From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models

Luca Dini[1,2], Lucia Domenichelli[1,2] Dominique Brunato[1], Felice Dell'Orletta[1],
[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa
[2]University of Pisa
{name.surname}@ilc.cnr.it

Abstract

Cognitive signals, particularly eye-tracking data, offer valuable insights into human language processing. Leveraging eye-gaze data from the Ghent Eye-Tracking Corpus, we conducted a series of experiments to examine how integrating knowledge of human reading behavior impacts Neural Language Models (NLMs) across multiple dimensions: task performance, attention mechanisms, and the geometry of their embedding space. We explored several fine-tuning methodologies to inject eye-tracking features into the models. Our results reveal that incorporating these features does not degrade downstream task performance, enhances alignment between model attention and human attention patterns, and compresses the geometry of the embedding space [1].

1 Introduction and Motivations

Understanding the inner workings of Neural Language Models (NLMs) remains a fundamental challenge in NLP. While these models achieve remarkable performance across a variety of tasks (Hendrycks et al., 2021; Bubeck et al., 2023), their decision-making processes are largely opaque due to their complex architectures and increasingly large number of parameters. To address these challenges, cognitively informed approaches that draw on insights from human language acquisition and processing have gained increasing attention in both the training and evaluation stages (Ettinger, 2020; Evanson et al., 2023). These approaches not only offer potential strategies for enhancing model interpretability but also provide frameworks for developing efficient learning methods, which are particularly useful in small-scale settings where data and computational resources are limited (Huebner et al., 2021; Warstadt et al., 2023).

In this scenario, eye-tracking (ET) data, which reflect human gaze patterns during reading, offers a potential bridge between human cognition and model behavior. Gaze has long been widely considered a rich source of cognitive information, as it reflects both early and late stages of text processing (Just and Carpenter, 1980; Rayner, 1998). This makes ET an important tool for formulating theoretical accounts on a variety of linguistic phenomena, from sentence complexity (Staub, 2010), to syntactic ambiguity resolution (Frazier and Rayner, 1982) and reading proficiency (Ashby et al., 2005).

Given their informativeness, a line of research in NLP has focused on leveraging these signals to **improve model's performance** on multiple downstream tasks, including core linguistic tasks like part-of-speech tagging and dependency parsing

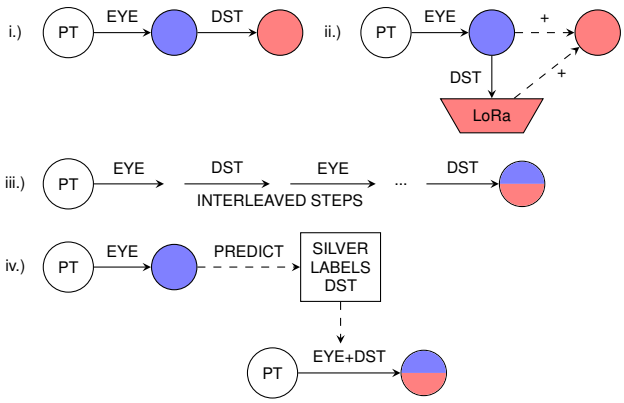

Figure 1: Scheme of the eye-tracking injection strategies. i.) is the intermediate fine-tuning; ii.) is the fine-tuning using LoRa adapters for the downstream task; iii.) is the multitask fine-tuning with interleaved steps (one on eye-tracking and one in downstream task); iv.) is the multitask fine-tuning with eye-tracking silver labels on the downstream task dataset. In the diagram, nodes correspond to models and edges represent fine-tuning processes, with EYE and DST indicating eye-tracking data or downstream task fine-tuning, respectively. PT stands for the pre-trained model. Blue colored nodes indicate that the model is specialized on the prediction of eye-tracking features, while red colored nodes indicate the specialization on the downstream task.

[1]Code to reproduce our experiments is available on Github.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17796–17813
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

(Barrett et al., 2016), as well as more applied tasks such as sentiment analysis and sarcasm detection (Mishra et al., 2016; Tiwari et al., 2023), question answering (Malmaud et al., 2020; Zhang and Hollenstein, 2024), text readability (Singh et al., 2016). More recently, interest has shifted toward also using ET data for **interpretability** purposes, with the goal of clarifying the potential connection between language learning and processing in both humans and models, thus providing a transparent and cognitively grounded framework for analyzing model performance and representational capacity. This includes exploring parallels between human gaze patterns and model attention mechanisms (Wang et al., 2024a; Morger et al., 2022; Sood et al., 2020; Wang et al., 2024b), as well as investigating the effectiveness of model perplexity related metrics to predict reading estimates (Oh and Schuler, 2023a,b; Hao et al., 2020).

This work aims to advance this line of research by conducting a comprehensive investigation of NLMs informed by data representative of human reading behavior. We explore the integration of ET signals into the model by designing and testing various strategies for injecting gaze-related features and evaluate them on an encoder-based model, i.e. RoBERTa-base (Liu et al., 2019). Despite the growing prominence of large language models (LLMs), we focus on encoder-only transformers, as they remain the foundation of many real-world applications (Tunstall et al., 2022; Zaratiana et al., 2024). Encoders are particularly well-suited for scenarios where computational resources are limited, making LLMs impractical for deployment in every task (Karpukhin et al., 2020; Lewis et al., 2020). Accordingly, interpretability approaches are especially crucial for understanding their behavior and increasing transparency.

Our goal is to analyze the impact of ET data on these models across multiple dimensions, which, to the best of our knowledge, have never been assessed together. Specifically, we focus on these main dimensions and research questions:

i.) **Effect on downstream task performance**: Does injecting ET features into the model affect its performance on a downstream task, and how do task-specific characteristics influence this effect?

ii.) **Impact on model attention patterns**: What is the impact of injecting ET features on the model's attention mechanism? Additionally, how does the attention mechanism change when the model is also trained to solve downstream tasks?

iii.) **Changes in representation space**: How does the ET injection affect the geometric properties of the model's representation space?

For all aspects, we examined the impact of different approaches to inject ET features into the model to determine whether a particular approach consistently yields better results than others and whether its effectiveness is consistent across all dimensions.

2 Related Work

Eye-tracking data in NLMs In recent years, a growing line of research in NLP has focused on using physiological signals to study the properties of language models and enhance their performance. Among these signals, ET data recorded during reading has gained particular prominence, largely due to its relative ease of collection using non-invasive equipment compared to other physiological measurement techniques such as fMRI. The availability of many publicly accessible ET datasets (Cop et al., 2017; Siegelman et al., 2022; Raymond et al., 2023) has further facilitated advancements in this field.

Research leveraging ET data has followed two main directions. One focuses on improving performance on downstream tasks by augmenting models with gaze-related features (Hollenstein et al., 2019). The other, which is more directly relevant to our study, investigates the intersection between ET data and the internal mechanisms of NLMs, offering insights into both model interpretability and the cognitive plausibility of language processing in artificial systems (Beinborn and Hollenstein, 2023). In this context, Sood et al. (2020) were among the first to compare the attention mechanisms of neural models based on different architectures with ET data. Their findings revealed that, although transformers achieved the highest task performance, they exhibited significantly lower correlation with human gaze patterns compared to other architectures. Similarly focusing on task-oriented reading, Eberle et al. (2022) explored whether learned self-attention functions in large transformers correlated to eye fixation patterns across two task-specific reading datasets for sentiment analysis and relation extraction. They observed that task-specific fine-tuning does not increase the correlation with human reading. Contrasting with these findings, Bensemann et al. (2022) demonstrated that gaze dwell times are closely aligned with the early layers of pre-trained transformers like BERT, with this correlation remaining consistent regardless of the

model's parameter count. Likewise, Morger et al. (2022), in their cross-lingual study on human and model-derived word importance metrics, found robust correlations, particularly for monolingual models. Wang et al. (2024b) extended the experiments on the correlation between self-attention head values from LLMs and eye-tracking measures, finding similarities on the deeper layers of the model.

Embedding Space Studies The embedding space is a learned high-dimensional vector space where discrete tokens are mapped to dense representations capturing semantic and syntactic information. When a word is input to the model, it is first mapped to its embedding, which serves as its internal representation for further processing. A substantial body of research (Ethayarajh, 2019; Godey et al., 2023) has demonstrated that the embedding spaces induced by Transformer models often exhibit *anisotropy*. In other words, the embeddings do not fully occupy all available dimensions but cluster closely together, a phenomenon referred to as the *representation degradation problem* (Gao et al., 2019). In NLP, anisotropy is often seen as detrimental as it confines embeddings to a "narrow cone" in the embedding space, thereby masking linguistic information and diminishing expressive power (Cai et al., 2021; Zhang et al., 2020; Mickus et al., 2020). However, broader machine learning literature has noted that anisotropy naturally arises from stochastic gradient descent and may, in fact, enhance generalization. In this respect, studying model behavior in downstream tasks provides valuable insights (Diehl Martinez et al., 2024; Rudman and Eickhoff, 2024; Machina and Mercer, 2024). Specifically, models that compress data onto lower-dimensional manifolds often achieve superior performance in downstream tasks (Ansuini et al., 2019). In this work we extend the analysis of anisotropy to sentence embeddings.

3 Our Approach

We develop a comprehensive framework to investigate the impact of incorporating human reading behavior into a NLM across three key dimensions: **performance on downstream tasks**, modifications in **attention mechanisms**, and shifts in **embedding space**. To incorporate reading-related information into the model, we leverage a set of ET features extracted from the dataset described in 3.1. However, unlike most existing literature—which typically aggregate data across participants, with very few exceptions (Brandl and Hollenstein, 2022)— we analyze each reader separately, conducting experiments independently for each one. This choice is motivated by the great variability in reading behaviors among readers, even highly skilled ones (Parker and Slattery, 2021; Ashby et al., 2005; Slattery and Yates, 2018), and allows us to better model reader-specific dynamics.

One key feature of our work is the design and evaluation of multiple injection strategies for incorporating ET data into the model. These strategies, described in Section 3.3, include established techniques such as intermediate fine-tuning, in which the task of predicting ET features is performed before training the model on the target task, as well as novel ones proposed in this study, including two multitask fine-tuning approaches. While studies such as Weller et al. (2022) have shown that different transfer learning strategies can yield varying results depending on factors such as the relative size of the target and supporting tasks, to the best of our knowledge, no prior work has conducted an in-depth comparison of fine-tuning strategies using ET features prediction as a supporting task while also examining their impact on the model's inner mechanisms beyond task performance.

An additional peculiarity of our approach is the development of distinct evaluation strategies for each dimension under analysis. Specifically, for downstream task performance, we assess models using task-specific evaluation metrics. To examine attention mechanisms, we compute the correlation between human gaze patterns and model attention distributions. Lastly, to analyze the representation space, we investigate changes in *isotropy* and *linear effective dimensionality*, providing insights into how fine-tuning with ET signals reshapes the embedding space. Section 3.3 reports details on the implementation of these strategies.

To comprehensively evaluate the effect of ET injection strategies on each dimension, we compare the models fine-tuned with both ET and downstream task data, against i.) the pre-trained model; ii.) models fine-tuned only on ET data, and iii.) models fine-tuned solely on downstream tasks. Notably, to gain a deeper understanding of the effects of incorporating ET knowledge into the model, we evaluate these models across a range of downstream tasks, which were chosen as they provide insights into various aspects of language competence. In what follows, we describe the key components and methodological choices of our research.

3.1 Data

Eye-tracking dataset We leveraged ET data from the English section of the GECO corpus (Cop et al., 2017), which contains metrics from 14 participants (users) reading the novel *The Mysterious Affair at Styles* by Agatha Christie. The dataset consists of 56,410 words (588 sentences) read by all participants. However, two participants did not complete the entire novel and were excluded from our analysis. For each participant we extracted five features from GECO, corresponding to word-level reading time measures and serving as proxies for different stages of reading behavior: First Fixation Duration, the duration of the first fixation landing on the word; Gaze Duration, the summed duration of fixations on the word in the first pass reading; First-run Number of Fixations, the number of fixations on a word during the first pass; Total Reading Time, the summed duration of all fixations on the current word, including regressions; Total Number of Fixations, number of fixations on a word overall. To create a test dataset, we randomly selected 20% of sentences, while the remaining 80% were used for training. The split is consistent across all users. The features were scaled to the range $[0, 100]$ to balance the training process and facilitate the interpretability of the results.

Downstream task datasets As regards the downstream tasks on which models were tested, we chose the following dataset: i.) *Human Complexity Judgment* (COMP), which involves predicting the complexity score assigned by human annotators to a given sentence. For this purpose, we used the English section of the dataset described in Brunato et al. (2018). This task was chosen due to its close relationship with ET data, as it offers a complementary perspective— i.e. an *offline* perception of sentence complexity that may reflect cognitive processing during reading. ii.) The *GLUE benchmark* (Wang et al., 2018), a standard evaluation suite designed to test model performance across 9 natural language understanding tasks covering: Corpus of Linguistic Acceptability (*CoLa*), Stanford Sentiment Treebank (*SST-2*), Multi-Genre Natural Language Inference (*MNLI*), Question Natural Language Inference (*QNLI*), Recognizing Textual Entailment (*RTE*), Winograd Natural Language Inference (*WNLI*), Quora Question Pairs (*QQP*), Microsoft Research Paraphrase Corpus (*MRPC*), and Semantic Textual Similarity Benchmark (*STS-B*).

3.2 Models and Implementation

For all experiments, we used RoBERTa-base, an encoder-only transformer model with 12 hidden layers and an embedding size of 768. Fine-tuning on ET data was framed as a multi-label token-level regression task, where the model predicts the five ET features simultaneously for each token in a sentence. Since ET features are recorded at the word level, and the model tokenizes words into sub-tokens, we associated the features only with the first sub-token of a word, following the approach in (Hollenstein et al., 2021). In contrast, all downstream tasks were formulated as sentence-level classification or regression tasks, based on their output prediction. All training details are reported in Appendix A.

3.3 Eye-tracking injection strategies

To integrate knowledge about human reading behavior and train the model on downstream tasks, we designed several injection strategies, based on different fine-tuning techniques. We outline these strategies below. A visual summary is provided in Figure 1. i.) **Intermediate fine-tuning**: This strategy involves a fine-tuning on the ET dataset, followed by a fine-tuning on the downstream task. For the final fine-tuning on the downstream task, we fine-tuned the whole model (INT-FULL) or just the last layers: only the classification/regression head (INT-CLF), the classification/regression head and the last hidden layer (INT-LAST2) or classification/regression head and the last two hidden layers (INT-LAST3). ii.) **Fine-tuning with Adapters** (LORA): Similarly to the intermediate fine-tuning, we first finetune the model on the prediction of ET labels. Then we finetune the model on the downstream task using Low-Rank Adaptation (LoRA) (Hu et al., 2022), and a classification/regression head. Lastly, we add the obtained adapters to the first model and substitute the classification/regression head. iii.) **Multi-task interleaved fine-tuning** (MT-IL): In this method, the model alternates between two tasks during fine-tuning: one step on the ET dataset and the next on the downstream task dataset. To address the imbalance in dataset sizes—where the ET dataset is significantly smaller—we repeated the ET dataset multiple times to match the size of the downstream task dataset. iv.) **Multi-task fine-tuning with et silver labels on downstream task dataset** (MT-SILV): In this method, we first fine-tuned a model

on predicting ET features. We then used this model to generate silver labels (predicted ET features) for the sentences in the downstream task dataset. A new model was subsequently fine-tuned using silver labels as additional features alongside the downstream task resolution. We tested this strategy to observe what happens when multi-task fine-tuning is performed on the same data as the downstream task, rather than on two different datasets.

3.4 Evaluation Strategies

Performance on Downstream Tasks We evaluated the model's performance on the downstream tasks according to their corresponding official metrics. To improve the comprehensibility of the results, when a task could be evaluated using multiple metrics, we chose to report the most global or informative one. Specifically, we used the Spearman correlation coefficient for *COMP*, Matthews correlation coefficient for *CoLA*, accuracy for *MNLI*, *QNLI*, *RTE*, *WNLI*, and *SST-2*, a combined accuracy/F1 score for *MRPC* and *QQP*, and a combined Spearman/Pearson correlation score for *STS-B*. All these metrics yield scores in the range $[0, 1]$, where values closer to 1 indicate better performance.

Attention Patterns To evaluate whether injecting ET data modifies the model's attention mechanisms—potentially aligning them more closely with human attention—we calculated the Spearman correlation coefficient between human and model attentions across the model's layers. Human attention was approximated using ET features extracted from the dataset, while model attention was represented by the weights of the attention matrix, which indicate the distribution of focus across tokens during representation computation. For each sentence, model attention was calculated as the attention weight of each word contributing to the representation of the BOS token (<s> in the model used). To ensure consistency with the training process, when a token was split into sub-tokens by the tokenizer, only the first sub-token was used as a representative of the entire word. Moreover, to focus on the strength of the association we used the absolute values of the correlation coefficients.

This analysis was computed on GECO test set.

Embeddings Space To investigate the geometry of embedding space, we employ two scores: i.) **Linear effective dimensionality** (Lee et al., 2024), defined as the dimension of the minimal linear subspace that contains the embeddings. Note that linear approaches like PCA might miss non-linear structures, so our effective dimensionality analysis is only an initial step in understanding embedding geometry. ii.) **IsoScore*** (Rudman et al., 2022), which formalizes isotropy into a continuous score $\in$ [0,1], with higher values indicating more uniform usage of the embedding space. We opt for IsoScore* because, as the authors demonstrate, other existing isotropy metrics lack a rigorous mathematical foundation. We extract sentence embeddings via mean pooling from the GECO corpus and the English Universal Dependencies treebanks (de Marneffe et al., 2021), specifically using sentences from English-EWT (Silveira et al., 2014) matched in length to GECO, and compute scores for every model layer. Despite limited data, potential rank deficiency, and unstable singular values, Figure 2 in Appendix C shows that IsoScore* remains stable and effectively distinguishes isotropic from anisotropic point clouds.

4 Experimental Results

This section presents our experimental results. Due to space constraints, user-specific results are available on our GitHub page.

4.1 Impact on Downstream Task Performance

We first report the performance results of the ET injected models on the resolution of the downstream tasks, averaged across all users, and in comparison to models fine-tuned only on the downstream task. For reference, Appendix B reports the detailed performance of models fine-tuned exclusively on ET feature prediction.

As shown in Table 1, results suggest that **intermediate fine-tuning followed by full fine-tuning on the downstream task (INT-FULL) effectively preserves task performance**, with only minor exceptions. This is particularly interesting, as it is well established that sequential fine-tuning on multiple tasks typically leads to performance degradation on the final task due to *catastrophic forgetting*—a phenomenon where the model loses pretraining-acquired capabilities that are not directly relevant for the first fine-tuning task but could be crucial for the second (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017). While it is reasonable to expect that this issue would be mitigated in the *COMP* task—given that ET data provides a complementary perspective to the conscious judgment of sentence complexity—the fact that this

	Downstream Task										
Fine-tuning	**COLA**	**COMP**	**MNLI M/MM**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	0.56	0.90	0.88 / 0.88	0.90	0.93	0.90	0.70	0.92	0.91	0.56	0.82
INT-LAST3	0.25	0.88	0.70 / 0.71	0.80	0.82	0.81	0.54	0.88	0.81	0.56	0.71
INT-LAST2	0.15	0.85	0.62 / 0.64	0.77	0.75	0.77	0.53	0.86	0.74	0.56	0.66
INT-CLF	0.00	0.70	0.43 / 0.44	0.75	0.61	0.61	0.50	0.76	0.12	0.56	0.50
LORA	0.41	0.87	0.85 / 0.85	0.80	0.91	0.86	0.49	0.93	0.88	0.55	0.76
MT-IL	0.53	0.91	0.83 / 0.83	0.90	0.92	0.88	0.75	0.93	0.90	0.52	0.81
MT-SILV	0.51	0.91	0.88 / 0.87	0.88	0.93	0.90	0.60	0.93	0.91	0.50	0.76
DST-ONLY	0.60	0.91	0.88 / 0.88	0.90	0.93	0.90	0.77	0.93	0.90	0.56	0.83

Table 1: Performance on downstream tasks. Each row corresponds to a model fine-tuned using a distinct ET injection strategy. For comparison, DST-ONLY is the model fine-tuned exclusively on the downstream task. Highlighted cells in the table indicate a minimal performance drop (at most 0.02 points) compared to DST-ONLY. Note that *MNLI* is released with two test sets: Matched (*M*), with data coming from the same distribution as the training set, and MisMatched (*MM*), with out-of-domain test data.

pattern holds across nearly all tasks suggests that predicting ET features draws upon a broad set of capabilities acquired during pretraining. However, **when the fine-tuning on downstream task is partial or employs LoRa adapters, performance degradation is observed for almost all tasks**. This indicates that training only a small subset of parameters is insufficient for our model to learn and solve the downstream tasks effectively. In contrast, **multitask fine-tuning approaches (MT-IL and MT-SILV) consistently maintain performance** across nearly all tasks, showing their robustness in preserving downstream task capabilities while integrating ET knowledge. Regarding performance across tasks, we observed that the two sentence-level tasks from *GLUE*, i.e. *COLA* and *SST-2*, exhibit contrasting performance degradation across nearly all fine-tuning strategies. While *COLA* is the most affected, *SST-2* shows the least degradation. Similarly, although both *RTE* and *WNLI* are categorized as inference tasks in *GLUE*, *RTE* experiences the most significant performance degradation when fine-tuned on ET data, while for *WNLI*, most fine-tuning strategies yield results comparable to the model not fine-tuned on ET data. These findings highlight that the interaction between ET signals and downstream tasks is not straightforward and may depend on deeper task-specific factors rather than broad task categorizations.

4.2 Impact on Attention Patterns

To address our second research question, we assessed the impact of fine-tuning on ET data by analyzing the correlation between the model's attention and human attention before and after the ET injection. Table 2 presents these correlations, using Attention Weights as model attention and *Total Reading Time* (TRT) as human attention proxies. We focused on this feature as it serves as a comprehensive measure of the reading process, encompassing both full semantic integration and syntactic reanalysis (de Varda and Marelli, 2023). Results for all the other ET features are reported in Appendix E showing consistent patterns across all of them. The EYE-ONLY row displays the average correlation across all users, whereas the BASE row represents the average correlation between the attention weights of the pre-trained RoBERTa-base model and human attention, serving as a baseline for comparison. For each layer of the model, we report the Spearman correlation coefficient, along with the average correlation score across all layers (AVG column). Appendix E.1 reports the correlation scores for each reader. Our results indicate that fine-tuning on **eye-tracking data aligns the model's attention weights more closely with human attention patterns, and this is true for all readers**. As highlighted in the table, this effect emerges from the sixth layer onward and is particularly pronounced in the last four layers, whereas in the base model the highest correlation is observed in the initial layers. Furthermore, for the base model, the attention of only half of the readers shows a statistically significant correlation (p-value ≤ 0.05) with the model's attention.

We now examine whether this increased alignment between model attention and human attention persists after the model undergoes fine-tuning on downstream tasks. Table 3 reports the Spearman correlation coefficients between model attention and human attention for models fine-tuned on the downstream tasks, across all injection strategies. For each model, we provide the correlation at the last layer (the closest to the classification/regression head) with correlation scores av-

	Model Layer												
Model	1	2	3	4	5	6	7	8	9	10	11	12	AVG
EYE-ONLY	0.23	0.20	0.26	0.23	0.19	0.22	0.25	0.25	**0.32**	0.27	0.24	0.29	0.25
BASE	0.25	0.21	**0.28**	0.25	0.21	0.17	0.16	0.19	0.18	0.05	0.08	0.12	0.18

Table 2: Spearman correlation coefficients between model attention and user attention for the model fine-tuned on predicting user ET features (EYE-ONLY) and RoBERTa-base (BASE). The scores are averaged across all users. Differently from EYE-ONLY, only half of users lead to significant correlations with BASE; the others were excluded from the mean.

eraged across all users[2]. Our primary focus is to compare the correlation scores of ET injected models against those of models fine-tuned directly on the downstream task, in order to determine whether incorporating gaze-related signals enhances or preserves the alignment between model attention and human attention, even after fine-tuning on task-specific objectives.

The first observation is that the last-layer correlations between ET injected models and human attention are largely preserved even after fine-tuning on the downstream task. Across almost all tasks, the average correlation scores for different injection strategies (after downstream task fine-tuning) remain closely aligned with those of the model fine-tuned exclusively on ET feature prediction. Notably, only minimal differences (≤ 0.05) are observed for *COLA*, *MRPC*, *RTE*, and *WNLI*, suggesting that **integrating eye-tracking signals does not substantially disrupt the learned alignment with human attention**. Regarding the effect of different injection strategies, the highest last-layer correlations are achieved by partial downstream task fine-tuning applied after intermediate fine-tuning on ET data (INT-LAST3, INT-LAST2, and INT-CLF). This result is expected, as the most correlated layers of the EYE-ONLY model were frozen during the second fine-tuning step. Notably, the two **multi-task injection strategies (MT-IL and MT-SILV) and LORA prove effective in maintaining alignment with human attention patterns**. The worst-performing system is INT-FULL, as expected, since fine-tuning a model sequentially on two different tasks leads to a loss of competence in the first task. However, when compared to the model fine-tuned exclusively on the downstream task (DTS-ONLY), all models that incorporate ET data consistently achieve higher correlation scores.

4.3 Impact on Representation Space

Using RoBERTa-base (BASE) and the model fine-tuned solely on downstream tasks (DST-ONLY) as baselines, we evaluate their isotropy and linear intrinsic dimensionality in comparison to models fine-tuned either exclusively on ET data or on both ET data and downstream tasks. Tables 4 and 5 report the corresponding scores on the GECO dataset[3]. As known from literature (Rudman and Eickhoff, 2024; Cheng et al., 2023; Li et al., 2020), the BASE model exhibits an anisotropic subspace (IsoScore* ≈ 0.029 in a range [0,1]) and reduces the number of linearly independent dimensions from 768 to 297. Further fine-tuning drives both metrics even lower. In addition, Tables 4 and 5 **demonstrate that fine-tuning on eye-tracking data (EYE-ONLY) makes the representations more anisotropic and lower-dimensional** compared to the RoBERTa-base model and to the same extent as models fine-tuned exclusively on downstream tasks. This trend is generally observed across all ET-injected models, with a few exceptions found in the LORA and MT-SILV models, as well as in specific tasks such as *QQP*, *SST-2*, and *QNLI*. Notably, INT-FULL maintains strong performance (Table 1) while effectively "compressing" its representational space compared to other injection methods. This suggests that **fine-tuning with eye-tracking data can reduce isotropy and dimensionality while preserving downstream task performance**. More generally, with respect to the different injection strategies, when aiming to maximize representation compression, the intermediate fine-tuning INT-* models emerges as the most effective. Furthermore, we observe a strong Spearman correlation ($C = 0.75$) between model isotropy and linear intrinsic dimensionality, a finding not previously reported to our knowledge.

5 Discussion and Conclusion

The impressive abilities of deep learning models have fostered growing interest across disciplines in understanding how they work and represent language. One approach to interpretability draws inspiration from human language behavior, seeking to uncover connections between artificial models and human cognition, with the potential to improve

[2]Appendix E.2 reports the same results where correlations are averaged across all layers of the model.

[3]Appendix D provides a full analysis of the embedding space for both GECO and the English-EWT dataset.

Fine-tuning	Attention correlation (last layer)										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	0.19	**0.35**	0.05	0.16	0.06	0.08	0.12	0.04	0.09	0.18	0.13
INT-LAST3	**0.29**	**0.28**	0.24	**0.31**	0.16	**0.29**	0.26	0.21	0.20	0.23	0.25
INT-LAST2	**0.28**	0.26	0.19	**0.28**	**0.30**	0.24	**0.28**	**0.29**	**0.31**	**0.28**	0.27
INT-CLF	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**
LORA	**0.27**	0.22	0.13	0.20	0.13	0.20	**0.32**	0.16	0.21	**0.30**	0.21
MT-IL	0.26	0.23	0.22	**0.27**	0.16	0.21	**0.27**	0.20	**0.27**	**0.28**	0.24
MT-SILV	0.25	0.11	**0.28**	0.15	**0.31**	0.23	**0.33**	**0.31**	0.14	**0.27**	0.24
DST-ONLY	0.06	0.08	0.05	0.01	0.07	0.03	0.02	0.07	0.11	0.12	0.08

Table 3: Correlations between human attention (*TRT*) and model attention on the **last layer** for each injection strategy. The scores are averaged across all readers. Highlighted cells indicate that the correlation score of the ET injected model exceeds that of DST-ONLY by at least 0.02 points. Bold scores are the highest correlation coefficients: those exceeding 0.27, which is 0.02 points lower than the last-layer correlation of EYE-ONLY.

F-T	Linear ID										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	**127**	**89**	191	185	242	11	161	**4**	**32**	127	**117**
INT-LAST3	173	135	194	162	**148**	154	**154**	92	142	154	151
INT-LAST2	162	148	166	160	160	153	157	142	158	158	157
INT-CLF	160	160	160	160	160	160	160	160	160	160	160
LORA	184	144	310	**158**	279	256	166	202	146	163	201
MT-IL	232	154	**110**	179	228	88	155	251	228	152	178
MT-SILV	249	209	233	268	251	207	206	221	264	209	232
DST-ONLY	289	249	249	249	249	**3**	278	**4**	249	**16**	186
BASE	297										–
EYE-ONLY	160										–

Table 4: Layer 12 Linear ID values averaged over all users in the GECO dataset. Entries in bold mark the lowest value for each task.

F-T	IsoScore* $\times 10^3$										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	**0.74**	**1.19**	2.75	15.59	3.03	**0.35**	5.95	0.88	**0.71**	9.74	4.09
INT-LAST3	7.92	2.96	7.40	6.79	**2.10**	3.53	4.36	**0.69**	3.46	5.00	4.42
INT-LAST2	5.89	3.78	7.45	5.05	5.58	4.08	4.35	3.24	5.77	4.69	4.99
INT-CLF	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99
LORA	11.26	5.36	30.23	8.47	11.34	28.62	6.01	9.99	2.72	5.27	11.93
MT-IL	4.34	5.02	**1.39**	**2.71**	4.06	1.07	**2.52**	5.83	4.66	3.69	**3.53**
MT-SILV	17.38	10.76	8.28	12.00	11.89	6.57	10.14	11.26	21.56	11.97	12.18
DST-ONLY	6.53	35.94	4.58	15.08	4.69	0.40	28.03	1.17	11.14	**0.27**	10.78
BASE	28.69										–
EYE-ONLY	4.97										–

Table 5: Layer 12 IsoScore* values ($\times 10^3$ for better visualization) averaged over all users in the GECO dataset. Entries in bold mark the lowest value for each task.

both transparency and efficiency beyond task performance. In this context, our work provides novel insights by exploring the integration of human attention patterns—captured through real-time ET signals—into transformer-based language models. Through a comprehensive set of experiments, we examined the impact of injecting ET features into a transformer encoder model, assessing both interpretability and performance implications. Our findings reveal that fine-tuning on ET prediction enhances the alignment between the model's attention mechanisms and gaze patterns, making its internal processing more reflective of human reading behavior. This pattern holds consistently across all readers in our dataset. Crucially, this increased alignment does not come at the cost of task performance. Despite the well-known risk of catastrophic forgetting in sequential fine-tuning, we observe that overall models incorporating ET data maintain strong performance across a diverse set of downstream tasks. Moreover, the alignment between model and human attention persists even after the model is fine-tuned on downstream tasks. Beyond attention alignment, we also examined the impact of incorporating ET knowledge on the model's representation space showing that this process makes representations more anisotropic and reduces their intrinsic dimensionality. This compression effect, however, does not compromise downstream task performance, suggesting that incorporating ET signals may lead to a more efficient representation space while preserving task-relevant knowledge—potentially opening new ways for model optimization and compression. As for the effectiveness of different injection strategies, we showed that certain approaches, particularly those based on full intermediate fine-tuning, consistently outperform others across nearly all dimensions and tasks. These strategies prove to be especially robust, maintaining higher downstream task performance while efficiently compressing the model's representational space. Conversely, partial fine-tuning strategies, while less effective in preserving downstream task performance, show the highest alignment with human attention. Overall, our study highlights the potential of leveraging cognitive signals for both interpretability and task effectiveness, offering a promising pathway for designing more transparent and efficient models. Future research should focus on scaling these techniques to larger models, exploring different architectures, and incorporating additional cognitive signals to deepen our understanding of deep learning systems.

Limitations

While our study provides valuable insights into the integration of eye-tracking (ET) data into encoder-based language models, several limitations must be acknowledged.

First, our experiments focus exclusively on RoBERTa-base, a single encoder-only architecture. Although RoBERTa is one of the most prominent transformer-based language model, our findings may not generalize to other models, such as different encoder-based architectures (e.g., BERT, DeBERTa). Future work should explore whether the observed effects hold across a broader range of architectures, including multilingual and domain-specific models.

Second, our study is limited to the Ghent Eye-Tracking Corpus (GECO) as the source of cognitive signals. While GECO provides high-quality eye-tracking data, it is relatively small and primarily based on English reading behavior. Larger, more diverse datasets spanning different languages and reading conditions (e.g., task-specific reading, second-language readers) could offer a more comprehensive understanding of how ET signals influence model behavior.

Acknowledgments

This work has been supported by:

- FAIR - Future AI Research (PE00000013) projects under the NRRP MUR program funded by the NextGenerationEU.;
- The project "XAI-CARE" funded by the European Union - Next Generation EU - NRRP M6C2 "Investment 2.1 Enhancement and strengthening of biomedical research in the NHS" (PNRR-MAD-2022-12376692_VADALA' – CUP F83C22002470001)
- The project "Human in Neural Language Models" (IsC93_HiNLM), funded by CINECA3 under the ISCRA initiative;
- Language Of Dreams: the relationship between sleep mentation, neurophysiology, and neurological disorders - PRIN 2022 2022BNE97C_SH4_PRIN2022.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Jane Ashby, Keith Rayner, and Charles Clifton. 2005. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, 58(6):1065–1086.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584.

Lisa Beinborn and Nora Hollenstein. 2023. *Cognitive Plausibility in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Cham. EBook Packages Synthesis Collection of Technology (R0).

Joshua Bensemann, Alex Yuxuan Peng, Diana Benavides Prado, Yang Chen, Neşet Özkan Tan, Paul M. Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

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

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International conference on learning representations*.

Emily Cheng, Corentin Kervadec, and Marco Baroni. 2023. Bridging information-theoretic and geometric

compression in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12397–12420, Singapore. Association for Computational Linguistics.

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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Andrea Gregor de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Annual Meeting of the Association for Computational Linguistics*.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5999–6011, Miami, Florida, USA. Association for Computational Linguistics.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Annual Meeting of the Association for Computational Linguistics*.

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.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2023. Is anisotropy inherent to transformers? *arXiv preprint arXiv:2306.07656*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with cognitive language processing signals. *CoRR*, abs/1904.02682.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena A. Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *North American Chapter of the Association for Computational Linguistics*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. 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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural

networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. 2024. Geometric signatures of compositionality across a language model's lifetime. *arXiv preprint arXiv:2410.01444*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4892–4907, Mexico City, Mexico. Association for Computational Linguistics.

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

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24, pages 109–165. Academic Press.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? assessing bert as a distributional semantics model. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Leveraging cognitive features for sentiment analysis. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 156–166, Berlin, Germany. Association for Computational Linguistics.

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.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Conference on Empirical Methods in Natural Language Processing*.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Adam J Parker and Timothy J Slattery. 2021. Spelling ability influences early letter encoding during reading: Evidence from return-sweep eye movements. *Quarterly Journal of Experimental Psychology*, 74(1):135–149. PMID: 32705948.

Owen Raymond, Yelaman Moldagali, and Naser Al Madi. 2023. A dataset of underrepresented languages in eye tracking research. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA. Association for Computing Machinery.

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

William Rudman and Carsten Eickhoff. 2024. Stable anisotropic regularization. In *The Twelfth International Conference on Learning Representations*.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

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

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold

standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Timothy J. Slattery and Mark Yates. 2018. Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology*, 71(1):250–259.

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, Online. Association for Computational Linguistics.

Adrian Staub. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. 2023. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In *Conference on Empirical Methods in Natural Language Processing*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *ArXiv*, abs/2209.11055.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024b. Probing large language models from a human behavioral perspective. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 1–7, Torino, Italia. ELRA and ICCL.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Annual Meeting of the Association for Computational Linguistics*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

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

Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527.

A Training Hyperparameters

In general, all fine-tuning procedures used the following hyperparameter settings: a learning rate of $1e-05$, 10 epochs, a batch size of 16, a warmup ratio of 0.06, and a weight decay of 0.1. However, there were some exceptions:

- For tasks with larger datasets, such as *MNLI*, *QNLI*, *QQP*, and *SST-2*, fine-tuning was performed for only 3 epochs, except when using LoRa adapters or fine-tuning on *WNLI*, where training extended to 20 epochs.
- All fine-tuning with LoRa adapters employed a learning rate of $5e-05$ and lasted for 10 epochs, except for *WNLI*, where training was conducted for 20 epochs.

For LoRa adapter settings, we used the following default configuration: $r = 32$, $\alpha = 8$, and a dropout rate of 0.05. For all the experiments we employed the Transformers library of Huggingface[4].

B Performances on prediction of eye-tracking features

Table B.1 summarizes the model's performance on the eye-tracking feature prediction task for each user, reported in terms of Spearman correlation coefficient. All reported correlations are statistically significant at a p-value of 0.05. Correlations that are not significant are marked as NS. These scores refer to the pre-trained RoBERTa-base fine-tuned exclusively on this task, i.e., the first step of the injection process using intermediate fine-tuning (INT-*) and LoRa adapters (LORA).

From now on we will use these acronyms for eye-tracking features: FFD = First Fixation Duration, GD = Gaze Duration, FRNF = First-run Number of Fixations, TRT = Total Reading Time and TNF = Total Number of Fixations.

The results indicate that, for most readers, the model achieves consistent performance across all eye-tracking features. However, for some readers, such as User 29, the model performs significantly worse than the average, while for others, like User 30, it achieves substantially better results. These findings support our decision to treat data from each user separately rather than aggregating it, as this allows us to better capture user-specific patterns.

[4] https://huggingface.co/

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
21	0.34	0.38	0.39	0.37	0.38	0.37
22	0.36	0.36	0.36	0.36	0.36	0.36
23	0.37	0.42	0.42	0.42	0.42	0.41
26	0.29	0.34	0.34	0.33	0.34	0.33
28	0.32	0.36	0.37	0.35	0.36	0.35
29	0.21	0.25	0.26	-0.18	0.28	0.16
30	0.50	0.51	0.51	0.50	0.50	0.50
31	0.36	0.37	0.37	0.37	0.37	0.37
32	0.37	0.39	0.39	0.38	0.39	0.38
33	0.45	0.45	0.45	0.45	0.45	0.45
34	0.27	0.32	0.32	0.30	0.33	0.31
35	0.36	0.41	0.40	0.41	0.42	0.40
AVG	0.35	0.38	0.38	0.34	0.38	0.37

Table B.1: Spearman correlation coefficient for each user on the prediction of eye-tracking features of the RoBERTa-base fine-tuned only on this task.

Table B.2 reports the same evaluation, but measured using Mean Absolute Error (MAE). For comparison, Table B.3 reports the performance of a baseline that always predicts the average value of each feature.

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
21	5.01	4.97	4.98	5.00	5.00	4.99
22	3.55	3.54	3.55	3.55	3.55	3.55
23	4.92	4.88	4.90	4.91	4.90	4.90
26	4.88	4.84	4.85	4.86	4.86	4.86
28	4.48	4.44	4.43	4.47	4.45	4.45
29	7.24	7.23	7.23	7.24	7.23	7.23
30	2.80	2.79	2.79	2.80	2.80	2.80
31	3.04	3.04	3.03	3.04	3.04	3.04
32	3.83	3.82	3.82	3.83	3.83	3.83
33	2.66	2.66	2.66	2.66	2.66	2.66
34	6.35	6.31	6.32	6.34	6.32	6.33
35	5.38	5.31	5.34	5.35	4.34	5.14
AVG	4.59	4.56	4.56	4.57	4.57	4.57

Table B.2: Mean Absolute Error for each user on the prediction of eye-tracking features of the RoBERTa-base fine-tuned only on this task.

We can clearly see that each user has almost the same error across all the features. This is expected since the 5 features lay in the same value range, and their loss is averaged during model training. The results are in line with the evaluation using Spearman correlation coefficients.

Comparing the results with the baseline, we can see that the models outperform it.

Below, we report the performance of models trained using multi-task injection strategies on the prediction of eye-tracking features. Tables B.4 and B.5 present the results for the model injected using interleaved fine-tuning (MT-IL). The first table shows the scores averaged across all downstream tasks for each user, while the second table reports the scores for models trained on each downstream task, averaged across all users.

While the correlation scores are slightly lower on

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
21	7.42	5.31	5.14	5.12	5.49	5.70
22	8.64	3.95	5.18	5.22	3.21	5.24
23	7.79	5.39	7.65	8.26	6.49	7.12
26	4.96	5.14	7.37	6.14	6.17	5.96
28	5.41	4.64	7.13	4.18	2.98	4.87
29	9.85	7.15	10.34	5.55	5.09	7.59
30	9.50	3.34	4.60	5.56	4.51	5.50
31	2.94	3.24	6.14	6.39	3.81	4.50
32	11.86	4.32	5.14	4.71	3.65	5.94
33	9.12	3.04	7.37	5.07	3.53	5.62
34	7.31	6.79	8.94	9.43	8.14	8.12
35	7.56	5.90	6.55	4.33	3.64	5.60
AVG	7.70	4.85	6.80	5.83	4.73	5.98

Table B.3: Mean Absolute Error for each user on the prediction of eye-tracking features of the *baseline* always predicting the mean value.

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
21	0.37	0.37	0.38	0.37	0.38	0.37
22	0.28	0.29	0.29	0.29	0.29	0.29
23	0.34	0.35	0.36	0.33	0.36	0.35
26	0.28	0.29	0.29	0.27	0.30	0.29
28	0.34	0.35	0.35	0.34	0.35	0.35
29	0.35	0.35	0.35	0.35	0.36	0.35
30	0.40	0.40	0.41	0.38	0.42	0.40
31	0.30	0.30	0.30	0.30	0.31	0.30
32	0.30	0.31	0.30	0.29	0.31	0.30
33	0.29	0.31	0.30	0.27	0.33	0.30
34	0.25	0.26	0.26	0.23	0.27	0.25
35	0.40	0.41	0.40	0.41	0.41	0.41
AVG	0.32	0.33	0.33	0.32	0.34	0.33

Table B.4: Spearman correlation coefficient on the prediction of eye-tracking features on Interleaved Multitask injection (MT-IL), averaged across all tasks for each user.

average, with respect to those of the model solely finetuned on eye-tracking feature prediction, the relative ranking of performance across users is largely preserved. There is not a big difference on eye-tracking prediction, while varying the downstream task.

Lastly, Tables B.6 and B.7 contain the results of the models injected using multitask fine-tuning with silver labels (MT-SILV), respectively aggregated across tasks and across users. It is important to note that, for this model, the eye-tracking features are not gold but are instead predicted by another model. Consequently, these results are less indicative of the model's true ability to predict eye-tracking features.

C Stability of IsoScore*

Figure 2 demonstrates that IsoScore* is stable and effectively differentiates between isotropic and anisotropic point clouds. This contrasts with the **Partition Score**, a widely recognized metric in literature (Arora et al., 2016; Mu and Viswanath,

	Eye-tracking feature					
Task	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
COLA	0.35	0.34	0.35	0.35	0.35	0.35
COMP	0.35	0.32	0.34	0.33	0.30	0.33
MNLI	0.31	0.31	0.31	0.31	0.31	0.31
MRPC	0.35	0.32	0.35	0.33	0.33	0.34
QNLI	0.34	0.34	0.34	0.34	0.34	0.34
QQP	0.31	0.31	0.31	0.31	0.31	0.31
RTE	0.34	0.29	0.32	0.32	0.31	0.32
SST-2	0.36	0.35	0.35	0.36	0.36	0.36
STSB	0.36	0.35	0.36	0.36	0.36	0.36
WNLI	0.32	0.29	0.30	0.32	0.25	0.29
AVG	0.34	0.32	0.33	0.33	0.32	0.33

Table B.5: Spearman correlation coefficient on the prediction of eye-tracking features on Interleaved Multitask injection (MT-IL), averaged across all user for each task.

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
21	0.90	0.91	0.90	0.90	0.90	0.90
22	0.92	0.93	0.92	0.92	0.92	0.92
23	0.91	0.92	0.91	0.91	0.91	0.91
26	0.93	0.93	0.93	0.92	0.93	0.93
28	0.88	0.89	0.88	0.88	0.88	0.88
29	0.78	0.81	0.76	0.78	0.78	0.78
30	0.94	0.94	0.94	0.94	0.94	0.94
31	0.92	0.92	0.91	0.92	0.91	0.92
32	0.92	0.93	0.93	0.92	0.92	0.92
33	0.92	0.92	0.92	0.92	0.92	0.90
34	0.90	0.91	0.90	0.90	0.90	0.93
35	0.93	0.93	0.93	0.93	0.93	0.90
AVG	0.90	0.91	0.90	0.90	0.90	0.90

Table B.6: Spearman correlation coefficient on the prediction of eye-tracking features on silver labels multitask fine-tuning (MT-SILV) averaged across all tasks.

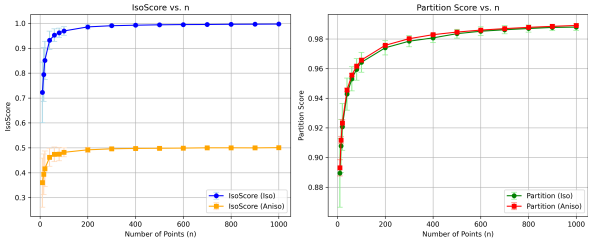

Figure 2: IsoScore*compared to the Partition Score known in literature (1000 runs per number of points).

	Eye-tracking feature					
User	**FFD**	**GD**	**FRNF**	**TNF**	**TRT**	**AVG**
COLA	0.91	0.91	0.91	0.91	0.91	0.91
COMP	0.93	0.92	0.93	0.92	0.90	0.92
MNLI	0.93	0.93	0.93	0.93	0.93	0.93
MRPC	0.91	0.91	0.92	0.91	0.91	0.91
QNLI	0.91	0.91	0.91	0.91	0.91	0.91
QQP	0.95	0.95	0.95	0.95	0.95	0.95
RTE	0.89	0.89	0.89	0.89	0.89	0.89
SST2	0.93	0.93	0.93	0.93	0.93	0.93
STSB	0.85	0.85	0.86	0.85	0.86	0.86
WNLI	0.83	0.84	0.87	0.83	0.84	0.84
AVG	0.90	0.90	0.91	0.90	0.90	0.90

Table B.7: Spearman correlation coefficient on the prediction of eye-tracking features on silver labels multitask fine-tuning (MT-SILV) averaged across all users.

2018).

D Embedding space

This Section extends Section 4.3, by reporting Linear ID (Tables D.8) and IsoScore* (D.9) computed on the English-UD dataset, to study the impact of eye-tracking injection on the embedding space, when representing out-of-domain sentences.

	Linear ID										
F-T	**COLA**	**COMP**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	243	153	243	234	284	30	254	**3**	**159**	143	**175**
INT-LAST3	207	182	246	214	209	226	204	183	210	197	208
INT-LAST2	**191**	183	209	**196**	**192**	200	204	185	199	198	196
INT-CLF	198	198	198	198	198	198	**198**	198	198	198	198
LORA	258	166	304	216	309	322	218	309	246	200	255
MT-IL	268	231	**168**	265	272	146	286	267	301	207	241
MT-SILV	199	**138**	203	269	236	202	208	207	245	198	211
DST-ONLY	305	249	265	273	288	**4**	304	4	264	**24**	198
BASE					308						–
EYE-ONLY					198						–

Table D.8: The 12-layer Linear ID values averaged over all users for English-ETW. Entries in **bold** mark the lowest value for each task.

	IsoScore* $\times 10^3$										
F-T	**COLA**	**COMP**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	5.74	**2.95**	4.54	11.79	6.56	**0.63**	12.66	0.56	**2.78**	3.74	5.20
INT-LAST3	9.84	5.71	8.28	9.01	**4.91**	5.31	5.89	1.84	9.53	4.71	6.50
INT-LAST2	5.96	5.20	6.24	**4.92**	4.85	4.53	**5.12**	4.18	6.54	4.52	5.21
INT-CLF	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
LORA	15.17	4.10	21.15	7.46	28.98	34.77	6.34	27.13	16.74	5.23	16.71
MT-IL	8.88	6.70	6.07	12.28	13.06	3.21	22.62	5.86	32.22	5.65	11.66
MT-SILV	**4.48**	2.63	**3.39**	8.33	5.19	3.25	5.91	3.61	7.76	5.66	**5.02**
DST-ONLY	12.89	5.34	5.22	6.09	6.21	1.83	17.31	**0.53**	12.11	**0.67**	6.82
BASE					11.90						–
EYE-ONLY					5.25						–

Table D.9: Layer 12 IsoScore* ($\times 10^3$ for better visualization) averaged over users for English-ETW. Entries in **bold** mark the lowest value for each task.

The results show a significant correlation (c_{id} = 0.74 and c_{iso} = 0.40) with those obtained on GECO. Overall, fine-tuning the models injecting eye-tracking features in the model reduces both linear dimensionality and isotropy without compromising performance (see Table 1).

Table D.10 and D.11 present an average over users, the two datasets, and, for some models, the downstream tasks, of our scores. They confirm the findings discussed in 4.3. In almost every layer, INT-FULL appears to be the model that reduces the embedding dimensions the most and exhibits the highest anisotropy.

E Additional Results on Attention correlation

This appendix provides supplementary results related to Section 4.2. Specifically, Section E.1 presents the correlation coefficients between model attention and human attention for all users, using the Total Reading Time (TRT) feature, as in the main body of the paper. Section E.2 reports correlation scores averaged across all model layers rather than reporting the last layer correlation. Finally, Section E.3 reports the correlation scores aggregated across users but computed using the four remaining eye-tracking features.

E.1 Correlations with TRT for all users

Table E.12 extends Table 2 by reporting correlation scores for each individual user. Highlighted cells indicate that the correlation score for a given user at a specific layer is higher than the average correlation score of the RoBERTa model without fine-tuning on eye-tracking data (BASE). Bold values represent the highest correlation scores for each user.

We observe that, in most cases, the correlation increases as we move toward the final layers of the model, peaking around layers 9 and 10. Interestingly, there is no clear relationship between a model's performance in predicting eye-tracking features and the average correlation between model and human attention across different users.

E.2 Attention correlation with TRT with model attention averaged across layers

Table E.13 is analogous to Table 3, but instead of reporting the correlation between human attention and the model's last-layer attention, it presents the correlation scores averaged across all model layers. The overall correlation scores remain similar on average, and the results are largely consistent with those in Section 4.2. One difference is that when averaging attention across all layers, models injected using LoRa adapters (LORA) exhibit slightly higher correlations than those injected via multitask fine-tuning with interleaved steps (MT-IL).

E.3 Attention correlations with other eye tracking features

This section extends Table 3 by using the remaining four eye-tracking features to represent human attention. Specifically, Table E.14 reports correlation scores obtained using First Fixation Duration

	Linear ID											
Layer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
INT-FULL	340	338	336	327	**317**	**300**	**275**	**248**	**205**	**179**	**152**	**143**
INT-LAST3	340	340	342	337	329	317	293	272	250	228	201	197
INT-LAST2	340	340.	342	337	329	317	293	272	250	227.8	204	198
INT-CLF	340	340	342	337	329	317	293	272	250	228	204	198
LORA	340	341	342	337	330	318	294	273	251	229	206	200
MT-IL	341	341	344	339	332	320	299	280	258	237	217	207
MT-SILV	341	343	344	340	333	317	297	272	256	236	204	199
DST-ONLY	**339**	340	347	350	348	336	312	290	277	262	251	213
BASE	340	342	349	351	353	349.	347	345	340	332	321.0	303
EYE-ONLY	**337**	**332**	**333**	**326**	320	309	287	263	238	216	193	179

Table D.10: Average Linear ID (over user, dataset and, for some models, downstream task). Entries in **bold** mark the lowest value for each task.

	IsoScore* $\times 10^3$											
Layer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
INT-FULL	**31.35**	**25.23**	**20.91**	**14.91**	**10.75**	**8.73**	**7.61**	**5.81**	**4.13**	**3.39**	**2.59**	**3.74**
INT-LAST3	31.57	25.69	23.88	17.49	12.52	10.80	10.13	7.21	5.70	4.45	2.84	4.72
INT-LAST2	31.57	25.69	23.88	17.49	12.52	10.80	10.13	7.21	5.70	4.45	3.09	4.52
INT-CLF	31.57	25.69	23.88	17.49	12.52	10.80	10.13	7.21	5.70	4.45	3.09	5.26
LORA	31.56	25.69	23.84	17.44	12.54	10.86	10.19	7.23	5.71	4.43	3.11	5.23
MT-IL	31.72	25.70	24.86	18.19	13.74	12.10	12.03	9.03	7.19	5.55	3.78	5.66
MT-SILV	31.90	26.66	23.63	17.58	13.15	10.58	8.53	7.94	6.01	5.80	3.14	5.67
DST-ONLY	34.26	30.12	33.32	30.99	27.94	25.20	22.15	23.19	22.29	19.89	16.57	8.80
BASE	34.73	30.96	33.25	29.30	27.68	26.33	26.09	26.33	25.67	23.99	21.54	20.29
EYE-ONLY	33.02	27.21	26.31	21.21	18.73	17.07	16.85	13.71	9.73	7.40	5.05	5.13

Table D.11: Average IsoScore* (over user, dataset and, for some models, downstream task), all values multiplied by 10^3. Entries in **bold** mark the lowest value for each task.

(FFD), Table E.15 uses Gaze Duration (GD), Table E.16 employs First Run Number of Fixations (FRNX), and Table E.17 presents results based on the Total Number of Fixations (NFIX). Once again, the results remain consistent with those in Section 4.2.

	Model Layer												
User	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**AVG**
21	0.25	0.16	0.25	0.29	0.20	0.19	0.25	0.27	**0.34**	NS	0.32	0.33	0.26
22	0.18	0.14	0.19	0.16	0.15	0.19	0.19	0.19	**0.33**	0.31	0.29	0.24	0.21
23	0.25	0.23	0.29	0.25	0.21	0.25	0.33	0.32	**0.39**	0.29	0.20	0.35	0.28
26	0.19	0.18	0.21	0.17	0.13	0.14	0.21	0.26	0.24	NS	0.26	**0.32**	0.20
28	0.24	0.19	0.25	0.21	0.12	0.21	0.21	0.27	0.27	0.25	**0.33**	0.20	0.23
29	0.23	0.24	0.28	0.21	0.23	0.21	0.24	**0.27**	**0.27**	0.10	0.10	0.13	0.21
30	0.24	0.27	0.32	0.26	0.25	0.29	0.27	0.19	0.39	**0.44**	**0.44**	0.40	**0.31**
31	0.23	0.16	0.23	0.20	0.10	0.17	0.20	0.22	**0.32**	0.20	0.10	**0.32**	0.20
32	0.25	0.24	0.29	0.24	0.24	0.26	0.29	0.28	0.33	**0.34**	**0.34**	0.34	0.29
33	0.26	0.23	0.28	0.25	0.24	0.22	0.19	0.27	0.36	**0.37**	0.33	0.23	0.27
34	0.19	0.18	0.23	0.18	0.13	0.20	**0.27**	0.23	0.21	0.13	0.02	0.26	0.19
35	0.28	0.24	0.32	0.27	0.24	0.28	0.36	0.37	0.37	0.22	0.16	**0.38**	0.29
AVG	0.23	0.20	0.26	0.23	0.19	0.22	0.25	0.25	**0.32**	0.27	0.24	0.29	0.25
BASE	0.25	0.21	**0.28**	0.25	0.21	0.17	0.16	0.19	0.18	0.05	0.08	0.12	0.18

Table E.12: Spearman correlation coefficients between model attention and user attention for the RoBERTa-base model fine-tuned on predicting user eye-tracking features.

	Attention correlation TRT (layers avg)										
Fine-tuning	**COLA**	**COMP**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	0.17	0.22	0.16	0.18	0.16	0.16	0.17	0.15	0.15	0.20	0.17
INT-LAST3	**0.25**	**0.24**	**0.24**	**0.25**	**0.23**	**0.24**	**0.25**	**0.23**	**0.24**	**0.24**	**0.24**
INT-LAST2	**0.24**	**0.24**	**0.24**	**0.24**	**0.25**	**0.24**	**0.24**	**0.25**	**0.25**	**0.24**	**0.24**
INT-CLF	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
LORA	0.21	**0.24**	0.18	**0.24**	0.18	0.19	**0.25**	0.20	0.21	**0.25**	**0.21**
MT-IL	0.19	**0.24**	0.14	0.20	0.17	0.14	0.21	0.20	**0.23**	**0.23**	0.19
MT-SILV	0.22	**0.23**	0.20	0.20	0.22	0.18	**0.25**	0.21	0.21	**0.24**	0.22
DST-ONLY	0.17	0.17	0.16	0.16	0.16	0.13	0.17	0.14	0.17	0.17	0.16
BASE					0.18						–
EYE-ONLY					0.25						–

Table E.13: Correlations between human attention and model attention **averaged acros all models' layers** for each eye-tracking injection strategy. The used eye-tracking features is TRT The scores are averaged across all readers.

	Attention correlation FFD (last layer)										
Fine-tuning	**COLA**	**COMP**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	0.17	**0.31**	0.04	0.14	0.06	0.05	0.09	0.03	0.10	0.16	0.12
INT-LAST3	**0.26**	**0.26**	0.21	**0.28**	0.15	**0.26**	**0.24**	0.18	0.19	0.21	0.22
INT-LAST2	**0.25**	0.23	0.18	**0.25**	**0.29**	**0.24**	0.23	**0.26**	**0.27**	**0.25**	**0.25**
INT-CLF	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**	**0.26**
LORA	**0.24**	0.21	0.12	0.19	0.12	0.18	**0.28**	0.15	0.18	**0.27**	0.19
MT-IL	0.22	0.23	0.19	**0.24**	0.15	0.19	0.23	0.20	**0.24**	**0.25**	0.22
MT-SILV	0.22	0.10	**0.25**	0.14	**0.28**	0.21	**0.30**	**0.28**	0.13	**0.25**	0.21
DST-ONLY	0.05	0.05	NS	0.10	NS	NS	0.03	0.06	0.10	0.11	0.07
BASE					0.11						–
EYE-ONLY					0.26						–

Table E.14: Correlations between human attention and model attention on the **last layer** for each eye-tracking injection strategy. The used eye-tracking features is FFD The scores are averaged across all readers.

	Attention correlation GD (last layer)										
Fine-tuning	**COLA**	**COMP**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STSB**	**WNLI**	**AVG**
INT-FULL	0.18	**0.33**	0.04	0.16	0.06	0.05	0.09	0.03	0.10	0.18	0.12
INT-LAST3	**0.28**	**0.29**	0.23	**0.30**	0.16	**0.28**	0.24	0.20	0.20	0.22	0.24
INT-LAST2	**0.27**	0.25	0.20	**0.27**	**0.31**	0.25	0.25	**0.28**	**0.29**	**0.27**	**0.27**
INT-CLF	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
LORA	**0.26**	0.22	0.13	0.20	0.13	0.19	**0.31**	0.16	0.20	**0.29**	0.21
MT-IL	0.24	0.25	0.21	**0.26**	0.15	0.20	0.25	0.19	**0.26**	0.27	0.23
MT-SILV	0.24	0.11	**0.27**	0.15	**0.30**	0.22	**0.32**	**0.30**	0.14	**0.26**	0.23
DST-ONLY	0.08	0.11	0.03	0.08	0.04	NS	0.04	0.05	0.11	0.16	0.08
BASE					0.13						–
EYE-ONLY					0.28						–

Table E.15: Correlations between human attention and model attention on the **last layer** for each eye-tracking injection strategy. The used eye-tracking features is GD The scores are averaged across all readers.

Fine-tuning	Attention correlation FRNF (last layer)										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	0.17	**0.32**	0.04	0.17	0.06	0.05	0.10	0.03	0.11	0.17	0.12
INT-LAST3	**0.27**	**0.28**	0.23	**0.29**	0.15	**0.25**	**0.25**	0.21	0.19	0.22	0.24
INT-LAST2	**0.27**	0.24	0.19	**0.26**	**0.28**	0.23	0.24	**0.28**	**0.29**	**0.26**	**0.25**
INT-CLF	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**
LORA	**0.26**	0.21	0.12	0.20	0.12	0.19	**0.30**	0.16	0.20	**0.28**	0.20
MT-IL	0.23	0.24	0.20	**0.25**	0.14	0.20	0.24	0.20	**0.25**	**0.28**	0.22
MT-SILV	0.24	0.11	**0.26**	0.14	**0.29**	0.21	**0.31**	**0.29**	0.13	0.25	0.22
DST-ONLY	0.08	0.10	0.03	0.08	0.04	0.03	0.04	0.05	0.10	0.16	0.07
BASE					0.12						–
EYE-ONLY					0.27						–

Table E.16: Correlations between human attention and model attention on the **last layer** for each eye-tracking injection strategy. The used eye-tracking features is FRNF The scores are averaged across all readers.

Fine-tuning	Attention correlation NFIX (last layer)										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	0.18	**0.33**	0.04	0.16	0.06	0.05	0.10	0.04	0.11	0.17	0.12
INT-LAST3	**0.28**	**0.29**	0.23	**0.30**	0.16	**0.29**	**0.26**	0.19	0.19	0.23	0.24
INT-LAST2	**0.28**	0.25	0.19	**0.28**	**0.27**	**0.26**	**0.27**	**0.28**	**0.30**	**0.27**	**0.26**
INT-CLF	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
LORA	**0.27**	0.21	0.13	0.20	0.13	0.19	**0.31**	0.17	0.20	**0.29**	0.21
MT-IL	0.24	0.25	0.21	**0.26**	0.16	0.20	**0.26**	0.20	**0.26**	**0.27**	0.23
MT-SILV	0.24	0.11	**0.27**	0.15	**0.30**	0.22	**0.32**	**0.31**	0.14	**0.26**	0.23
DST-ONLY	0.09	0.11	NS	0.08	0.04	NS	0.05	0.05	0.11	0.16	0.09
BASE					0.12						–
EYE-ONLY					0.28						–

Table E.17: Correlations between human attention and model attention on the **last layer** for each eye-tracking injection strategy. The used eye-tracking features is NFIX The scores are averaged across all readers.