Comparing Eye-gaze and Transformer Attention Mechanisms in Reading Tasks

Maria Mouratidi
Utrecht University
m.mouratidi@students.uu.nl

Massimo Poesio
Utrecht University
Queen Mary University of London
m.poesio@uu.nl

Abstract

As transformers become increasingly prevalent in NLP research, evaluating their cognitive alignment with human language processing has become essential for validating them as models of human language. This study compares eye-gaze patterns in human reading with transformer attention using different attention representations (raw attention, attention flow, gradient-based saliency). We employ both statistical correlation analysis and predictive modeling using PCA-reduced representations of eye-tracking features across two reading tasks. The findings reveal lower correlations and predictive capacity for the decoder model compared to the encoder model, with implications for the gap between behavioral performance and cognitive plausibility of different transformer designs.

1 Introduction

The impressive capabilities of Transformer models in linguistic tasks have revolutionized Language Models in Natural Language Processing (NLP) research. A key difference in their architecture from previous models is the incorporation of an attention mechanism, which assigns a degree of relevance between words in the input. Previous work has shown that transformer models show signs of processing steps similar to humans (Clark et al., 2019; Voita et al., 2019), and tend to mirror the structure of the classic NLP pipeline (Tenney et al., 2019).

Attention during reading has also been extensively studied in human eye-movement research. Eye-movements track much of linguistic processing, including both lower-level word processing (Just and Carpenter, 1980; Clifton Jr. et al., 2007) and higher-level comprehension (Reichle et al., 2010; Southwell et al., 2020).

While transformer attention is not explicitly modeled after human attention in text processing, both mechanisms seem to process text by allocating resources on relevant linguistic targets. This similarity, combined with the broader effort of explainability research to explain artificial models in human terms, has driven comparisons of model attention to human eye-movement patterns. From a cognitive science perspective, the goal of this comparison is to determine the cognitive plausibility of computational models like transformers. This involves understanding whether they merely achieve high, human-like performance due to genuinely assimilating the human cognitive process, or due to other artificial processes learned independently.

Previous work (Kozlova et al., 2024; Bensemann et al., 2022; Eberle et al., 2022; Morger et al., 2022; Wu et al., 2024; Hollenstein and Beinborn, 2021; Brandl and Hollenstein, 2022) has investigated this parallel using various techniques to extract attention scores from transformers and compare them to eye-movements from established eye-tracking datasets. However, several knowledge gaps exist. Most existing literature has focused on encoder transformer models, leaving open the question of whether more advanced and recent decoder models can equally align with eye-movements. Additionally, since many studies neglect the impact of low-level text properties on eye-movements, any correlation driven primarily by these surface-level features would be insufficient evidence of deeper cognitive alignment. Finally, eye-tracking datasets consist of multiple eye-tracking features that provide informative signals regarding reading patterns. However, these features are often intercorrelated, so they may capture redundant aspects of the same underlying attention mechanism. The literature has not been able to consolidate overlapping information from multiple features into a single analysis, where previous studies most often focus arbitrarily on a single metric.

To address the identified knowledge gaps, this

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 26–36, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-004

study compares the attention mechanism of a decoder-only model with human attention during reading. It employs both correlation analysis and predictive modeling using PCA-reduced representations of eye-tracking features and accounting for surface-level properties of the text. Moreover, the effect of different attention representation methods (raw attention, attention flow, and gradient-based saliency) is investigated on the results. The results provide insights into how the model architecture, attention method, and reading task collectively influence the similarity of model attention patterns to eye-movement behavior.

2 Background

2.1 Human attention and Eye-Movements

Eye-movement research has a long and successful history in studying human cognitive tasks. Eye-movements in reading are shown to provide information about cognitive language processing, like syntactic parsing and semantic integration (Frazier and Rayner, 1982), expectations about the text (Ehrlich and Rayner, 1981), and reading goals (Rayner, 2009).

Eye-movements consist of fixations and saccades (Rayner et al., 2006). Saccades are short, rapid movements to other parts of the text, while fixations occur when eyes remain stationary in between saccades (Reichle et al., 2003) and are considered the key point of information processing. Eye-tracking measures focus on different aggregations of these movements, such as the total fixated time on each word. Both low-level bottom-up processing and higher-level comprehension are reflected in eye-tracking measures. Familiar (Clifton Jr. et al., 2007), or frequently occurring words (Inhoff and Rayner, 1986) are subject to faster processing, whereas rare occurring words such as novel proper nouns tend to have longer fixations (Barrett and Hollenstein, 2020). Longer words also receive longer fixations, while shorter words are more likely to be skipped. Word length additionally interacts with the functional role of a word, where function words are fixated less than content words (Rayner, 2009). More top-down processes like assessing the predictability of the text given the preceding context draws shorter gaze durations and the reverse holds for unpredictable, surprising words (Ehrlich and Rayner, 1981).

2.2 The Rise of Transformer Models and the Quest for Interpretability

Transformer models were introduced when Vaswani et al. (2017) proposed attention as a novel method for handling contextual relations in language models. The attention mechanism compares different embedding representations of the sequence to determine the degree of relevance between each pair of words. The model then attends to important parts of the sequence depending on these relevance scores. The exceptional performance of these models and the intuition and transparency of the attention mechanism drew much attention from both deep learning and interpretability research. As transformer models are advancing, there is a growing demand to interpret not only their outputs, but also the internal mechanisms that lead to those outputs. The first advancements in interpretability emerged from the Computer Vision field (Simonyan et al., 2014; Zeiler and Fergus, 2014), where saliency maps were used to trace model decisions back to input pixels. This technique estimates the contributions of input raw data or intermediate activations to model predictions (Li et al., 2022) and is also commonly applied to NLP research. For transformer architectures specifically, attention-based and gradient-based methods have gained popularity for representing importance allocated to the input sequence.

2.2.1 The interpretability debate

Raw attention scores extracted directly from the model provide an easily understandable weighting of the input sequence. For example, in the original paper introducing attention, Vaswani et al. (2017) showed that examining the raw attention towards ambiguous pronouns like ”its” could reveal how anaphora resolution is represented in the model.

Previous work correlating raw attention and human eye-gaze shows mixed findings. While Sood et al. (2020) reported non-significant correlations for later layers of models like XLNet, Bensemann et al. (2022), Eberle et al. (2022) and Morger et al. (2022) found strong correlations in early Transformer layers. Bensemann et al. (2022) noted that correlation strength is generally higher in early layers and not dependent on the model’s size, though it can be influenced by the training process. Kozlova et al. (2024) similarly found strong early-layer correlations in the context of anaphora resolution. Eye-gaze features like First Fixation Duration (FFD) are

often found to align better with single-pass model behavior than cumulative measures like Fixation Count (F) or Total Reading Time (TRT), as FFD reflects initial processing (Ikhwantri et al., 2023). Furthermore, research has explored the ability of language models to predict human eye-movements as an indicator of their cognitive plausibility (Hollenstein et al., 2022).

However, the increased focus on faithful explanations opened a debate about the effectiveness of the attention mechanism as an explanation method. Some critics (Jain and Wallace, 2019) argue that raw attention weights do not always strongly correlate with gradient-based measures of feature importance, and that different attention distributions can lead to effectively identical model predictions (Jain and Wallace, 2019; Serrano and Smith, 2019), questioning whether attention provides a unique explanation for the model's behavior. However, Wiegreffe and Pinter (2019) argued that producing identical explanations to gradient-based methods is not necessary for plausible model explanations, especially when the goal shifts from explaining the model's predictions to broadly understanding the model's internal behavior (Bastings and Filippova, 2020). Nevertheless, attention flow and saliency-based methods have been proposed as more suitable for quantifying word importance in sentence processing.

Attention flow (Abnar and Zuidema, 2020) is an interpretation method based on flow networks from graph theory. This method tackles the problem of uniform raw attention in higher layers and models a global view of attention, as it captures the entire information propagation through the network layers. In attention flow, the raw attention graph is treated as a flow network that consists of nodes connected by directed edges. A flow function assigns values to edges such that the maximum total flow from a source node reaches a target node, under some capacity and conservation constraints.

Gradient-based saliency differs from attention-based methods as it does not utilize the transformer's attention mechanism[1]. Instead, it measures how sensitive the model output is to changes in each input token's embeddings. For each target token in the sequence, gradients are computed with respect to all input tokens and are normalized to produce a saliency score for each token.

Hollenstein and Beinborn (2021) found that fixation durations correlate better with saliency-based than with attention-based importance, suggesting saliency as a more cognitively plausible metric for interpretation. Morger et al. (2022) supported this finding for gradient-based saliency and for attention flow, across multiple languages. Similarly, Eberle et al. (2022) observed strong alignment of attention flow with human fixation times in natural reading, competitive with a specialized cognitive model of human reading (E-Z reader).

2.2.2 Alignment in task-specific contexts

Human reading strategies are task-dependent and influence how attention is allocated to different parts of the sequence. Task specificity thus plays a crucial role in the alignment between human and model attention. While Wu et al. (2024) found that finetuning models on task-specific objectives can enhance correlations with human gaze when using saliency methods, Eberle et al. (2022) showed that task-specific finetuning did not significantly increase correlation, and models aligned better with natural reading patterns than with task-specific ones. Brandl and Hollenstein (2022) further demonstrated that more in-depth reading (characterized by longer total reading times and lower skipping rates) generally correlates better with model attention compared to faster, shallow reading.

2.3 Our contribution

Despite evidence that transformer attention patterns align with human reading behavior, most existing work has focused on encoder-only or encoder-decoder architectures, leaving questions about newer decoder-only models that process text left-to-right (Hollenstein and Beinborn, 2021). Additionally, many studies overlook text properties' influence on eye-movements and lack methods for integrating multiple eye-tracking features in the analysis. While Wu et al. (2024) investigates an early decoder-only model (GPT-2), their analysis focused on gradient-based saliency in a task-specific setting. This study compares decoder-only model attention with human attention using raw attention, attention flow, and gradient-based saliency across both natural and task-specific reading. We address three research questions: **(RQ1)** To what extent do human eye-movements correlate with decoder model attention? **(RQ2)** Can decoder models predict eye-movements independently of text fea-

[1]Even though gradient-based saliency relies on input-output gradients rather than attention scores, it is loosely referred to as an *attention method* in this study for brevity, in the sense of attributing importance to the input.

tures like word frequency, length, and surprisal? **(RQ3)** How does Principal Component Analysis of eye-tracking features and task-specificity affect these correlations and predictions?

3 Methods

3.1 Eye-tracking Data

This study uses the Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018). ZuCo combines EEG and eye-tracking recordings from English native speakers reading natural sentences. 12 participants read sentences under different conditions (tasks). In Task 2 (”Normal Reading”) the participants were asked to read 300 sentences containing certain relations and answer a comprehension question after each sentence. In Task 3 (”Task-specific Reading”), the participants were instructed to focus on a specific relation type before reading the sentence. 407 sentences were presented in blocks of the same relation so the subjects knew what relation to look for. For each sentence, the participants had to indicate whether the specified relation was present in the sentence or not.

The raw eye-tracking data consists of the following eye-tracking features on the word-level: gaze duration (GD), total reading time (TRT), first fixation duration (FFD), single first duration (SFD), go-past time (GPT), fixation count (F) and mean pupil size (mPS). These features are normalized to their relative value in each sentence and then are averaged across participants to ensure robustness across different sentences and reading behaviors.

This study compares human and model data in two ways: (1) by analyzing each gaze feature individually, and (2) by combining the most informative aspects of these features using Principal Component Analysis (PCA). A PCA representation is derived separately for each task across all sentences within that task, with each word represented by its normalized eye-tracking values. We experimented with different numbers of components to identify the optimal balance between compact representation and captured variance. Ideally, a single component is preferred, as it encodes the most compact and efficient representation of the multidimensional eye-tracking data.

3.2 Model Attention

This study uses Llama 3.1-8B for investigating transformer attention, which is a latest generation decoder model, and BERT-base-uncased for the encoder comparison. Both BERT and Llama models remain in their pretrained states without task-specific fine-tuning because the goal is to investigate the fundamental model alignment with human attention, rather than deliberately optimizing it. For both models, explicit instructions similar to those the participants received are prepended to the input sentences to better resemble the original experiment and guide model attention closer to the human cognitive task. Attention patterns are extracted using standard forward passes without masking to reveal how each model’s attention mechanism responds to the same instructional context during inference.

3.2.1 Raw attention

The input is tokenized and passed through the model to obtain the raw attention scores for each layer, averaged across attention heads. Because tokenization can split original words into smaller subtokens, the attention scores are aligned with the human data by assigning each original word the maximum attention score among its subtokens, following the approach of Sood et al. (2020). The final score for each word in each layer is computed as the average attention it receives from all other words (including instruction words). A normalization by the sum of the total attention is applied to obtain relative attention scores per word and to account for variability across sentences.

3.2.2 Attention flow

Attention flow is calculated using Edmonds-Karp’s maximum-flow algorithm (Edmonds and Karp, 1972). The last token is considered the target ”sink” token in each sentence, where reading presumably ”stops”. For the Llama model, attention flow is implemented under a reduced number of paths to respect its causal attention structure. Finally, a decay is applied to account for the inherent bias to early input tokens in decoder models, using the position-based weighting proposed by Metzger et al. (2022).

3.2.3 Gradient-based saliency

Saliency is calculated by taking the L1 normalized gradient of the model’s output logit with respect to each input token embedding. This process is repeated with each token serving as the prediction target, and the resulting saliency scores are averaged across all targets to obtain a global saliency score for each token. As with previous methods, token-

level saliency scores are combined at the word level and normalized to reflect relative saliency within each sentence.

3.3 Analysis

The eye-tracking data (both PCA-reduced and individual features) are compared word by word with the transformer scores using Spearman's correlation. In addition to correlation analysis, we use linear regression models (ordinary least squares) to assess whether there is a predictive relationship between model attention and eye-gaze, and to measure how additional text features linked to eye-movements may influence this relationship. The predictive relationship is assessed through adjusted R^2 on unseen data (20% of the dataset). We incorporate 5 text-related features as additional predictors alongside model attention: word frequency (across many corpora), word length, functional category (function vs. content words) and surprisal derived from the respective transformer model.

Four regression models are fitted for each combination of transformer model (BERT, Llama), attention method (raw attention, attention flow, gradient-based saliency) and reading task (Task 2, Task 3). To determine whether model attention improves predictive capacity, we compare performance to a baseline model that uses only text features as independent variables. We similarly compare the PCA model to the average performance of models predicting individual eye-gaze features, with PCA predictions transformed back to the original feature space. Both comparisons use the Wilcoxon signed-rank test of mean squared errors between regression models. To further analyze the contribution of attention and text features to the regression models, we visualize feature importance using absolute t-values[2].

4 Results

4.1 Experiment 1: Replication

The correlation analysis between BERT raw attention and human eye-movements successfully replicates previous findings. BERT's first layer shows the highest correlations with Task 2 eye-movements ($\mu = 0.69, \sigma = 0.02$), as indicated by the blue line in Figure 1, which is consistent with results from Morger et al. (2022). Correlation

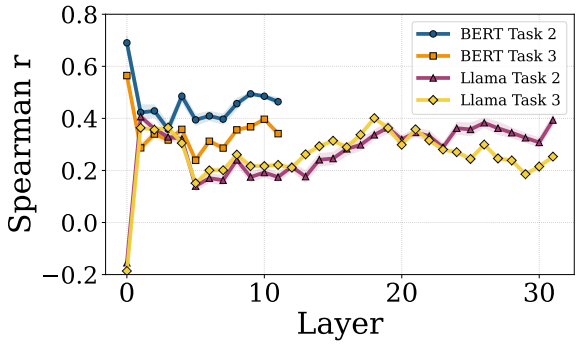

Figure 1: Average layer-wise correlations of each transformer's raw attention across 6 eye-tracking features.

strength generally decreases across subsequent layers. A similar trend is observed in Task 3, where the first layer again exhibits the strongest correlation ($\mu = 0.56, \sigma = 0.008$), as shown by the orange line.

The results with the alternative attention methods also replicate previous findings. Attention flow shows the strongest correlations with human eye-movements. For Task 2, the correlation is $\mu = 0.74, \sigma = 0.007$ (blue plain bar, Figure 2), while for Task 3 $\mu = 0.62, \sigma = 0.007$ (orange plain bar). These results align with Eberle et al. (2022), who found that attention flow produces better alignment with human eye-movements than the strongest correlating layer of raw attention.

Gradient-based saliency performs less strongly across both tasks. Task 2 correlations reach $\mu = 0.68, \sigma = 0.02$ (blue hatched bar, Figure 2), matching the score reported by Hollenstein and Beinborn (2021). Task 3 correlations are $\mu = 0.53, \sigma = 0.005$ (orange hatched bar), similar to results from Wu et al. (2024). This makes saliency the least correlating attention method with human eye-movements for the BERT model. For all attention methods, Task 3 correlations mirror Task 2 patterns but at reduced magnitudes, consistent with previous work (Eberle et al., 2022). All reported correlations are statistically significant ($\alpha < 0.05$).

4.2 Experiment 2: Extension

4.2.1 What about Llama?

Correlations of Llama's raw attention with human eye-movements fall visibly lower than BERT correlations. Llama's first layer shows almost negative correlations with Task 2 eye-movements, while the second layer is the one with the highest correlations ($\mu = 0.4, \sigma = 0.009$), as shown by the purple

[2]Code is available in Github: https://github.com/maria-mouratidi/thesis

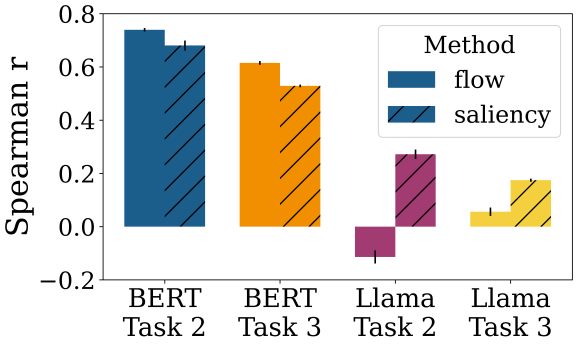

Figure 2: Average correlations across 6 eye-tracking features with each transformer's *attention flow* and *gradient-based saliency*.

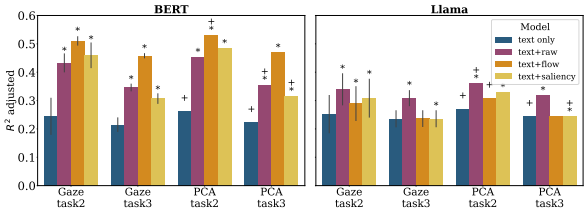

Figure 3: R^2 adjusted scores of the regression models, with or without attention as a feature. On the x-axis are the prediction targets and task conditions of each model. For the Gaze models, performance is averaged over each eye-tracking target. Asterisks indicate significant improvement over baseline, while crosses indicate significant improvement of PCA over the Gaze variant.

line in Figure 1. Like BERT, correlations decrease in subsequent layers, though with some upward trend toward the final layers. Task 3 correlations in Llama remain close to Task 2 correlations, and even exceed them in some layers (yellow line, Figure 1) showing that task-specific patterns are not affecting Llama as much as BERT.

Similarly to raw attention, attention flow and gradient-based saliency produce more moderate correlations with Task 2 and Task 3 eye-movements compared to BERT, with saliency ($r = 0.27, r = 0.18$) showing a clear advantage over flow ($r = -0.11, r = 0.05$) (purple and yellow bars, Figure 2). All correlations for the Llama model are statistically significant ($\alpha < 0.05$).

4.2.2 Predicting eye-movements

The regression models demonstrate clear benefits from incorporating transformer attention as a feature. For BERT (left panel, Figure 3), all attention methods significantly outperform the text-only baseline (blue bars). Moreover, attention flow contributes to the highest model fit for all tasks and DV conditions, reaching an overall $R^2 \approx 0.5$ (orange bars). This result is predictable from the higher correlations of attention flow with the eye-tracking features in Figure 2.

Llama (right panel, Figure 3) shows more modest performance than BERT. The models incorporating raw attention achieve R^2 values between 0.3 and 0.4 and both raw attention and saliency outperform the baseline across all conditions. As expected, attention flow does not significantly improve predictive capacity, except for the Task 2 Gaze condition. A clear pattern that emerges for both models is that Task 2 eye-movements are more predictable than Task 3 from all attention methods.

4.2.3 Reducing eye-tracking features

During the PCA exploration phase, we noticed that the SFD and GPT features accounted for most of the variance in a potential second PCA component. This is likely due to the nature of these features: SFD becomes zero when words receive multiple fixations, so it is often sparse, and GPT is likely more noisy as an intermediate feature between immediate processing (like FFD) and overall processing measures (like TRT). To maintain model simplicity, we removed SFD and GPT from the analysis entirely[3]. This reduction resulted in 94% explained variance in the first PCA component for Task 2 and 97% for Task 3. The simplified approach allows us to retain only one PCA component per task while preserving the most informative gaze patterns.

The cross annotations in Figure 3 demonstrate that all text-only PCA baselines show statistically significant improvement over their corresponding Gaze variants. This pattern extends to models using attention as well, where more than half outperform their Gaze counterparts in both BERT and Llama and both tasks. For the remaining PCA models that do not reach significance, the performance differences are minimal. This finding indicates that a single-PCA representation of the most important eye-tracking features can successfully replace the training procedure of multiple Gaze models.

4.2.4 Attention's role in prediction

To gain some perspective of the contribution of attention methods to predicting eye-movements, we examine the average feature importances over all

[3]This makes a total of 5 eye-tracking features included in the analysis. Whenever 6 features are mentioned, it means that the PCA component is also considered as a feature.

linear regression models. As seen in the upper panels of Figure 4, all attention methods for the BERT model receive the highest significance compared to the other text features. Attention flow demonstrates the greatest difference from other features, with only length and surprisal showing significant contributions. Llama models present a different

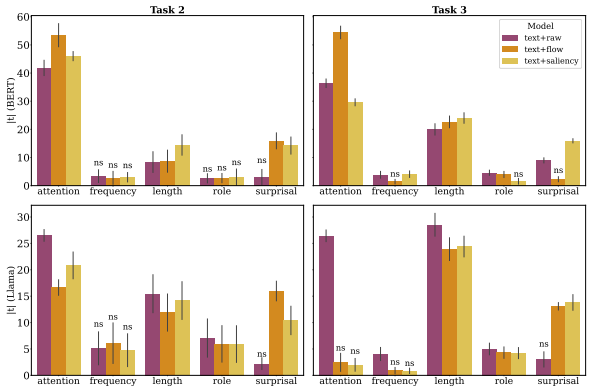

Figure 4: Feature importances based on mean absolute t-value across models predicting 6 eye-tracking features. "Ns" signifies non significant t-values ($p > 0.05$)

pattern, in the lower panels of Figure 4. Raw attention shows greater contribution than other attention methods but competes closely with word length in Task 3. Other attention methods maintain high contributions in Task 2 but become insignificant in Task 3. For non-raw attention methods in Llama, word length becomes a large contributor, followed by surprisal.

To further explore relationships between features that may influence their respective relative importance in the predictions, we examine correlations between attention methods and text features in Figure 5. BERT shows attention patterns that align with established eye-movement research. Higher frequency words receive smaller BERT attention values, while longer, content words draw more attention than short, function words. Notably, surprisal appears only slightly represented in BERT's attention mechanism. When it comes to Llama, similar correlation directions appear for raw attention, but in smaller magnitudes. Here, model attention using any method is more correlated to surprisal than any other text feature.

5 Discussion

We return to our research questions to briefly reiterate the key findings: Regarding correlation strength **(RQ1)**, decoder-only models like Llama show

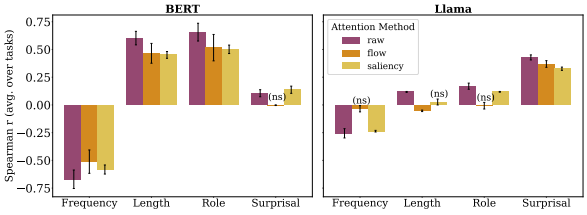

Figure 5: Correlations of text features with each attention method on a word-level.

medium-strength correlations with human eye-movements when using raw attention or gradient-based saliency. For **RQ2**, regression models combining Llama raw attention or gradient-based saliency with other text features achieve moderate performance in predicting human reading behavior. The regression models' predictive success relies heavily on attention, followed by word length and surprisal features. Concerning feature reduction and task effects **(RQ3)**, a single PCA component successfully replaces individual gaze targets while maintaining equivalent, or more favorable alignment results. Across all attention methods, both BERT and Llama show stronger alignment with Task 2 eye-movements than Task 3, suggesting that task-specific departures from normal reading are not equally well encoded in pretrained model attention mechanisms. The following sections provide deeper discussion of these findings and their implications.

5.1 What do layer-wise correlations imply?

The layer-wise raw attention correlations may hint at the type of linguistic processing that is most similar between humans and models. Early layers typically process surface-level features before higher-level semantic integration occurs, in both encoder (Tenney et al., 2019) and decoder models (Vig and Belinkov, 2019). Thus, the fact that higher correlations occur in early layers (first layer for BERT and second for Llama) suggests that better alignment can be found in bottom-up processes. This finding also has theoretical grounding in eye-movement research. The oculomotor control system can guide saccades before full lexical identification occurs (Rayner et al., 2011), and other early processing features like word length and frequency (Inhoff and Rayner, 1986) have robust independent effects on word skipping and fixation durations.

However, the correlation patterns are not monotonic across layers. Later layers show stronger correlations than middle layers. This suggests that

final layers may be more similar to the higher-level processing that also influences gaze behavior, for example when expectations about the text are formed using contextual information (Ehrlich and Rayner, 1981), or when the reader is more engaged in next-word prediction (Goldstein et al., 2022).

5.2 Why is Llama falling behind?

We consider two primary explanations for the alignment gap between the encoder and decoder model.

5.2.1 Pre-training objectives

First, Llama is a generative model optimized for next-token prediction, while BERT is trained in masked language modeling to capture bidirectional contextual representations. This fundamental difference in training objectives may explain why BERT aligns more with human reading behavior, which primarily involves comprehension rather than generation. Although the brain engages in next-word prediction during reading (Goldstein et al., 2022), the autoregressive nature of decoder models may not fully capture the integrative parts of human language comprehension that involve both forward and backward contextual dependencies. This difference is empirically supported by our feature correlation analysis, where Llama attention correlates most strongly with surprisal (a prediction-based feature) while BERT attention correlates mostly with the other text features (Figure 5).

5.2.2 The role of model size

This study's comparison between BERT-base (110M parameters) and Llama 3.1-8B (8B parameters) confounds architecture type with model scale. The substantial size difference may contribute to the observed alignment gaps, as larger models can distribute attention-relevant information across more parameters and layers. The specific model variants were chosen because of 1) BERT-base-uncased for replication and validation of previous work and 2) Llama 3.1-8B for the best performance-efficiency tradeoff among available decoder models. Nevertheless, future work should compare models of similar sizes to isolate architectural effects from scale effects.

5.3 Why is alignment with task-specific reading more difficult?

Task 3 eye-movements occur under fundamentally different reading conditions than Task 2, leading to smaller alignment with pretrained transformers. In the task-specific condition, readers form expectations about which words or syntactic structures might signal the specified relation, leading to more selective attention distribution among words (Hollenstein et al., 2020). When searching for specific words, reading resembles visual search, and certain text-level influences on eye-movements like word frequency disappear. (Rayner, 2009). Even when receiving explicit instructions, the models cannot replicate this type of selective attention. This limitation appears to be fundamental to current pretraining techniques rather than architecture-specific, as both encoder (BERT) and decoder (Llama) models show consistently better alignment with natural reading than task-specific reading. This suggests that neither masked language modeling nor autoregressive prediction objectives adequately prepare models for goal-directed attention strategies.

5.4 To each their own attention method

The results of different attention extraction methods vary significantly between the two models. Attention flow aligns best with eye-movements for BERT, while raw attention performs better for Llama. This may relate to the original motivation for attention flow, which was proposed as a way to represent attention in encoder models (Abnar and Zuidema, 2020). In decoder models, attention is restricted to preceding tokens, leading to an early token bias. When this effect is normalized as recommended by Abnar and Zuidema (2020), the attention signal may become more diluted. This highlights the need to carefully match the attention explanation method to both the model architecture and the explanation task.

5.5 One dimension for all eye-tracking features

Models using the PCA representation match or outperform Gaze models in correlation and regression analyses. This approach serves primarily as a methodological efficiency tool rather than aiming to increase predictive power. By capturing the shared variance across multiple eye-tracking features in a single component, PCA removes redundancy inherent in correlated features while preserving the essential reading patterns. So we were able to remove this practical challenge of determining which of the many available features are suitable for the comparison, without distorting them using averaging techniques. When PCA models

show similar alignment patterns to individual feature models, this suggests that much of the variance relevant to the comparison is captured by a common underlying dimension of reading behavior. This has implications for future eye-tracking studies, where researchers may be able to focus their analysis on this common dimension rather than examining all traditional eye-tracking measures individually, when the goal is understanding attention alignment with computational models.

6 Conclusion

This study examined the alignment between decoder-only models and human attention during reading. Overall, eye-movement data correlated with and was predictable from transformer attention, suggesting partial model alignment with human language processing. Early layers showed stronger alignment with eye-movements, hinting that bottom-up processes are more consistent with human reading behavior. However, lower alignment with task-specific reading suggests these pretrained models lack human-like flexibility to adapt attention based on task goals. Despite these shared patterns across architectures, the decoder model underperformed compared to the encoder model, showing lower correlations, weaker predictive power, and different patterns of feature prioritization, likely due to architectural differences. Finally, different methods of representing transformer attention significantly impact alignment comparisons, which emphasizes the importance of well-motivated, model and task-specific choices in explaining transformer mechanisms.

6.1 Limitations

This work assumes eye-movements provide sufficient information about cognitive language processing, though eye-tracking misses covert cognitive mechanisms and information processed outside the fixation region (Rayner, 2009; Reingold et al., 2016). Additionally, different attention explanation methods produce variable results, creating uncertainty about their faithfulness in explaining transformer attention mechanisms. Ultimately, any attention method is only a proxy to the true model representation.

Acknowledgments

We gratefully acknowledge the John S. Latsis Public Benefit Foundation for their generous support through the Postgraduate Scholarship Program.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. Association for Computational Linguistics.

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

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

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87. 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, pages 72–77. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Charles Clifton Jr., Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, editors, *Eye Movements: A Window on Mind and Brain*, pages 341–371. Elsevier.

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, pages 4295–4309. Association for Computational Linguistics.

Jack Edmonds and Richard M. Karp. 1972. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of Association for Computing Machinery*, 19(2):248–264.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

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

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

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.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Fariz Ikhwantri, Jan Wira Gotama Putra, Hiroaki Yamada, and Takenobu Tokunaga. 2023. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing & Management*, 60(2):103195.

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

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, pages 3543–3556. 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.

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. Association for Computational Linguistics.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234.

Niklas Metzger, Christopher Hahn, Julian Siber, Frederik Schmitt, and Bernd Finkbeiner. 2022. Attention flows for general transformers.

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.

Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

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

Keith Rayner, Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of experimental psychology. Human perception and performance*, 37(2):514–528.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The e-z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476.

Erik D. Reichle, Andrew E. Reineberg, and Jonathan W. Schooler. 2010. Eye movements during mindless reading. *Psychological Science*, 21(9):1300–1310.

Eyal M. Reingold, Heather Sheridan, K. L. Meadmore, Denis Drieghe, and S. P. Liversedge. 2016. Attention and eye-movement control in reading: The selective reading paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, 42(12):2003–2020.

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. Association for Computational Linguistics.

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

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. 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. Association for Computational Linguistics.

Rosy Southwell, Julie Gregg, Robert Bixler, and Sidney K. D'Mello. 2020. What eye movements reveal about later comprehension of long connected texts. *Cognitive Science*, 44(10).

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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Kaiser. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NIPS'17*, pages 6000—6010. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76. Association for Computational Linguistics.

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

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Guojun Wu, Lena Bolliger, David Reich, and Lena Jäger. 2024. An eye opener regarding task-based text gradient saliency. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 255–263. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing.