Comparing Transformer Model Interpretability with Human Cognition: A Dual Analysis of Attention and Attribution

Lingchen Kong, Jinnie Shin, Pavlo Antonenko
University of Florida
l.kong@ufl.edu, jinnie.shin@coe.ufl.edu, p.antonenko@coe.ufl.edu

Abstract

Understanding whether transformer-based language models exhibit processing patterns that correspond to human cognition remains a central question in cognitively motivated computational linguistics. This study examines the statistical correspondence between model interpretability measures and human cognitive signals in a sentence-level relation classification task. We analyze attention weights and attribution scores (Integrated Gradients, Leave-One-Out, and LIME) from six fine-tuned transformer models and compare them to gaze duration derived from the ZuCo 2.0 corpus and human-annotated word importance. Results show that early-layer attention exhibits stable, moderate correlations with gaze duration that are largely unaffected by fine-tuning. Attribution methods, by contrast, demonstrate stronger correlations with human word-importance judgments, particularly for correctly predicted instances, though higher predictive accuracy does not consistently imply stronger human alignment. These findings suggest that interpretability measures capture distinct aspects of model computation, only some of which overlap with human cognitive signals. Systematic, multi-level evaluation against human benchmarks is therefore essential for assessing cognitive plausibility and for developing more transparent, trustworthy, and human-aligned NLP systems.

Keywords: Transformer Models, Explainable AI, Cognitive Modeling, Eye-tracking, Cognitive Plausibility

1. Introduction

Transformer-based language models have substantially advanced natural language processing (NLP) and demonstrated strong performance across multiple downstream tasks, including machine translation (Wang et al., 2022), automatic essay scoring (Shin and Gierl, 2024), and question answering (Nassiri and Akhloufi, 2022). Despite these successes, open questions remain regarding whether the internal computations of these models correspond to patterns observed in human language processing (Gu et al., 2025; Schrimpf et al., 2021). Within cognitive science and computational linguistics, this issue extends beyond performance optimization and relates directly to whether artificial systems can serve as empirically informative models of behavioral patterns observed in human language processing (Goldstein et al., 2025). Recent interdisciplinary work has therefore shifted toward evaluating not only model accuracy but also the extent to which model behaviour corresponds to human attentional and reasoning patterns (Eberle et al., 2022; Gao et al., 2025).

Interpretability methods have become a primary tool for investigating model decision processes (Mi et al., 2020). Among these, attention weights in transformer architectures are often interpreted as indicators of token importance. Because attention distributes representational focus across input tokens, it has been proposed as a potential computational correlate of token salience that may relate to human perceptual attention during reading (Zou et al., 2023). However, whether attention reliably explains model decisions remains debated. Some studies demonstrate correlations between transformer attention distributions and human eye-tracking signals, suggesting partial correspondence (Bensemann et al., 2022; Sen et al., 2020), whereas others argue that attention weights may not faithfully represent causal feature importance (Jain and Wallace, 2019; Serrano and Smith, 2019).

Beyond attention, attribution-based interpretability methods aim to provide more direct explanations of model predictions (Schwalbe and Finzel, 2023). Techniques such as Integrated Gradients (IG, Sundararajan et al., 2017), Leave-One-Out (LOO, Li et al., 2016) feature perturbation, and Local Interpretable Model-Agnostic Explanations (LIME, Ribeiro et al., 2016) attempt to quantify the contribution of individual tokens to model outputs. While these approaches have been widely applied for model transparency, limited research has examined whether attribution scores correspond to human judgments about linguistic relevance or reasoning importance within controlled comprehension tasks. This gap limits the extent to which interpretability findings can be connected to cognitive modeling objectives.

Human cognitive signals provide a valuable external benchmark for evaluating model explanations. Eye-tracking research has demonstrated that fixation duration and gaze patterns reflect attention allocation and lexical processing during reading (Degno and Liversedge, 2020). Similarly, human annotation of word importance can approximate reasoning-level judgments about semantic rele-

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 86–98
16 May 2026.

vance (Lottridge et al., 2023). Integrating these signals enables a multimodal evaluation of whether model interpretability methods capture behavioral patterns associated with perceptual and decision-related processing.

This study investigates statistical correspondence between transformer model interpretability measures and human cognitive signals in a sentence relation classification task. Using six fine-tuned transformer models and the ZuCo 2.0 corpus (Hollenstein et al., 2023), we systematically compare attention weights and attribution-based explanation scores against human gaze duration and annotated word importance. We examine alignment across model layers, interpretability methods, and prediction correctness to determine when model explanations approximate human attentional and decision-related signals and when they diverge.

This work makes three primary contributions. First, it provides a comprehensive comparison between attention-based and attribution-based interpretability methods using both perceptual and reasoning-oriented human benchmarks. Second, it investigates how fine-tuning and model performance influence measured correlation strength with human cognitive signals. Third, it provides a multimodal evaluation framework combining eye-tracking, human annotation, and interpretability analysis to assess whether transformer explanations reflect human-like language processing. By identifying the conditions under which alignment emerges, this study provides empirical constraints on how transformer processing patterns relate to human cognition and informs the development of more transparent and cognitively grounded NLP systems.

2. Related Work

2.1. Neural Language Models as Cognitive Processing Systems

Neural network models have long served as computational approximations of human language processing via distributed representations learned from data (Mikolov et al., 2013; Rumelhart et al., 1986). Transformer-based language models extend this paradigm through self-attention mechanisms that enable global contextual integration and the formation of progressively abstract representations across layers (Vaswani et al., 2017). Emerging evidence suggests that transformer representations exhibit structural properties that parallel functional characteristics associated with multi-stage human language processing (Caucheteux et al., 2021). Empirical studies show that earlier layers tend to encode lexical and syntactic information, whereas deeper layers capture abstract semantic and task-level representations (Jawahar et al., 2019; Tenney et al., 2019). These findings are consistent with cognitive theories proposing hierarchical processing during reading, in which early perceptual and lexical decoding stages precede semantic interpretation and higher-level comprehension processes (Dien, 2009; Hauk et al., 2006). However, predictive success alone does not guarantee cognitive plausibility. As a result, interpretability methods have been used to examine whether internal model computations reflect human-like information processing.

2.2. Interpretability Metrics in Transformer Models

2.2.1. Attention Weights

In transformer architectures, attention weights regulate how information flows between tokens by determining the relative influence of each token when contextualizing others (Vaswani et al., 2017). Through this mechanism, the model learns to distribute focus across the input sequence, enabling it to capture syntactic and semantic dependencies during representation construction (Kobayashi et al., 2020). Because attention explicitly distributes representational focus, it has frequently been interpreted as an indicator of token importance and model reasoning. However, the explanatory validity of attention remains contested. Jain and Wallace (2019) demonstrated that alternative attention distributions can frequently produce identical model predictions, suggesting that attention may not uniquely determine decision outcomes. Similarly, Serrano and Smith (2019) showed that zeroing out high-attention inputs does not substantially change the model's output, and attention-based rankings often fail to identify the most decision-critical inputs relative to gradient-based rankings. These findings indicate that attention may primarily reflect internal representational routing rather than direct causal contribution to predictions. This limitation motivates investigating complementary interpretability methods that more directly capture decision-level influence.

2.2.2. Attribution Scores

Attribution scores quantify each input feature's contribution to the model's output by assigning an importance value relative to a baseline (Hao et al., 2021). Unlike attention, which reflects contextual encoding processes, attribution methods are intended to capture how much each token influences the final prediction, providing insight into model decision processes. In this study, we computed attribution scores in transformer models using three complementary methods. Integrated Gradients (IG)

estimates feature importance by integrating gradients along a continuous path from a baseline input (e.g., zero embedding) to the observed input, producing attributions that reflect the accumulated sensitivity of the model output to each token (Sundararajan et al., 2017). Leave-One-Out (LOO) representation erasure measures importance by removing or masking individual tokens and evaluating the resulting change in model predictions, providing a perturbation-based estimate of each token's contribution (Li et al., 2016). LIME approximates local model decision boundaries by fitting interpretable surrogate models to perturbed versions of the input, where samples closer to the original input are weighted more heavily, yielding locally faithful approximations of token importance (Ribeiro et al., 2016).

Together, these attribution techniques provide complementary perspectives on features that influence model predictions, capturing gradient sensitivity, perturbation-based effects, and local decision approximations. Despite their widespread application in explainable AI (Azad et al., 2025; Gaspar et al., 2024), relatively few studies have examined whether attribution scores correlate with human judgments during language comprehension. Understanding this relationship helps clarify whether attribution-based explanations reflect cognitively meaningful decision processes.

2.3. Evaluating Transformer Attention via Human Gaze Correlations

Eye-tracking research provides fine-grained measurements of attentional allocation during reading, with gaze duration, fixation frequency, and total reading time serving as established indicators of lexical processing difficulty and perceptual attention (Rayner, 2009, 1998; Staub et al., 2010). These gaze signals have therefore been adopted as external benchmarks for evaluating whether neural language models exhibit human-like attention patterns.

Multiple studies have reported moderate to strong correlations between transformer attention distributions and human gaze signals. Bensemann et al. (2022) found that early-layer attention weights in pretrained BERT models moderately align with human dwell time patterns, although they cautioned that better alignment does not necessarily improve model performance. Similarly, Wang et al. (2024) reported moderate to strong positive correlations $(0.39 - 0.78)$ between five different human gaze signals (e.g. gaze duration and the number of fixations) and the BERT model's attention patterns across different layers.

Cross-linguistic studies provide further support for the attention-gaze alignment. Morger et al. (2022) found correlations exceeding 0.5 across German, Dutch, English, and Russian reading tasks when comparing total reading time with attention patterns and the gradient-based saliency. Kozlova et al. (2024) observed positive early-layer correlations $(0.45 - 0.77)$ between attention patterns and gaze signals in Russian texts but found that task-specific fine-tuning did not significantly improve alignment. Consistent with these findings, Eberle et al. (2022) reported minimal changes in attention-gaze alignment following BERT fine-tuning. Together, these studies indicate that, in several reading-related settings, attention distributions from pretrained transformer models have shown statistically significant correlations with selected human gaze measures.

2.4. Integrating Human Gaze Signals into Transformer Training

Recent research has sought to actively integrate predicted or measured gaze signals into transformer architectures. Dong et al. (2022) introduced the GazBy, a joint model that integrates human gaze fixation estimation with BERT and showed improved performance in passage re-ranking tasks. Sood et al. (2020) proposed a hybrid text saliency model that mimics the human gaze patterns and demonstrated improved performance in paragraph generation and sentence compression when integrated into transformer attention layers. Wang et al. (2024) demonstrated that incorporating predicted gaze signals into BERT-based models can improve performance across sentiment analysis and semantic similarity tasks, while Zhang and Hollenstein (2024) reported improved question-answering performance using eye-tracking attention masks.

Parallel findings in vision tasks further highlight the potential of integrating human cognitive signals during training. Sharan et al. (2019) demonstrated improved visual question-answering performance by training models to attend to human-identified salient image regions, while Rong et al. (2021) showed similar improvements in visual classification using gaze-based saliency maps. In software engineering, EyeTrans incorporated human attention into transformer-based code summarization, yielding improved code summarization performance (Zhang et al., 2024).

However, integrating human gaze signals remains methodologically challenging. Kozlova et al. (2024) found inconsistent performance improvements when incorporating gaze signals as training objectives, while Dong et al. (2022) demonstrated that performance gains depend strongly on where gaze information is integrated within model architectures. These findings highlight the complexity of effective human-data integration and emphasize

that alignment with human cognition is beneficial only when operationalized with consideration for model architecture and task objectives.

3. Methods

3.1. Data Collection

This study utilized the ZuCo2.0 Corpus (Hollenstein et al., 2023), which provides synchronized eye-tracking and EEG data from 16 native English-speaking adults during sentence reading. The corpus includes both a normal reading and a task-specific relation identification paradigm. The present study focused exclusively on the task-specific paradigm, in which participants actively identify semantic relations between entities within sentences. Sentences in the task paradigm were collected from Wikipedia and annotated with relation categories such as "Political Affiliation", "Employment", and "Education". Sentences that did not express a target relation were labelled as "Control" and served as a reduced relational-demand baseline condition.

Two complementary human-derived measures were used as evaluation benchmarks. First, attentional allocation during reading was operationalized using word-level gaze duration, defined as the total fixation time on each word aggregated across fixations and averaged across participants. Gaze duration is a well-established indicator of lexical processing difficulty and reading-time allocation (Carpenter and Just, 2017; Strandberg et al., 2023).

Second, decision-relevant salience was measured through human annotation of word importance for relation classification. Two annotators independently ranked the top 20% of words they judged to be most critical for identifying each sentence relation, following interpretability thresholding practices established in prior attribution research (Ju et al., 2022). Rankings were converted into graded importance weights based on selection order, with unselected words assigned zero weight. Inter-rater agreement was measured using Spearman's rank correlation, as annotators produced ordinal word-importance rankings. Agreement was 0.77 across all sentences and 0.84 when Control sentences were excluded, indicating substantial consistency in relational word-importance judgments.

3.2. Transformer Models and Fine-Tuning

Six pretrained transformer models were evaluated to examine whether human-model alignment is consistent across architectural and representational differences. The selected models were grouped into three categories. Base models included BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019), representing standard transformer architectures. Compact models included ALBERT and DistilBERT (Lan et al., 2020; Sanh et al., 2019), which reduce parameters via weight sharing or compression. Alternative models included DeBERTa variants (He et al., 2021) with disentangled attention for improved relational encoding.

To fine-tune transformer models for sentence relations prediction, we collected a training set of 3,353 English sentences from Wikipedia, matching the source of the ZuCo 2.0 task-specific sentences in category distribution, sentence length, and lexical complexity (measured by Flesch Reading Ease). The dataset was split into 80% for model training and 20% for validation. Fine-tuning was conducted for one and three epochs, which were treated as controlled manipulations of task adaptation strength and performance optimization. Training used AdamW optimization, a learning rate of 2e-5, batch size 16, and 10% warmup steps. Model performance was evaluated using F1 score and accuracy.

3.3. Model Interpretability Measures

Model-derived interpretability measures were extracted to parallel the two human cognitive benchmarks. Attention weights were obtained from both the first and last transformer layers to capture early and late contextual integration. Within each layer, attention weights were averaged across attention heads and aggregated column-wise to estimate the total incoming attention weight assigned to each token. Sub-word attention values were summed to produce word-level attention scores compatible with gaze measurements. Token-level attribution scores were estimated using IG, LOO, and LIME. For IG and LOO, sub-token scores were aggregated to the word level through summation, while LIME produced word-level scores directly. Attribution scores were ranked within each sentence, and the top 20% of words were retained to match the human annotation protocol, enabling direct comparison between model-derived and human salience patterns.

3.4. Analysis Plan

The analyses followed the unified evaluation pipeline illustrated in Figure 1, which summarizes data preparation, model training, and comparison procedures across experiments. Experiment 1 evaluates the alignment between transformer attention weights and human gaze duration, assessing whether the distribution of attention across tokens corresponds to human gaze patterns during relational reasoning. Correlation magnitudes are compared across model architectures and fine-tuning

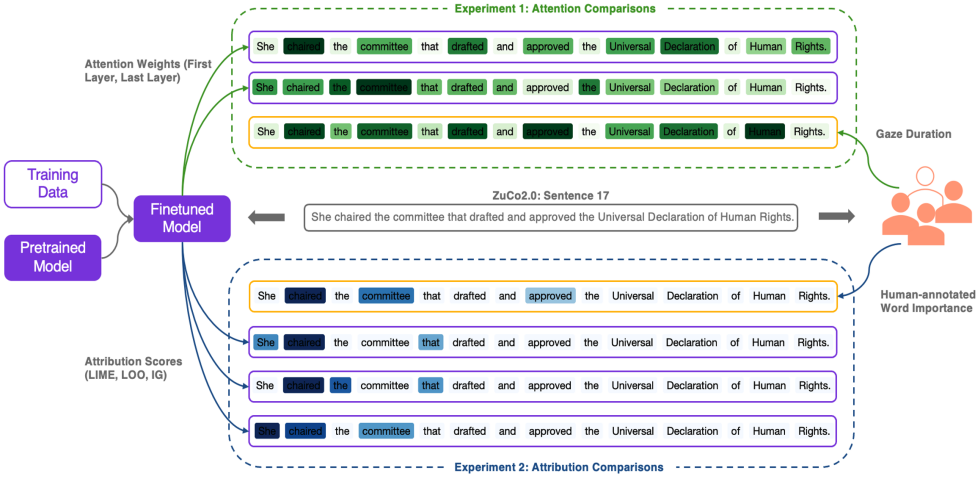

Figure 1: Conceptual overview of the analysis pipeline for evaluating human–model alignment in sentence-level relation classification. Darker shades indicate higher values (attention weight or attribution score). Blue shading corresponds to attribution scores, and green shading corresponds to attention weights.

stages to examine how differences in task performance relate to differences in human-model alignment.

Experiment 2 examines alignment between attribution scores and human-annotated word importance, evaluating whether words deemed influential by the model coincide with those selected by human annotators as critical for identifying relational meaning. The analysis further compares correlation strength across attribution methods and model correctness (correct vs. incorrect predictions) to determine whether stronger alignment is associated with higher prediction accuracy.

4. Results

4.1. Attention-Gaze Alignment

Table 1 summarizes average sentence-level correlations between transformer attention weights and human gaze duration across architectures, layers, and training epochs. The strongest attention-gaze alignment emerged in the first transformer layer, particularly for BERT and DistilBERT. BERT exhibited the highest correlation (0.657 across epochs), followed closely by DistilBERT (≈ 0.61). DeBERTa models showed weaker but consistently positive first-layer correlations ($\approx\ 0.28 - 0.34$), whereas RoBERTa demonstrated negligible alignment in both layers. ALBERT diverged from this pattern, showing higher correlations in the last layer than the first (last layer $= 0.298$ and 0.248 vs. first layer $= -0.044$ and 0.032), despite its parameter-sharing architecture.

Figure 2 illustrates the distribution of these correlations. First-layer distributions for BERT and DistilBERT are clearly right-shifted, with medians

Model	Epoch	F1	Corr. (Layer)	
			First	Last
ALBERT-base	1	0.73	-0.044	**0.298**
ALBERT-base	3	0.76	0.032	**0.248**
DistilBERT	1	0.76	**0.608**	0.068
DistilBERT	3	0.80	**0.613**	0.020
BERT-base	1	0.71	**0.657**	0.167
BERT-base	3	0.78	**0.657**	0.106
RoBERTa-base	1	0.75	**0.018**	0.054
RoBERTa-base	3	0.77	**0.017**	0.005
DeBERTa-base	1	0.74	**0.283**	0.183
DeBERTa-base	3	0.77	**0.287**	-0.011
DeBERTa-large	1	0.69	**0.337**	0.174
DeBERTa-large	3	0.77	**0.328**	0.157

Note. Strongest correlations per model are in bold.

Table 1: Average sentence-level correlations between human gaze duration and transformer attention weights by model, layer, and fine-tuning epoch

near the reported means and relatively tight dispersion, indicating relatively stable moderate alignment across many sentences. DeBERTa variants show weaker but still positive first-layer alignment, while RoBERTa's distributions cluster around zero in both layers, reflecting uniformly low correlations with human gaze duration. Last-layer distributions generally shift toward lower values or include broader negative tails, with ALBERT as the notable exception whose last-layer distribution centers in the positive range.

To understand whether better model performance in F1 translates into improvement in attention-gaze alignment from first layer, we tested the sentence-level paired differences in correla-

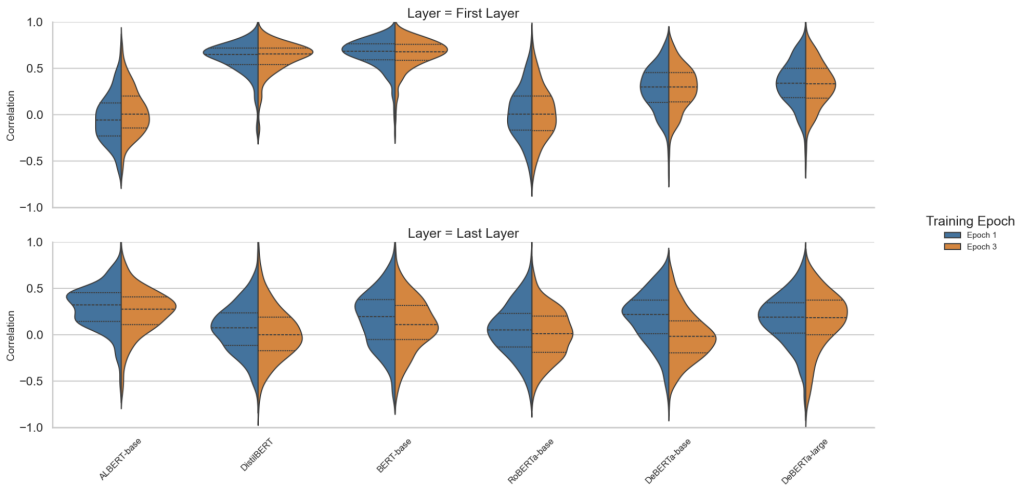

Figure 2: Distribution of sentence-level correlations between human gaze duration and transformer attention weights by model, layer, and fine-tuning epoch.

tion between epoch 1 and epoch 3 using Wilcoxon signed-rank tests. Results show no significant improvement in correlation between attention weights and human gaze duration after fine-tuning for any model except ALBERT ($Z = 11.95, p < .001, r = 0.66$). For all other models, the Wilcoxon signed-rank test did not support the hypothesis that fine-tuning improves alignment. Despite small increases in mean correlation for some models (e.g., DistilBERT: 0.608 to 0.613), the signed-rank distribution indicated that most sentences experienced reduced alignment ($Z = -3.83, p < .001, r = -0.21$). These findings suggest that attention-gaze alignment is more strongly associated with model architecture and pretraining conditions than with task-specific fine-tuning.

4.2. Attribution-Word Importance Alignment

Table 2 summarizes average correlations between human-annotated word importance and model attributions across methods (IG, LOO, LIME) and training epochs. IG consistently yields the strongest alignment, with RoBERTa achieving the highest overall mean correlation (0.374), followed by BERT (0.352) and DistilBERT (0.333). LOO generally performs mid-range and is the top method only for ALBERT, where it slightly exceeds IG (0.283 vs. 0.255 at epoch 3). LIME produced the weakest alignment overall, with correlations near zero for several models.

Table 3 reports correlations between human-annotated word importance and model attributions for sentences the models classified correctly versus incorrectly. The results showed a consistent pattern where the attribution scores from IG and LOO demonstrated higher correlations with human word

Model	Epoch	Attribution Scores		
		IG	LOO	LIME
ALBERT-base	1	0.244	**0.282**	0.150
ALBERT-base	3	0.255	**0.283**	0.117
DistilBERT	1	**0.333**	0.270	0.133
DistilBERT	3	**0.332**	0.277	0.135
BERT-base	1	**0.304**	0.251	0.094
BERT-base	3	**0.352**	0.272	0.124
RoBERTa-base	1	**0.345**	0.237	0.176
RoBERTa-base	3	**0.374**	0.253	0.176
DeBERTa-base	1	0.159	**0.227**	0.104
DeBERTa-base	3	**0.229**	0.222	0.069
DeBERTa-large	1	**0.333**	0.181	0.111
DeBERTa-large	3	**0.308**	0.247	0.142

Note. Strongest correlations per model are in bold.

Table 2: Average sentence-level correlations between human word-importance judgments and model attribution scores across attribution methods, models, and fine-tuning epochs

importance in correctly predicted sentences compared to incorrectly predicted ones. For example, BERT showed an IG correlation of 0.194 for incorrect predictions versus 0.400 for correct predictions at epoch 3. And ALBERT exhibited a notable LOO correlation increase from 0.137 to 0.336 at epoch 1. LIME exhibited smaller and more inconsistent differences across prediction outcomes. Group differences were evaluated using Mann-Whitney U tests, which confirmed significantly higher attribution–importance alignment for correct predictions across models and epochs for both IG and LOO ($p < .001$). No consistent significant differences were observed for LIME. These results indicate that when models make correct decisions, their attribu-

Model	Epoch	Attribution Scores					
		IG		LOO		LIME	
		IC	C	IC	C	IC	C
ALBERT-base	1	0.139	**0.283**	0.137	**0.336**	0.104	**0.166**
ALBERT-base	3	0.192	**0.274**	0.149	**0.323**	**0.156**	0.105
DistilBERT	1	0.167	**0.382**	0.157	**0.303**	0.133	**0.135**
DistilBERT	3	0.177	**0.372**	0.180	**0.302**	**0.136**	0.135
BERT-base	1	0.159	**0.361**	0.128	**0.300**	0.064	**0.105**
BERT-base	3	0.194	**0.400**	0.184	**0.297**	0.110	**0.179**
RoBERTa-base	1	0.170	**0.406**	0.123	**0.276**	**0.194**	0.169
RoBERTa-base	3	0.207	**0.426**	0.170	**0.280**	**0.211**	0.165
DeBERTa-base	1	0.019	**0.208**	0.104	**0.270**	0.082	**0.111**
DeBERTa-base	3	0.088	**0.271**	0.128	**0.249**	0.051	**0.075**
DeBERTa-large	1	0.215	**0.389**	0.078	**0.230**	0.101	**0.116**
DeBERTa-large	3	0.206	**0.339**	0.182	**0.267**	**0.183**	0.130

Note. Strongest correlations per condition are in bold.

Table 3: Average sentence-level correlations between human word-importance judgments and model attribution scores, separated by incorrect (IC) and correct (C) predictions

tion scores (particularly IG and LOO) more closely reflect human judgments of word importance, while LIME provides comparatively limited alignment.

5. Discussion

Evaluating whether neural language models approximate human cognitive processes during reading remains a challenge for cognitively grounded natural language processing (Kozlova et al., 2024). The present study contributes to this objective by systematically comparing transformer interpretability measures (i.e. attention weights, attribution scores) and two complementary human cognitive indicators (i.e., gaze duration, human-annotated word importance) during sentence-level relation classification. Across models, early-layer attention weights showed moderate and consistent alignment with gaze duration, whereas later-layer attention exhibited reduced alignment as the models progressed toward making final predictions. On the other hand, attributions scores, particularly Integrated Gradients (IG) and Leave-One-Out (LOO), showed more consistent and higher alignment with human-annotated word importance, especially for correctly predicted sentences. However, alignment patterns varied across model architectures and attribution methods, and perturbation-based LIME explanations showed comparatively unstable alignment.

These findings refine previous studies that early-layer attention aligns with human gaze behavior (Bensemann et al., 2022; Eberle et al., 2022; Sen et al., 2020; Brandl and Hollenstein, 2022) by showing that such alignment is stable and largely unaffected by fine-tuning, supporting the interpretation that early attention captures general perceptual or lexical features formed during pre-training (Kozlova et al., 2024). From a cognitive modeling perspective, this pattern supports the view that attention often reflects positional and distributional prominence, which aligns with human perceptual focus under the eye-mind hypothesis (Just and Carpenter, 1980), particularly when cognitive engagement is low. Our analysis of gaze duration reveals that participants tend to allocate more attention to visually distinct or complex words, such as numbers or uncommon terms, a pattern partially reflected in early transformer attention distributions, which is influenced by token frequency, orthographic complexity, and relative positioning. The weaker alignment observed in final-layer attention likely reflects increasing abstraction toward task-specific output representations (Clark et al., 2019; van Aken et al., 2019), which may rely on non-explicit or aggregate cues rather than human-salient information.

Our results extend prior work by demonstrating that attribution scores more closely approximate human decision-level reasoning. This finding is consistent with prior research emphasizing the utility of attribution scores in approximating human judgment in automated scoring studies. For example, Lottridge et al. (2023) found that IG-based saliency maps (i.e., word importance maps) aligned closely with human annotations in automated scoring tasks, though challenges persisted for longer responses and phrase-level distinctions. Similarly, Poulton and Eliens (2021) discovered that gradient-based techniques, especially Input X Gradient and Integrated Gradients, demonstrated the highest consistency in aligning with human-selected keywords when used on transformer-based Automated Short Answer Grading (ASAG) models trained with the

SQuAD 2.0 dataset.

We further observed that alignment between attribution scores and human word importance improves substantially for correctly predicted sentences. In these cases, both humans and models may rely on highly diagnostic semantic cues, which is consistent with principles of bounded rationality, where decision-makers prioritize the most informative elements when processing information under cognitive constraints (Gigerenzer and Goldstein, 1996). However, improved task performance did not uniformly translate into stronger human-model alignment. Notably, models such as DeBERTa-large often achieved higher F1 scores by emphasizing frequent but semantically trivial function words (e.g., was, of, the, a), instead of relying on meaning-bearing content. This divergence suggests that, while humans selectively attend to conceptually relevant cues to support semantic reasoning, models may instead exploit superficial statistical regularities to optimize classification. This pattern reflects the Clever Hans effect (Lapuschkin et al., 2019), in which performance gains arise from reliance on data set-specific artifacts rather than cognitive strategies relevant to the task, ultimately compromising human reasoning alignment and limiting generalizability.

Beyond empirical findings, this study contributes conceptually by integrating both process-level (gaze duration) and decision-level (attribution) indicators to evaluate transformer interpretability. Existing frameworks often compare explanations to post hoc human annotations (Doshi-Velez and Kim, 2017) without accounting for the cognitive processes underlying those judgments. Our dual-alignment approach addresses this gap by linking model attention to real-time human reading behavior and attribution scores to deliberative reasoning. This distinction reflects dual-process views of reading comprehension, where initial lexical processing precedes meaning integration (Rayner, 1998), and supports ongoing discussions that neural representations may align with cognitive operations at different architectural layers (Nonaka et al., 2021; Schirrmeister et al., 2017; Yamins et al., 2014). In addition, we show that higher task accuracy does not necessarily imply better cognitive alignment, reinforcing arguments that explainability must be evaluated on human-relevant criteria rather than performance alone (Lipton, 2018). By identifying where alignment emerges and where it breaks down, such as when models exploit shallow structural cues for prediction, our findings fill a critical gap in understanding how attribution methods approximate human reasoning (Garcia et al., 2024). This offers a pathway for refining explanation techniques to produce cognitively meaningful and trustworthy model interpretations.

From a practical perspective, these results provide actionable guidance for the development of cognitively informed transformer systems. The observed positive association between attribution-word importance alignment and model performance, particularly for IG and LOO, suggests that incorporating human-annotated importance signals into training or fine-tuning may promote models that reason in ways more consistent with human cognitive judgment (Bhatia and Richie, 2024; Hong et al., 2024). Such alignment is especially relevant in applications where justification quality matters as much as accuracy, such as automated essay scoring, clinical text assessment, and educational decision-support systems (Schneider et al., 2022). Our results also demonstrate that early-layer attention naturally aligns with human gaze patterns irrespective of task-specific tuning, implying that pre-trained models already encode perceptual regularities that approximate low-level human reading behavior. This suggests that approximate gaze signals could be simulated using early attention mechanisms in contexts where eye-tracking data is unavailable (Hollenstein et al., 2021), reducing data collection costs while preserving cognitive relevance. Importantly, these implications should be approached with caution: human-model alignment is not guaranteed by high accuracy but emerges more reliably in cases of correct, semantically grounded reasoning. This underscores the value of integrating human-centered interpretability evaluations into model development workflows to discourage reliance on superficial cues and promote trustworthy decision-making.

6. Conclusion

This study examined the extent to which transformer interpretability measures statistically correspond to human-derived indicators of reading and decision processes in a sentence relation classification task. By jointly evaluating attention–gaze and attribution–word importance correlations across architectures, layers, and fine-tuning stages, we provide a structured assessment of human–model alignment at both process and decision levels. Results show that early-layer attention weights exhibit stable, moderate correlations with gaze duration that are largely unaffected by task-specific fine-tuning, suggesting that pretrained representations encode perceptual or lexical regularities that overlap with human reading-time allocation. In contrast, attribution methods, particularly Integrated Gradients and Leave-One-Out, demonstrate stronger correlations with human-annotated word importance, especially for correctly predicted sentences. However, higher predictive performance does not uniformly imply stronger alignment, and

some high-performing models rely on superficial cues that diverge from human-identified relevance. Taken together, these findings indicate that different interpretability measures capture distinct aspects of model computation, only some of which overlap with human cognitive signals. Systematic evaluation against human-derived benchmarks is therefore necessary to assess the cognitive plausibility of transformer explanations and to inform the development of more transparent and human-aligned NLP systems.

7. Limitations and Future Work

Despite offering new insights into human-model alignment, this study is subject to several methodological and conceptual constraints. First, the operationalization of human cognition was limited to gaze duration and word-importance ratings from a single eye-tracking dataset (ZuCo2.0). While gaze duration is an established indicator of attentional allocation, it does not fully capture deeper reasoning or semantic integration processes. Similarly, binary top 20% annotations simplify human judgment into discrete relevance tiers. Future research should incorporate temporally dynamic and multivariate cognitive signals such as fixation regressions, saccade patterns, or EEG measures (Hollenstein et al., 2020), and expand to datasets featuring greater linguistic, demographic, and task diversity to strengthen generalizability.

Second, the analysis was restricted to encoder-based transformers within the BERT family. While this constraint enabled controlled architectural comparison, it limits conclusions about autoregressive decoder-only LLMs or hybrid architectures. Evaluating models trained under different objectives (e.g., causal language modeling) would clarify whether observed alignment patterns depend on training signals or architectural inductive biases.

Third, interpretability measures were evaluated primarily through correlation analysis. While correlation provides an interpretable and comparable summary metric, it compresses structured alignment into a single scalar statistic. This aggregation obscures distributional properties such as sparsity, positional concentration, or higher-order interaction patterns across tokens. Future work should complement correlation-based metrics with distribution-sensitive similarity measures, permutation controls, and intervention-based evaluations to more fully characterize explanatory correspondence.

Finally, the scope of investigation was limited to sentence-level relation classification. Extending this research to more complex, open-ended, or generative NLP tasks (e.g., reasoning, automated scoring, long-form comprehension), as well as non-text domains such as computer vision or multimodal learning, would enable broader evaluation of human-model alignment across cognitive demands and representational structures.

8. Acknowledgements

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

9. Bibliographical References

Mohammad Azad, Md Faraz Kabir Khan, and Sameh Abd El-Ghany. 2025. Xai-enhanced machine learning for obesity risk classification: A stacking approach with lime explanations. *IEEE Access*.

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.

Sudeep Bhatia and Russell Richie. 2024. Transformer networks of human conceptual knowledge. *Psychological Review*, 131(1):271–306.

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.

Patricia A. Carpenter and Marcel Adam Just. 2017. *Cognitive Processes in Reading: Models Based on Readers' Eye Fixations*, page 177–213. Routledge.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pages 1336–1348. PMLR.

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

Federica Degno and Simon P. Liversedge. 2020. Eye movements and fixation-related potentials in reading: A review. *Vision*, 4(1):11.

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.

Joseph Dien. 2009. The neurocognitive basis of reading single words as seen through early latency erps: A model of converging pathways. *Biological Psychology*, 80(1):10–22.

Sibo Dong, Justin Goldstein, and Grace Hui Yang. 2022. Gazby: Gaze-based bert model to incorporate human attention in neural information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 182–192.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

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

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090.

Basile Garcia, Crystal Qian, and Stefano Palminteri. 2024. The moral turing test: Evaluating human-llm alignment in moral decision-making.

Diogo Gaspar, Paulo Silva, and Catarina Silva. 2024. Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, 12:30164–30175.

Gerd Gigerenzer and Daniel G. Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, Omer Levy, Roi Reichart, and Uri Hasson. 2025. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1).

Chanyuan Gu, Samuel A. Nastase, Zaid Zada, and Ping Li. 2025. Reading comprehension in l1 and l2 readers: neurocomputational mechanisms revealed through large language models. *npj Science of Learning*, 10(1).

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

O. Hauk, M.H. Davis, M. Ford, F. Pulvermüller, and W.D. Marslen-Wilson. 2006. The time course of visual word recognition as revealed by linear regression analysis of erp data. *NeuroImage*, 30(4):1383–1400.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. 2023. The zuco benchmark on cross-subject

reading task classification with eeg and eye-tracking data. *Frontiers in Psychology*, 13.

Seung-Kyu Hong, Jae-Seok Jang, and Hyuk-Yoon Kwon. 2024. Enhancing performance of transformer-based models in natural language understanding through word importance embedding. *Knowledge-Based Systems*, 304:112404.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

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*.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2022. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, Dublin, Ireland. Association for Computational Linguistics.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Anastasia Kozlova, Albina Akhmetgareeva, Aigul Khanova, Semen Kudriavtsev, and Alena Fenogenova. 2024. Transformer attention vs human attention in anaphora resolution. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 109–122, Bangkok, Thailand. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1).

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Susan Lottridge, Sherri Woolf, Mackenzie Young, Amir Jafari, and Chris Ormerod. 2023. The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning*, 39(3):787–803.

Jian-Xun Mi, An-Di Li, and Li-Fang Zhou. 2020. Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, 8:191969–191985.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Khalid Nassiri and Moulay Akhloufi. 2022. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. 2021. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013.

Andrew Poulton and Sebas Eliens. 2021. Explaining transformer-based models for automatic short answer grading. In *2021 5th International Conference on Digital Technology in Education*, ICDTE 2021, page 110–116. ACM.

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

Keith Rayner. 2009. Eye movements in reading: Models and data. *Journal of eye movement research*, 2(3):28.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. 2021. Human attention in fine-grained classification.

David E Rumelhart, James L McClelland, PDP Research Group, et al. 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

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

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420.

Johannes Schneider, Robin Richner, and Micha Riser. 2022. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 33(1):88–118.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101.

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? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online. Association for Computational Linguistics.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2931–2951.

Komal Sharan, Ashwinkumar Ganesan, and Tim Oates. 2019. *Improving Visual Reasoning with Attention Alignment*, page 219–230. Springer International Publishing.

Jinnie Shin and Mark J. Gierl. 2024. *Automated Short-Response Scoring for Automated Item Generation in Science Assessments*, page 504–534. Routledge.

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.

Adrian Staub, Sarah J White, Denis Drieghe, Elizabeth C Hollway, and Keith Rayner. 2010. Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1280.

Andrea Strandberg, Mattias Nilsson, Per Östberg, and Gustaf Öqvist Seimyr. 2023. Eye movements are stable predictors of word reading ability in young readers. *Frontiers in Education*, 8.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions?: A layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1823–1832. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bingbing Wang, Bin Liang, Lanjun Zhou, and Ruifeng Xu. 2024. Gaze-infused bert: Do human gaze signals help pre-trained language models? *Neural Computing and Applications*, 36(20):12461–12482.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

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.

Yifan Zhang, Jiliang Li, Zachary Karas, Aakash Bansal, Toby Jia-Jun Li, Collin McMillan, Kevin Leach, and Yu Huang. 2024. Eyetrans: Merging human and machine attention for neural code summarization. *Proceedings of the ACM on Software Engineering*, 1(FSE):115–136.

Jiajie Zou, Yuran Zhang, Jialu Li, Xing Tian, and Nai Ding. 2023. Human attention during goal-directed reading comprehension relies on task optimization. *eLife*, 12.