Do Causal Language Model Attention Patterns Mirror Human Reading Fixations?
A Multi-Model, Multi-Dataset Analysis

Anonymous ACL submission

Abstract

Do the internal attention distributions of causal language models reflect how humans allocate visual attention during reading? We systematically compare self-attention patterns from eight pretrained autoregressive models (GPT-2 family, Llama 3.2 1B/3B, Qwen 2.5 0.5B/1.5B) with human eye-tracking fixation data across three established corpora (ZuCo, Provo, GECO). Using word-level Spearman correlation between model attention (received attention per word) and human total reading time, we find that individual attention heads achieve surprisingly strong alignment with human fixation patterns, up to $\rho = 0.589$ for Llama-3.2-1B on ZuCo. This alignment is robust across datasets and statistically significant ($p < 0.001$, permutation test). Critically, we discover an architectural divide: Llama models consistently outperform GPT-2 models regardless of parameter count, suggesting that architecture matters more than scale for human-like attention. Regression analysis reveals that model attention captures psycholinguistic effects including word frequency and surprisal, while position bias analysis uncovers a parallel to the human sentence wrap-up effect. Our results provide the first comprehensive comparison of modern causal LM attention with human reading across multiple corpora and model families.

1 Introduction

The attention mechanism (Vaswani et al., 2017) is the computational core of modern language models. While its primary purpose is computational, routing information between token representations, a natural question arises: do these learned attention patterns bear any resemblance to how humans allocate attention during language processing?

Human reading, as measured by eye-tracking, provides a rich signal of cognitive attention allocation. When reading, humans do not fixate every word equally; instead, fixation durations reflect processing difficulty driven by word frequency, predictability, syntactic complexity, and other factors (Rayner, 1998; Just and Carpenter, 1980). If language model attention captures similar patterns, this would suggest that statistical language learning gives rise to attention distributions that parallel human cognitive processing, even without any explicit pressure to do so.

Prior work has explored this question primarily with BERT (Devlin et al., 2019). Eberle et al. (2022) and Bensemann et al. (2022) compared BERT's attention with human gaze data, finding modest correlations in early layers. Sood et al. (2020) examined attention in reading comprehension models. However, these studies share important limitations: they analyzed only masked language models (primarily BERT), used a single eye-tracking dataset, and did not systematically vary model architecture or scale.

We address these gaps with a comprehensive analysis that makes three contributions:

1. **Modern causal models.** We analyze eight autoregressive LMs spanning three architectures (GPT-2, Llama 3.2, Qwen 2.5) and parameter counts from 124M to 3.2B, the first such analysis of modern causal models against human fixation data.

2. **Multi-dataset validation.** We evaluate on three established eye-tracking corpora, ZuCo (Hollenstein et al., 2018), Provo (Luke and Christianson, 2018), and GECO (Cop et al., 2017), demonstrating that our findings generalize across datasets with different participants, materials, and languages of collection.

3. **Architecture vs. scale.** We discover that model architecture is a stronger predictor of human-like attention than parameter count:

Llama models with 1.2B parameters outperform GPT-2 models at all sizes (124M–1.5B), revealing that grouped-query attention and modern training produce more human-like attention patterns.

2 Related Work

Attention and human cognition. The relationship between transformer attention and human cognition has been studied through several lenses. Clark et al. (2019) showed that BERT attention heads specialize for syntactic functions. Sood et al. (2020) compared attention in machine reading comprehension models with human eye-tracking during the same task, finding partial alignment. Hollenstein et al. (2021) demonstrated that multilingual models can predict human reading times across languages.

Attention and eye-tracking. Eberle et al. (2022) compared BERT attention with human gaze in sentiment analysis, finding that first-layer attention heads correlate with fixation patterns. Bensemann et al. (2022) performed a more systematic layer-by-layer analysis of BERT attention against ZuCo fixation data. Oh and Schuler (2023) compared attention across transformer variants. However, all of these studies focused on bidirectional (masked) models, leaving autoregressive models unexplored.

Surprisal and reading. A separate line of work connects language model predictions to reading difficulty. Surprisal theory (Hale, 2001; Levy, 2008) predicts that word reading time is proportional to the negative log-probability of the word given its context. This has been confirmed empirically (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Shain et al., 2024). Our work bridges these two threads by examining whether attention, a distinct internal mechanism from next-word prediction, also captures these psycholinguistic effects.

Attention as explanation. The debate on whether attention provides faithful explanations of model behavior (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) is relevant but orthogonal to our question. We do not claim that attention explains *why* models make specific predictions; rather, we examine the empirical correlation between attention distributions and human reading patterns as a window into shared computational strategies for language processing.

3 Methodology

3.1 Eye-Tracking Datasets

We use three established eye-tracking corpora that record human reading behavior:

ZuCo (Hollenstein et al., 2018, 2020) provides simultaneous EEG and eye-tracking from 12 subjects reading 300 English sentences (Normal Reading task). We average fixation measures across subjects and retain 298 sentences after filtering for word count consistency.

Provo (Luke and Christianson, 2018) contains eye-tracking data from 84 participants reading 55 short English texts, yielding 133 sentences after segmentation and filtering (3–60 words).

GECO (Cop et al., 2017) provides eye-tracking from 14 monolingual English participants reading an entire novel. We split the reading trials into sentences using punctuation-based segmentation and subsample 1,000 sentences.

For each dataset, we extract four word-level reading measures: **Total Reading Time** (TRT), **First Fixation Duration** (FFD), **Gaze Duration** (GD), and **Number of Fixations** (nFix), averaged across participants.

3.2 Language Models

We analyze eight pretrained causal language models spanning three architectural families:

- **GPT-2 family** (Radford et al., 2019): GPT-2 (124M), GPT-2-medium (355M), GPT-2-large (774M), GPT-2-XL (1.5B). Standard multi-head attention.
- **Llama 3.2** (Grattafiori et al., 2024): 1B and 3B parameter variants. Grouped-query attention (GQA), RoPE positional encoding, SwiGLU activation.
- **Qwen 2.5** (Yang et al., 2024): 0.5B and 1.5B parameter variants. Grouped-query attention, RoPE, SwiGLU.

All models are loaded with `attn_implementation="eager"` to obtain full attention weight matrices (required since SDPA and FlashAttention do not return attention weights).

3.3 Attention Extraction and Alignment

Subword-to-word alignment. Modern language models operate on subword tokens, while eye-tracking data is word-level. We align these using the tokenizer's `offset_mapping`, which maps each token to character spans in the original text. Each token is assigned to the word whose character span it overlaps with. BOS/EOS tokens are excluded.

Received attention aggregation. For each sentence, we extract the full attention tensor $\mathbf{A} \in \mathbb{R}^{L\times H\times T\times T}$ where L is the number of layers, H is the number of heads, and T is the sequence length. To obtain a word-level attention distribution, we:

1. Compute *received attention* per token: $a_j = \sum_i A_{i,j}$ (column sum over the attention matrix), representing how much each token is attended to by all other tokens.

2. Aggregate subword tokens to words by summing the received attention of all tokens belonging to the same word.

3. Normalize to a probability distribution over words.

This "received attention" framing parallels fixation data: just as TRT measures how much total processing a word receives from the reader, received attention measures how much total attention a word receives from the model.

3.4 Evaluation Metrics

Spearman rank correlation (ρ). Our primary metric, following Eberle et al. (2022) and Bensemann et al. (2022). For each sentence, we compute the Spearman correlation between the model's word-level attention distribution and the human fixation distribution. We report the mean correlation across all sentences.

Permutation test. To assess statistical significance, we use a permutation test with 1,000 iterations. For each permutation, we shuffle the human fixation values within each sentence (breaking the word-level alignment) and recompute the mean cross-sentence correlation. The p-value is the proportion of permuted correlations $\geq$ the observed correlation.

Table 1: Best single-head Spearman ρ (TRT) on ZuCo for each model, with architectural details. Layer Mean = mean ρ across all heads in the best layer.

Model	Params	L	H	Best ρ	Best Head
GPT-2	124M	12	12	0.490	L0H4
GPT-2-med	355M	24	16	0.455	L0H10
Qwen-0.5B	494M	24	14	0.498	L0H6
GPT-2-lg	774M	36	20	0.496	L3H18
Llama-1B	1.2B	16	32	**0.589**	L6H6
Qwen-1.5B	1.5B	28	12	0.505	L0H7
GPT-2-XL	1.5B	48	25	0.450	L0H21
Llama-3B	3.2B	28	32	0.573	L21H11

Table 2: Permutation test results (1,000 permutations, ZuCo TRT). All models show alignment significantly above chance.

Model	Observed ρ	Null $\mu \pm \sigma$	p
GPT-2	0.490	0.001 ± 0.014	< 0.001
GPT-2-med	0.455	0.000 ± 0.014	< 0.001
Qwen-0.5B	0.498	0.001 ± 0.014	< 0.001
GPT-2-lg	0.496	0.000 ± 0.014	< 0.001
Llama-1B	0.589	0.000 ± 0.015	< 0.001
Qwen-1.5B	0.505	0.000 ± 0.014	< 0.001
GPT-2-XL	0.450	0.000 ± 0.014	< 0.001
Llama-3B	0.573	0.000 ± 0.015	< 0.001

Regression analysis. To understand what drives the attention–fixation alignment, we regress the difference between model attention and human fixation on word properties: word length, log word frequency, relative position in the sentence, content vs. function word status, and surprisal (negative log-probability under GPT-2).

4 Results

4.1 Best-Head Alignment on ZuCo

Table 1 shows the best single-head Spearman ρ for each model on ZuCo (TRT). The strongest alignment comes from Llama-3.2-1B ($\rho = 0.589$, layer 6, head 6), followed by Llama-3.2-3B ($\rho = 0.573$, layer 21, head 11). All correlations are statistically significant ($p < 0.001$; Table 2).

Architecture dominates scale. A striking finding is that Llama-1B ($\rho = 0.589$) substantially outperforms GPT-2-XL ($\rho = 0.450$) despite similar parameter counts (1.2B vs. 1.5B). In fact, Llama-1B outperforms *every* GPT-2 variant, including GPT-2-large ($\rho = 0.496$). This suggests that the architectural innovations in Llama, grouped-query attention, RoPE positional encoding, and SwiGLU activations, or differences in training data and methodology produce fundamentally more human-

Table 3: Cross-dataset comparison: Best single-head Spearman ρ (TRT). The architectural advantage of Llama holds across all datasets.

Model	ZuCo (298)	Provo (133)	GECO (1,000)
GPT-2 (124M)	0.490	0.166	0.415
GPT-2-med (355M)	0.455	0.130	0.337
GPT-2-lg (774M)	0.496	0.158	0.366
Llama-1B (1.2B)	**0.589**	**0.330**	0.475
Llama-3B (3.2B)	0.573	0.272	**0.502**

like attention distributions.

Qwen models fall between the two families: Qwen-0.5B ($\rho = 0.498$) matches GPT-2-large despite having fewer parameters, and Qwen-1.5B ($\rho = 0.505$) slightly exceeds all GPT-2 variants. Since Qwen shares architectural features with Llama (GQA, RoPE, SwiGLU), this intermediate performance supports the role of architecture.

Layer localization. GPT-2 and Qwen models show best heads concentrated in the first few layers (Layer 0 for most), while Llama models exhibit best heads at intermediate depths (Layer 6 for 1B, Layer 21 for 3B). This suggests that different architectures develop human-like attention at different stages of processing.

4.2 Cross-Dataset Generalization

Table 3 shows that the alignment patterns replicate across all three corpora.

Llama models consistently outperform GPT-2 models across all three datasets. The absolute magnitudes differ across corpora: ZuCo shows the strongest correlations, likely because it averages over fewer (12) subjects with more controlled conditions, while Provo shows the weakest, possibly due to its shorter texts and different recording conditions. The ranking across models is highly consistent, confirming that these patterns are not artifacts of a particular dataset.

On GECO, Llama-3B ($\rho = 0.502$) slightly outperforms Llama-1B ($\rho = 0.475$), the reverse of ZuCo. This may reflect the longer, more naturalistic texts in GECO (full novel reading) benefiting from the larger model's capacity for long-range dependencies.

4.3 Multiple Reading Measures

Table 4 compares alignment across four eye-tracking measures on ZuCo.

Table 4: Best Spearman ρ by eye-tracking measure (ZuCo). Bold = best measure per model. TRT and nFix typically yield strongest alignment.

Model	TRT	FFD	GD	nFix
GPT-2	0.490	0.435	0.483	**0.496**
GPT-2-med	0.455	0.375	0.419	**0.460**
GPT-2-lg	0.496	0.407	0.445	**0.504**
GPT-2-XL	**0.450**	0.365	0.403	0.447
Qwen-0.5B	**0.498**	0.424	0.467	0.497
Qwen-1.5B	**0.505**	0.441	0.496	0.495
Llama-1B	**0.589**	0.527	0.582	0.582
Llama-3B	0.573	0.538	**0.578**	0.566

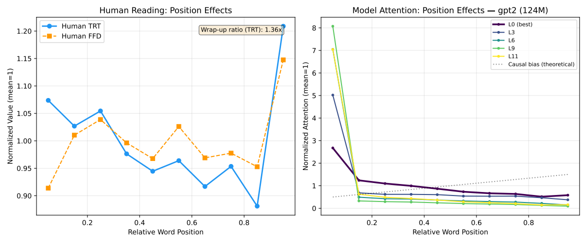

Figure 1: Position bias for GPT-2 (Layer 0, Head 4): model attention (blue) and human fixation (orange) as a function of relative word position. Both show elevated attention at sentence-final positions, paralleling the human wrap-up effect.

TRT and nFix consistently yield the strongest correlations, while FFD shows the weakest. This makes theoretical sense: TRT and nFix are cumulative measures that capture total processing effort (including regressions), while FFD reflects only the initial encounter with a word. The model's received attention, a global measure of how much all other positions attend to a given word, naturally corresponds more closely to cumulative rather than first-pass measures.

For GPT-2 models, nFix slightly outperforms TRT, while for Llama models, TRT is best. This may reflect that GPT-2's early-layer attention captures coarser word salience (how many times attention is directed to a word), while Llama's deeper attention captures more graded processing effort.

4.4 Position Bias and the Wrap-Up Effect

Causal (left-to-right) attention introduces a systematic position bias: later words mechanically receive attention from more preceding tokens. However, human readers also show increased fixation times at sentence-final positions, the well-documented "wrap-up effect" (Just and Carpenter, 1980; Rayner, 1998).

Figure 1 plots model attention and human fixation as a function of relative word position for

GPT-2's best head. Both curves show elevated values at sentence-final positions. Regression analysis confirms that position is a significant predictor ($p < 10^{-8}$) with a negative coefficient for GPT-2's best head, reflecting that the model assigns *less* attention to later words (relative to what causal masking would predict) except at sentence boundaries. This suggests that the model's best attention head has learned to counteract the causal position bias, producing a distribution that more closely resembles human reading.

4.5 Surprisal Analysis

To investigate what drives the attention–fixation alignment, we compute per-word surprisal (negative log-probability under GPT-2) and examine its relationship with both human fixation and model attention.

The Spearman correlation between GPT-2 surprisal and human TRT is $\rho = 0.566$ on ZuCo, confirming the well-established surprisal–reading-time link (Smith and Levy, 2013; Goodkind and Bicknell, 2018). We then correlate surprisal with each model's best-head attention:

- GPT-2 attention–surprisal: $\rho = 0.523$
- Llama-1B attention–surprisal: $\rho = 0.371$

GPT-2's best attention head correlates more strongly with surprisal than Llama's does. Since Llama's attention correlates *more strongly* with human fixation overall ($\rho = 0.589$ vs. 0.490), this reveals a qualitative difference: GPT-2's best head appears to track prediction difficulty (surprisal), while Llama's best head captures aspects of human reading attention that go beyond surprisal, potentially including frequency effects, syntactic processing, and other cognitive factors.

Adding surprisal to the regression model (alongside word length, frequency, position, and content/function word status) increases the explained variance from $R^2 = 0.066$ to $R^2 = 0.074$ for GPT-2, with surprisal as a significant predictor ($p < 10^{-12}$). For Llama-1B, word frequency ($p < 10^{-26}$) and position ($p < 10^{-8}$) are the strongest predictors of the attention–fixation residual, while content word status is not significant ($p = 0.14$), suggesting that the model has already implicitly captured content/function distinctions through its attention patterns.

5 Discussion

Why Llama? The consistent superiority of Llama over GPT-2 is our most striking finding. Several factors may contribute: (1) **Grouped-query attention** (GQA) forces key-value sharing across heads, potentially encouraging more diverse head specialization; (2) **RoPE positional encoding** provides more nuanced position information than GPT-2's absolute positional embeddings; (3) **Training data and scale**, Llama was trained on substantially more data with better curation; (4) **SwiGLU activation** enables more expressive intermediate representations. Disentangling these factors is an important direction for future work.

Is attention a valid comparison? The debate over attention as explanation (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) focuses on whether attention weights faithfully represent model reasoning. Our question is different: we ask whether attention weights, regardless of their explanatory power for model predictions, empirically correlate with human cognitive processing. The strong and consistent correlations we find (ρ up to 0.589) demonstrate that this correlation exists. Whether it arises because attention serves similar computational purposes in models and brains, or as a byproduct of shared statistical structure in language, remains an open question.

Causal vs. bidirectional models. Prior work (Eberle et al., 2022; Bensemann et al., 2022) focused on BERT, a bidirectional model. Our causal models face an inherent limitation: the triangular attention mask means early words can only be attended to by few tokens. Despite this, we find strong correlations, and the best heads appear to have learned to counteract position bias. This suggests that causal models develop compensatory attention strategies that, perhaps surprisingly, align with human reading patterns.

Implications for psycholinguistics. The finding that model attention captures psycholinguistic effects (word frequency, position, surprisal) without explicit training on reading data supports the view that language processing demands, shared between humans and models, shape attention allocation. The architectural dependence of this alignment suggests that not all optimization paths lead to equally human-like representations, which may inform cognitive modeling efforts.

6 Conclusion

We have presented the first comprehensive comparison of causal language model attention with human reading fixation patterns across eight models and three eye-tracking datasets. Our key findings are: (1) individual attention heads achieve strong alignment with human fixation patterns (ρ up to 0.589); (2) model architecture matters more than scale, with Llama consistently outperforming GPT-2; (3) these patterns generalize across the ZuCo, Provo, and GECO datasets; (4) model attention captures established psycholinguistic effects including word frequency, surprisal, and position effects. These results demonstrate that modern causal language models, despite being trained solely on next-word prediction, develop internal attention distributions that substantially mirror human cognitive attention during reading.

Limitations

Our study has several limitations. First, we analyze only English-language eye-tracking data; the patterns may differ for other languages, particularly those with different word order or morphological complexity. Second, we compare attention to aggregate reading measures averaged across participants, obscuring individual variation. Third, we do not analyze attention in context of specific syntactic constructions, which could reveal more fine-grained (dis)agreements. Fourth, our selection of the "best head" per model optimizes for the highest correlation, which may overestimate alignment for models with more heads. Fifth, the causal attention mask introduces a systematic position confound that, while we address analytically, is difficult to fully disentangle from genuine content-based attention patterns.

Acknowledgments

We thank the creators of the ZuCo, Provo, and GECO datasets for making their data publicly available.

References

Joshua Bensemann, Nora Hollenstein, and Alex James Peng. 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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? an analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 276–286.

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(2):602–615.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

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*, pages 4295–4309.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8.

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

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. In *Scientific Data*, volume 5, page 180291. Nature Publishing Group.

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.

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*, pages 3543–3556.

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

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Steven G Luke and Kiel Christianson. 2018. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Byung-Doh Oh and William Schuler. 2023. A comparison of self-attention and gaze fixation patterns across transformer models and human readers. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 65–74.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

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

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 11–20.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

An Yang, Baosong Yang, Binyuan Hui, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

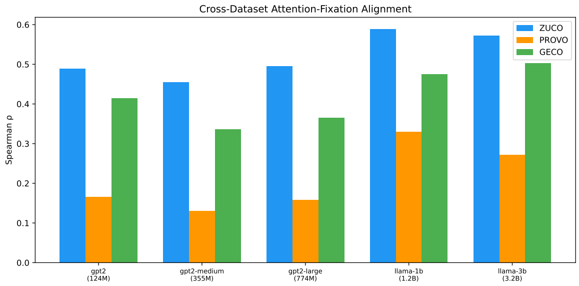

Figure 2: Cross-dataset comparison of best single-head Spearman ρ (TRT) across ZuCo, Provo, and GECO.

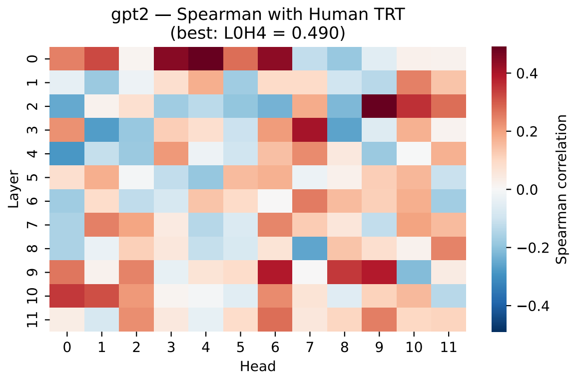

Figure 3: Layer $\times$ head Spearman ρ heatmap for GPT-2 on ZuCo (TRT). The best head is at Layer 0, Head 4.

A Detailed Cross-Dataset Results

Figure 2 shows the cross-dataset comparison as a grouped bar chart.

B Layer-Head Heatmaps

Figure 3 and Figure 4 show the full layer $\times$ head Spearman correlation heatmaps for GPT-2 and Llama-3.2-1B on ZuCo.

C Example Sentence Visualization

Figure 5 shows attention and fixation distributions for an example sentence, illustrating the word-level alignment.

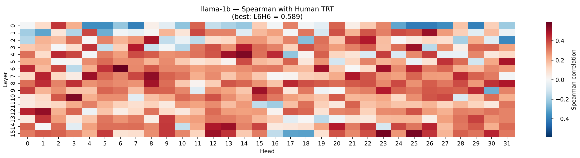

Figure 4: Layer $\times$ head Spearman ρ heatmap for Llama-3.2-1B on ZuCo (TRT). The best head is at Layer 6, Head 6, with broader high-correlation regions.

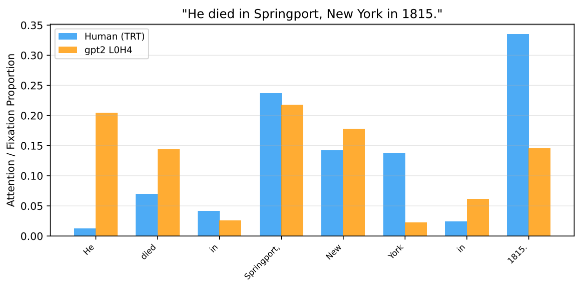

Figure 5: Example sentence from ZuCo: comparison of model attention (best head) and human fixation (TRT) distributions.