceeding to three blocks of 50 sentences each, with breaks provided between blocks. To enhance recording quality, each trial began with a calibration check, requiring participants to focus precisely on the point where the first word of the text would appear. Upon successful calibration, the text was displayed; otherwise, recalibration commenced. After responding to each question, participants automatically advanced to the next trial. The experiment duration averaged 45 minutes.

Dataset statistics Observations with missing values and parsing errors were excluded from the dataset. The final dataset consists of 296 sentence-question pairs, which contain 9319 words and 148 unique sentences. The average number of participants per word is 48. The total number of observations for each variable is 448047. The resulting fields of the dataset are presented in the App. B.

4 Comparative analysis of attention mechanisms

In order to investigate the potential advantages of incorporating attention mechanisms similar to human processes into transformer architecture, we first need to examine and compare different attention mechanisms. We carried out a set of experiments on various architectures, fine-tuned using the data, and compared them with data on human attention. Our aim was to provide a detailed comparative analysis of human and neural attention mechanisms on the Winograd schema challenge.

4.1 Human Attention

We use the three word-level gaze measures extracted from the eye-tracking dataset (see Sec. 3) to quantify human attention:

- **Total reading time, TRT**, the sum of all fixation durations on the current word, ms;
- **Gaze duration, GD**, the sum of all fixation durations on the current word in the first-pass reading, ms;
- **Fixations, F**, the number of all fixations on the current word.

We use TRT because it highly correlates with model attention in similar works (Eberle et al., 2022; Bensemann et al., 2022; Morger et al., 2022). GD and F reflect which words attracted the most attention. We use these measures to determine the relative importance of words in a sentence. Each word is assigned a value between 0 and 1, which is normalized for each participant. The sum of the values of all words in a sentence is 1. These values are averaged across all participants to obtain the human relative importance of the word in a sentence (w_i):

$$w_i = \frac{1}{N}\sum_{j=1}^{N}\frac{m_{ij}}{\sum_{i=1}^{T} m_{ij}} \tag{3}$$

where m_{ij} is the gaze observation of the j-th participant for the i-th word, N – a number of participants, T – a number of words in a sentence.

For each example, we aggregate participants' responses using majority voting. The percentage of correct answers is 97.97%.

4.2 Transformer Attention

We use attention scores from the encoder layers of pre-trained and fine-tuned models across various transformer architectures to describe model attention.

4.2.1 Models

Multilingual models represent multiple languages within a shared space, aiming for a more universal understanding of language. The Russian language is well-represented in the pre-training corpus of various multilingual language models. To evaluate the impact of multilingual data in the training set on the model's attention distribution, we compare the performance of monolingual and multilingual models that have the same architecture and similar size. We use six publicly available language models from 3 model families:

BERT-based models include ruBERT-base (Zmitrovich et al., 2023) and mBERT-base (Devlin et al., 2019)

RoBERTa-based models include ruRoberta-large (Zmitrovich et al., 2023) and XLM-R-large (Conneau et al., 2020).

T5-based models include ruT5-base (Zmitrovich et al., 2023) and mT5-base (Xue et al., 2021).

Refer to Tab. 1 for the statistical details.

4.2.2 Datasets

Fine-tuning datasets The fine-tuning data represents a collection of Winograd schemas from various data sources.

For the Russian-language models, we used data from the RWSD task from the MERA benchmark (Fenogenova et al., 2024) and the Winograd task from the TAPE benchmark (Taktasheva et al., 2022). From the TAPE dataset, we exclude duplicates that were included in the eye-tracking dataset.

For the multilingual models, we combined Russian-language data and the XWINO dataset (Tikhonov and Ryabinin, 2021) without Russian to avoid duplication. Japanese and Chinese languages were excluded due to the special preprocessing required for this task.

For the comparative experiments of models on the anaphora task, we use the eye-tracking dataset for evaluation and the RWSD test set from the MERA benchmark.

Preprocessing Since we conducted an evaluation process for both the model and humans under the same conditions, all datasets were preprocessed to replicate the human experiment. For each sentence, an antecedent and an anaphoric pronoun were identified. The corresponding pronoun was highlighted in the text using uppercase. We formulated the question about the presumed antecedent of the anaphora using the human experiment design described in Sec. 3 and the answer for this question ("Yes" or "No"). The question and the answer were formulated in the language of the **text**. Each example also contains information about whether the question is about the correct or incorrect antecedent, with labels equal to 1 and 0, respectively.

- **text:** *"Bob collapsed on the sidewalk. Soon he saw Carl coming to help. HE was very concerned."*
- **question:** *"Does the highlighted pronoun refer to Carl ?"*
- **antecedent:** *"Carl"*
- **reference:** *"He"*
- **answer:** *"Yes"*
- **label:** *1*

The datasets were filtered so that the reference attribute was a pronoun and contained no more than one word. For example, "there" and "he does/did" were excluded from the dataset.

Finally, the training dataset was balanced with respect to the labels and filtered from duplicates. Tab. 2 provides the number of examples by language in the final datasets.

4.2.3 Fine-tuning

We fine-tune pre-trained models using train sets presented in Sec. 4.2.2. The original case of the input text is preserved during tokenization.

The encoder-only models are fine-tuned using a sequence classification head on top. We add a [SEP] token between the text and the question to get the input text for the models during the training process.

The encoder-decoder models are fine-tuned using a language modeling head on top. The text was concatenated with the question about antecedent to get the input text for the models.

Implementation The models are fine-tuned using AdamW optimizer (Loshchilov and Hutter, 2017) and a linear learning rate scheduler.

For the encoder-only models, we use a context window of 256, learning rate of $1e^{-5}$, batch size of 8, and 12 epochs.

For the encoder-decoder-based models, we use a context window of 200 and a batch size of 8. We also use a learning rate of $1e^{-5}$ and 35 epochs for ruT5-base, a learning rate of $1e^{-4}$ and 25 epochs for mT5-base. We use the generation hyperparameters: $max\_length = 20$, $temperature = 1$, $top\_k = 50$, $top\_p = 1$.

Metrics Models' performance is evaluated using the Accuracy score. Accuracy measures the percentage of correct predictions. This metric was chosen due to the balance of classes.

Results We take the checkpoints with the best performance on the validation set to evaluate them on the eye data test, and RWSD test set from Sec. 4.2.2. The results are presented in Tab. 3. The models demonstrate higher accuracy after fine-tuning, especially ruBERT-base, ruRoberta-large, ruT5-base and mT5-base. The encoder-decoder model mT5-base appears to outperform other models in solving the question-answering task.

Model	Architecture	Language	Parameters	Layers	Heads	Hugging Face Hub
ruBERT-base	Encoder-only	Russian	178M	12	12	ai-forever/ruBert-base
mBERT-base	Encoder-only	Multi	178M	12	12	google-bert/bert-base-multilingual-cased
ruRoberta-large	Encoder-only	Russian	355M	24	16	ai-forever/ruRoberta-large
XLM-R-large	Encoder-only	Multi	560M	24	16	FacebookAI/xlm-roberta-large
ruT5-base	Encoder-decoder	Russian	222M	12	12	ai-forever/ruT5-base
mT5-base	Encoder-decoder	Multi	580M	12	12	google/mt5-base

Table 1: Summary of the model architecture configurations.

Language	Train	Val	Test
English	2846	1216	-
French	108	56	-
Portuguese	358	116	-
Russian	872	326	260*
Total	4184	1714	260

Table 2: The sets statistics. The sizes of the set in the number of examples. * – the RWSD test set.

Model	Checkpoint	Eye data	RWSD
ruBERT-base	pre-trained	49.7	51.2
	fine-tuned	63.2	**58.5**
mBERT-base	pre-trained	49.0	52.3
	fine-tuned	50.0	50.0
ruRoberta-large	pre-trained	50.3	49.6
	fine-tuned	61.5	51.5
XLM-R-large	pre-trained	50.0	50.0
	fine-tuned	50.0	50.0
ruT5-base	pre-trained	50.0	50.0
	fine-tuned	57.4	55.8
mT5-base	pre-trained	50.0	50.0
	fine-tuned	**71.3**	56.2

Table 3: The models' performance (Accuracy) on the Winograd schema challenge task for the Russian language. The best score is in bold, and the second score is underlined.

4.2.4 Word-level attention

We use attention weights from the encoder layers to obtain the importance of a word in a sentence for the model. The decoder attention layers are only allowed to process earlier positions in the sequence, so we exclude them from the analysis.

We convert the texts into the format presented in Sec. 4.2.3 and tokenize them, preserving the original case of the words. The tokenized data is fed into the model. We extract the attention weights for each layer and average them across all attention heads.

$$\begin{aligned} \mathrm{A}' &= \mathrm{Average}(\mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{h}) \\ \mathrm{A}_\mathrm{i} &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}}) \end{aligned} \quad (4)$$

where A_i is an attention score of the i-th head with a dimension of $n \times n$, n – a length of the input sequence, h – a number of attention heads.

Each row $a^{(t)}$ of matrix $A^{'}$ is an attention vector for token t. We use the following matrix aggregations to obtain a vector of token importance:

- *mean* – the average of all the rows in each column.
- *row* – the average of pronoun tokens in each column: we extract only the rows corresponding to pronoun tokens from matrix $A^{'}$ and average these rows.

The special tokens are used for the attention calculations but are excluded from the final vector. During tokenization, some words are encoded as multiple tokens. The weights of the tokens that make up each word are summed to obtain word-level attention weights. The final vector is interpreted as a relative word importance for the model.

4.3 Correlation Analysis

We matched human and model attention scores so that each word had a normalized attention score from both sources.

Since we assume a monotonic relationship between variables but do not assume that the variables are normally distributed, we calculate the Spearman's rank correlation coefficient ρ (Hollander et al., 2013) to analyze the correlation between human attention and model attention. The correlation coefficient quantifies the strength and direction of the relationship between two variables. It ranges from -1 to $+1$, where 0 indicates no correlation.

The p-value is used to determine the statistical significance of the correlation coefficient. It indicates the probability of observing the calculated correlation coefficient under the assumption that the variables are actually uncorrelated in the population. If the p-value is less than the significance level, the hypothesis of no correlation is rejected. This suggests that there is a statistically significant correlation between the variables. We use a significance level of 0.05.

5 Results

Correlations of human attention with model attention are reported in Tab. 4. We found significant correlations ($p > 0.05$) for all experiments. There are moderate correlations for T5-based and RoBERTa-based models and strong correlations for BERT-based models on the first layer.

The comparison of different aggregation setups for T5-based architectures underscores the prevalence of the *mean* aggregation with high correlations. Conversely, for other architectures, we noted a contrasting trend where *row* aggregation predominates.

The first layers have high correlations in comparison to the last layer in most setups. For example, there are extremely small correlations for ruRoberta-large on the last layer; meanwhile, the maximum correlation is almost even with other values. For several models, the highest correlation is noted on a particular layer. Moreover, the layer demonstrating the highest correlation varies notably across different architectures.

For most of the models, there is no difference between pre-trained and fine-tuned versions, except for a slight correlation decrease for multilingual mT5-base after tuning on the Winograd schemas. Furthermore, we can compare the outcomes across different eye-gaze metrics and observe minimal discrepancies among them in terms of correlation analyses. Finally, we highlight the model mBERT-base, which demonstrates the highest correlation with human attention. We conclude that task-specific fine-tuning did not enhance the correlation between human attention and machine attention.

The analysis suggests that encoder-only models provide more significant insights for evaluating attention correlation. A detailed visualization of the correlation between human attention and models' attention on different layers is presented in App. C. Additionally, App. D provides a visualization of the important words for the human and the models for one example from the eye-tracking dataset.

6 Integrating Human Gaze into Transformers

Based on the results obtained in Sec. 5, there are significant correlations between human attention and the models' attention during the task of anaphora resolution. It can be assumed that using eye movement data when training models for this task will increase their performance. We conducted experiments to integrate eye movement data into the model training process by using an additional term in the loss function to bring the model's attention closer to human attention.

6.1 Experimental setup

Data for human gaze integration For the experiments with human gaze integration during model training, we use the eye-tracking dataset as a training set. We use the Russian language sets from Sec. 4.2.2 as validation and test sets.

Method We use the procedure proposed by Bensemann et al. (2022) to investigate the effect of injecting human eye-gaze bias during training as the baseline. We introduce an additional loss function to align the distribution of model attention on a given layer with the distribution of human attention. The final loss function is calculated according to the following formula:

$$L = H(y, \hat{y}) + \alpha H(p, \hat{p}) \tag{5}$$

Where $H(y, \hat{y})$ is the cross-entropy loss that measures the model's performance on the anaphora resolution task. $H(p, \hat{p})$ is the cross-entropy loss that measures the difference between two probability distributions: the distribution of the model's attention values on a particular layer (p) and the distribution of the human relative word importance ($\hat{p}$). The hyperparameter α controls the weight of the second term in the loss function. We use the hyperparameter α of 0.05. The remaining hyperparameters for fine-tuning models are contained in Sec. 4.2.3.

We use the average of all the rows in each column (*mean*) and the average of all pronoun tokens in each column (*row*) to obtain the models' attention values. We conduct experiments with different layers: the first, the last, and the layers where the highest correlation values between model attention and human attention are observed.

6.2 Results

The findings on incorporating human gaze data into models are presented in Tab. 5. Based on the results, we can conclude that using an additional loss does not usually improve the model's performance. However, a significant increase in Accuracy is observed for the tuned mT5-base and tuned ruRoberta-large models with row aggregation when using human attention on layers 1 and 14, respectively. It can be concluded that, in most

Model	**Agg.**	**Checkpoint**	**Layer**	**Fixations**			**Gaze duration**			**Total reading time**		
			(max)	first	max	last	first	max	last	first	max	last
ruBERT-base	mean	pre-trained	1	0.601	←	0.382	0.606	←	0.394	0.592	←	0.376
		tuned	1	0.603	←	0.364	0.608	←	0.373	0.595	←	0.355
	row	pre-trained	1	0.722	←	0.578	0.71	←	0.568	0.719	←	0.581
		tuned	1	**0.723**	←	0.487	0.711	←	0.472	0.719	←	0.485
mBERT-base	mean	pre-trained	1	0.684	←	0.581	0.683	←	0.585	0.674	←	0.575
		tuned	1	0.684	←	0.59	0.683	←	0.592	0.673	←	0.582
	row	pre-trained	1	**0.771**	←	0.54	0.758	←	0.536	0.764	←	0.54
		tuned	1	**0.771**	←	0.601	0.758	←	0.597	0.765	←	0.598
ruRoberta-large	mean	pre-trained	16	0.485	0.543	0.076	0.495	0.551	0.088	0.475	0.538	0.067
		tuned	16	0.487	0.542	0.154	0.496	0.555	0.166	0.477	0.542	0.146
	row	pre-trained	16	0.452	**0.653**	0.064	0.453	0.641	0.067	0.445	0.652	0.058
		tuned	14	0.453	0.608	0.115	0.454	0.602	0.116	0.446	0.611	0.11
XLM-R-large	mean	pre-trained	14	0.498	0.592	0.394	0.506	0.605	0.404	0.491	0.593	0.385
		tuned	17	0.497	0.588	0.427	0.505	0.595	0.44	0.489	0.582	0.421
	row	pre-trained	11	0.556	0.703	0.424	0.554	0.688	0.414	0.551	0.701	0.418
		tuned	11	0.553	**0.717**	0.45	0.551	0.706	0.449	0.548	0.713	0.446
ruT5-base	mean	pre-trained	1	0.593	←	0.31	0.605	←	0.32	0.587	←	0.308
		tuned	1	0.594	←	0.323	**0.606**	←	0.333	0.588	←	0.321
	row	pre-trained	8	0.552	0.562	0.407	0.544	0.548	0.4	0.549	0.56	0.411
		tuned	8	0.552	0.577	0.442	0.544	0.563	0.434	0.549	0.576	0.445
mT5-base	mean	pre-trained	9	0.575	0.619	0.522	0.583	**0.63**	0.538	0.563	0.611	0.516
		tuned	1	0.573	←	0.471	0.58	←	0.484	0.561	←	0.468
	row	pre-trained	8	0.543	0.621	0.5	0.527	0.615	0.491	0.534	0.619	0.495
		tuned	7	0.535	0.569	0.437	0.518	0.561	0.436	0.526	0.569	0.436

Table 4: Spearman's rank correlations between human attention and models' attention on the first, last, and the layer with the highest correlation values. *Model* – the model's architecture. *Agg.* – the attention scores aggregation: the average of all the rows in each column (*mean*) and the average of all pronoun tokens in each column (*row*). *Checkpoint* – the configuration of the *Model* before (pre-trained) and after (tuned) tuning on the Winograd schema task. *Layer (max)* – the model's layer with the highest correlation value. *Fixations*, *Gaze duration*, *Total reading time* – the human attention characteristics. *first*, *max*, *last* – the model's layers. ← means that the first layer has the highest correlation value (see column *first*). The highest correlation values for each architecture are in bold.

cases, the Accuracy of the pre-trained model is lower than that of fine-tuned models. There are several exceptions for ruRoberta-large, ruT5-base, and XLM-R-large models with incorporated total reading time. The findings from comparing Accuracy between various eye-gaze measurements (*Fixations*, *Gaze duration*, *Reading time*) do not reveal a consistent trend, making it challenging to identify the optimal human signal for incorporating into loss functions.

7 Conclusion

In summary, this paper examines the transformer and human attention mechanisms in the anaphora resolution task. We collected a dataset for the anaphora resolution task using video oculography and released it under the MIT license [2]. We used this dataset to analyze the correlation between machine and human attention across various transformer architectures and network layers. The results show a strong correlation between human and machine attention, but fine-tuning did not enhance this correlation. Therefore, we conducted experiments integrating eye movement data into the model training process. This was done by adding an extra term to the loss function to align the model's attention more closely with human attention. However, the results did not show a consistent trend in the proposed setup, indicating that further research is needed for incorporation approaches.

Limitations

Data Specificity The study relies on an eye-tracking dataset limited to one specific coreference type with a relatively small number of instances. We investigate the results based on data specifically tailored to the Russian language. Therefore, the findings may not be generalizable to other languages or datasets with different linguistic structures and nuances.

We take the privacy and confidentiality of participants seriously when collecting eye-tracking data. All participants provided informed consent, fully understanding the nature of the study and how their data would be utilized. However, we acknowledge that such data may introduce linguistic biases that

[2] https://huggingface.co/datasets/RussianNLP/EyeWino

Model	Agg.	Checkpoint	Layer	Without	Fixations			Gaze duration			Total reading time		
			(max)	integration	first	max	last	first	max	last	first	max	last
ruBERT-base	mean	pre-trained	1	55.77	55.77	←	**59.23**	56.15	←	58.08	55.77	←	58.46
		tuned	1	58.08	58.08	←	58.08	58.08	←	58.08	58.08	←	58.08
	row	pre-trained	1	55.77	55.77	←	57.69	55.77	←	55.77	55.77	←	56.92
		tuned	1	58.08	58.08	←	58.08	58.08	←	58.46	58.08	←	58.08
mBERT-base	mean	pre-trained	1	54.62	56.15	←	56.92	53.08	←	50.00	56.15	←	53.08
		tuned	1	56.15	56.92	←	55.38	55.38	←	56.54	55.00	←	56.54
	row	pre-trained	1	54.62	53.85	←	53.85	55.00	←	54.62	54.62	←	52.69
		tuned	1	56.15	55.38	←	**57.31**	55.38	←	56.92	55.38	←	55.77
ruRoberta-large	mean	pre-trained	16	56.92	58.08	59.23	55.38	53.85	57.31	57.69	56.92	59.23	58.46
		tuned	16	55.38	55.38	54.23	53.85	55.38	58.08	54.23	55.77	56.92	**60.77**
	row	pre-trained	16	56.92	58.08	59.62	59.62	55.00	59.23	59.23	56.54	60.0	56.54
		tuned	14	55.38	53.08	59.62	56.54	58.46	59.23	55.77	54.23	56.54	56.54
XLM-R-large	mean	pre-trained	14	55.00	55.38	48.85	55.00	50.38	52.69	50.00	54.23	56.92	54.23
		tuned	17	54.62	52.31	55.77	56.54	53.46	56.54	55.00	56.92	51.54	56.92
	row	pre-trained	11	55.00	51.15	55.38	54.23	51.15	56.15	54.62	54.23	54.23	51.15
		tuned	11	54.62	53.46	56.54	56.15	54.23	**59.62**	58.85	55.00	55.0	56.92
ruT5-base	mean	pre-trained	1	52.69	**61.15**	←	53.46	56.54	←	58.46	57.31	←	63.46
		tuned	1	55.77	54.62	←	57.31	54.62	←	51.92	52.31	←	50.77
	row	pre-trained	8	52.69	57.69	58.46	51.92	56.54	60.0	53.08	58.08	56.15	49.23
		tuned	8	55.77	53.08	48.08	54.23	53.46	53.46	53.85	51.92	53.85	48.08
mT5-base	mean	pre-trained	9	53.08	53.08	54.23	54.62	54.23	51.54	54.62	54.62	52.69	54.62
		tuned	1	58.46	58.85	←	59.62	58.08	←	57.31	59.23	←	58.46
	row	pre-trained	8	53.08	54.62	54.62	52.69	57.31	51.15	53.46	55.38	52.69	53.08
		tuned	7	58.46	**64.23**	56.54	58.85	59.23	60.0	58.08	62.31	62.31	61.54

Table 5: Accuracy of the experiments with human gaze integration during model training on the first, last, and the layer with the highest correlation values. *Model* – the model's architecture. *Agg.* – the attention scores aggregation: the average of all the rows in each column (*mean*) and the average of all pronoun tokens in each column (*row*). *Checkpoint* – the configuration of the *Model* before (pre-trained) and after (tuned) tuning on the Winograd schema task. *Layer (max)* – the model's layer with the highest correlation value. *Without integration* - the Accuracy of the experiments without human gaze integration. *Fixations*, *Gaze duration*, *Total reading time* – the human attention characteristics. *first*, *max*, *last* – the model's layers. ← means that the first layer has the highest correlation value (see column *first*). The best scores for each architecture are in bold.

can be further transmitted to the neural model by incorporating the attention mechanisms.

Experimental setup The analysis was based on various transformer architectures, but it is important to note that we could not cover all possible attention mechanisms and neural approaches. We focused on the encoder attention layers in the paper, as these layers capture context from the entire input sequence. In contrast, the decoder attention layers can only process earlier positions in the sequence. Investigating the decoder's attention is an issue for future research. Additionally, the quantitative comparison between human and machine attention may be influenced by the intrinsic limitations of the experimental setups, such as the weaknesses of eye-tracking technology, the design of the Winograd schema tasks and the collected dataset, and the interpretability techniques applied to the neural models.

Human attention complexity is a multifaceted phenomenon influenced by numerous cognitive, cultural, and situational factors that have not been investigated. Thus, the current machine attention mechanisms are artificial approximations that are hard to compare. Our study, while comprehensive, only captures a subset of these factors, particularly those that are quantifiable through eye-tracking.

Acknowledgments

We would like to express our heartfelt gratitude to Olga Dragoy and Nina Zdorova from HSE University, as well as Tatiana Shavrina, for their invaluable support in preparing the datasets for our experiments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

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.

Stephanie Brandl and Nora Hollenstein. 2022. Every word counts: A multilingual analysis of individual human alignment with model attention. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 72–77, Online only. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2016. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49.

Armanda Costa, Gabriela Matos, and Paula Luegi. 2011. Using eye-tracking to study anaphoric relations processing in european portuguese.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan A Duffy and Keith Rayner. 1990. Eye movements and anaphor resolution: Effects of antecedent typicality and distance.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? pages 4295–4309.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, et al. 2024. Mera: A comprehensive llm evaluation in russian. *arXiv preprint arXiv:2401.04531*.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. 2013. *Nonparametric statistical methods*. John Wiley & Sons.

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

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. 2011. *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Kaidi Lõo, Marco Marelli, and Kerem Usal. 2022. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, 45:1–35.

Anna Laurinavichyute, Irina Sekerina, Svetlana Alexeeva, Kristina Bagdasaryan, and Reinhold Kliegl. 2019. (2019). article russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, 51:1161–1178.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Steven Luke and Kiel Christianson. 2017. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Shamita Chantherasarathy Naido and Nurjanah Mohd Jaafar. 2022. Anaphora resolution in reading among malaysian l2 english speakers: An eye-tracking investigation. *Jurnal Wacana Sarjana*, 6(4):1–13.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? pages 4596–4608.

Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, 54:1–21.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Alena Spiridonova, Valentina Kurenshchikova, Ekaterina Artemova, and Vladislav Mikhailov. 2022. TAPE: Assessing few-shot Russian language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain).

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Agata Wolna, Joanna Durlik, and Zofia Wodniecka. 2024. Correction: Pronominal anaphora resolution in polish: Investigating online sentence interpretation using eye-tracking.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

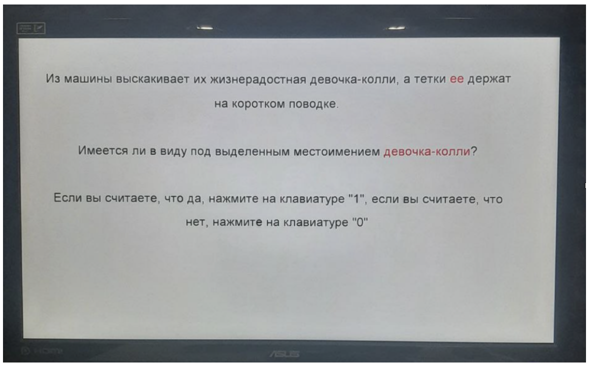
Leran Zhang and Nora Hollenstein. 2024. Eye-tracking features masking transformer attention in question-answering tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7057–7070, Torino, Italia. ELRA and ICCL.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, et al. 2023. A family of pretrained transformer language models for russian. *arXiv preprint arXiv:2309.10931*.

Appendix

A Participant Instructions

Fig. 1 contains an example format of a task for participants, consisting of the following parts: a text, a question about the text, and an instruction for the task.



Из машины выскакивает их жизнерадостная девочка-колли, а тетки ее держат на коротком поводке.

Имеется ли в виду под выделенным местоимением девочка-колли?

Если вы считаете, что да, нажмите на клавиатуре "1", если вы считаете, что нет, нажмите на клавиатуре "0"

Figure 1: The example of a task shown to participants on the screen.

B Eye-movement Measures

The eye-tracking dataset contains the following fields:

- **word**, a word in a sentence;
- **example_id**, id of the example in the dataset;

- **text_id**, id of the unique text in the dataset;
- **position_id**, position of the word in the sentence;
- **annotator_id**, experiment participant id;
- **is_answer_correct**, the correctness of the experiment participant's answer;
- **reading_time**, the sum of all fixation durations on the current word, ms;
- **gaze_duration**, the sum of all fixation durations on the current word in the first-pass reading, ms;
- **fixations**, the number of all fixations on the current word;
- **first_fixation_duration**, the duration of the first fixation on the word, ms;
- **x_coordinate_first_fixation**, the coordinate of the first fixation on the word along the x axis, where the screen is the coordinate plane;
- **y_coordinate_first_fixation**, the coordinate of the first fixation on the word along the y axis, where the screen is the coordinate plane;
- **amplitude_first_saccade**, the amplitude of the first saccade, deg;
- **correct_antecedent**, the correct antecedent for example_id;
- **incorrect_antecedent**, the incorrect antecedent for example_id;
- **pronoun**, an anaphoric pronoun for example_id;
- **is_pronoun**, an indicator of whether the word is the anaphoric pronoun;
- **label**, an indicator of whether the question is about the correct antecedent.

C Visualization of Correlations

Fig. 2 provides the correlations between the attention of different model architectures, aggregated using the *mean* approach, and eye-tracking data.

D Visualization of Attention Maps

Fig. 3 provides a visualization of the important words for the human and the models.

Human attention is characterized by the relative importance of words based on *Fixations*. Checkpoints, layers, and aggregations with the highest correlations with the relative importance of words for humans are used to describe the relative importance of words for the models.

The original examples and relative importance of words are in Russian. Below the Russian texts are the English translations of these texts and an adapted visualization of the relative importance of words.

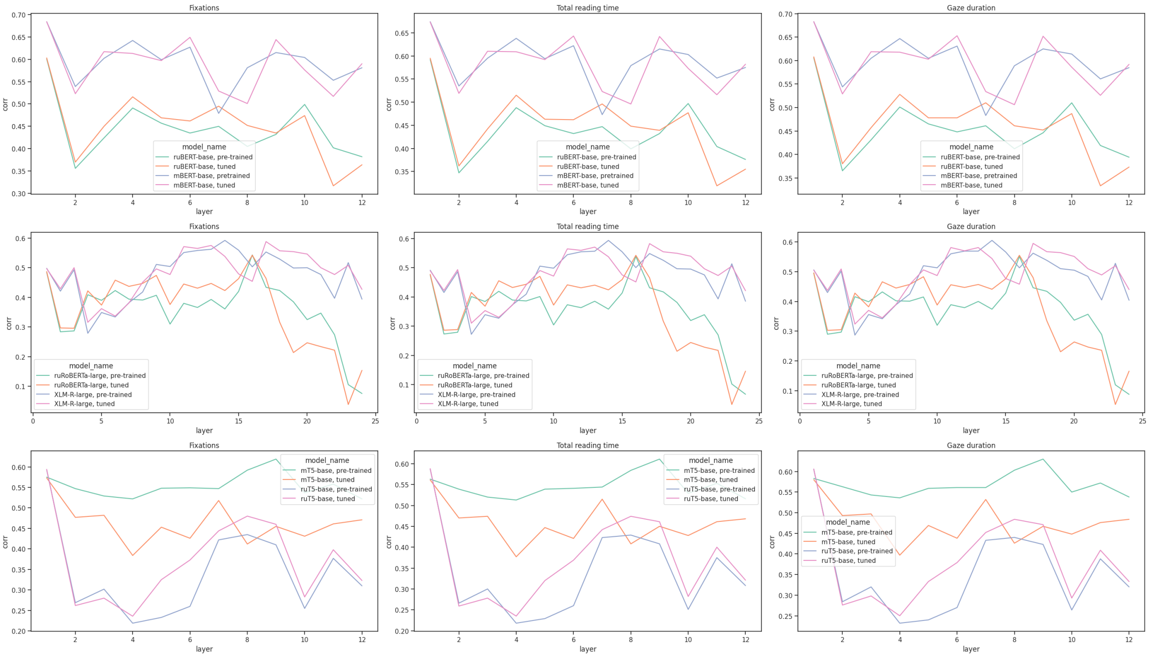
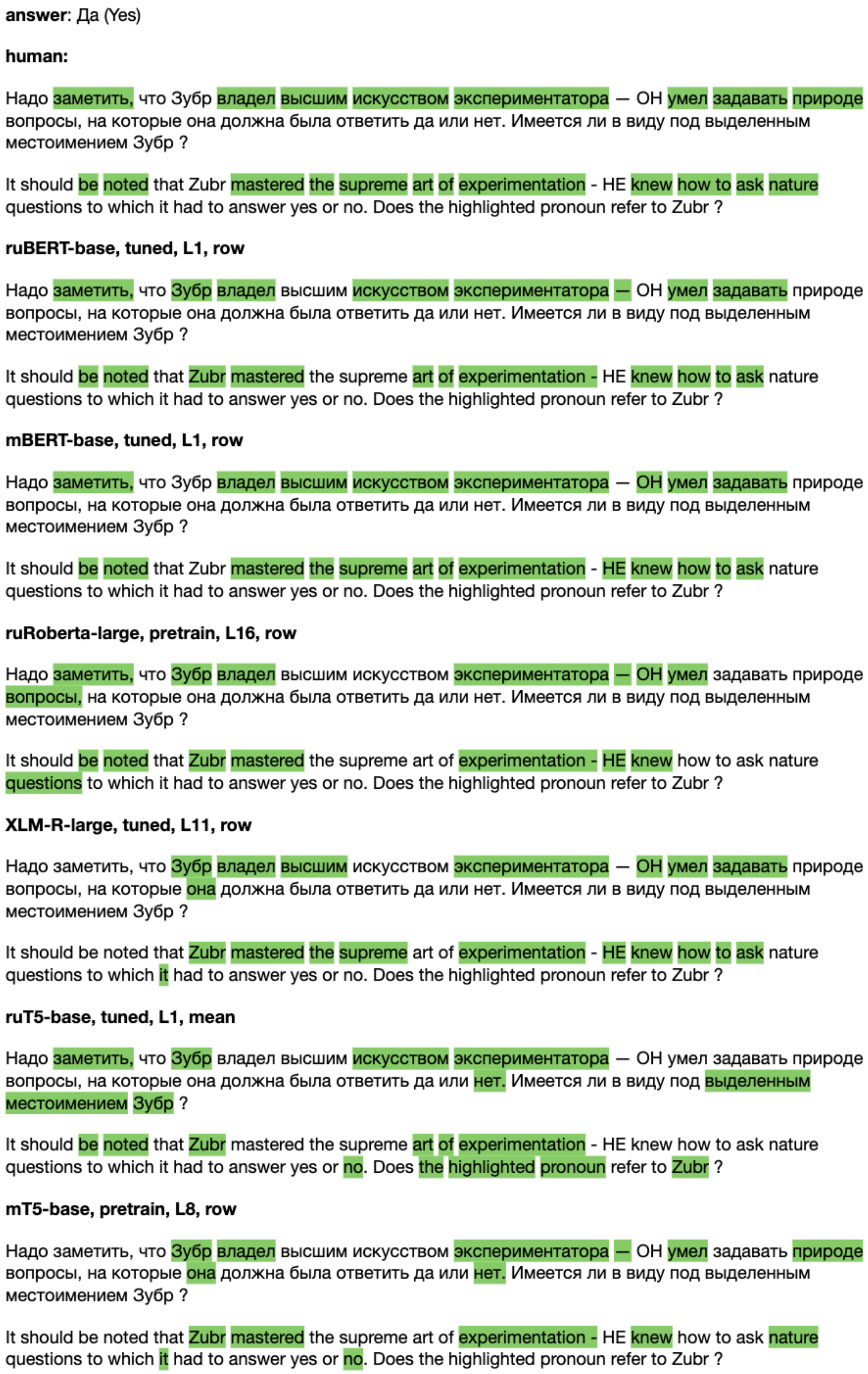


Figure 2: The correlations between models' attention on different layers and eye-tracking data.



answer: Да (Yes)

human:

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

ruBERT-base, tuned, L1, row

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

mBERT-base, tuned, L1, row

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

ruRoberta-large, pretrain, L16, row

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

XLM-R-large, tuned, L11, row

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

ruT5-base, tuned, L1, mean

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

mT5-base, pretrain, L8, row

Надо заметить, что Зубр владел высшим искусством экспериментатора — ОН умел задавать природе вопросы, на которые она должна была ответить да или нет. Имеется ли в виду под выделенным местоимением Зубр ?

It should be noted that Zubr mastered the supreme art of experimentation - HE knew how to ask nature questions to which it had to answer yes or no. Does the highlighted pronoun refer to Zubr ?

Figure 3: Visualizations of human and models' attentions. The words with high relative importance for Russian texts are highlighted in green. The third quartile is used to determine a word's importance.