Probing Large Language Models from A Human Behavioral Perspective

Xintong Wang[♠], Xiaoyu Li[♡], Xingshan Li[♢], Chris Biemann[♠]
[♠]Department of Informatics, Universität Hamburg
[♡]School of Computer Science and Technology, Beijing Institute of Technology
[♢]Institute of Psychology, Chinese Academy of Sciences
{xintong.wang, chris.biemann}@uni-hamburg.de,
demo.xyli@gmail.com, lixs@psych.ac.cn

Abstract

Large Language Models (LLMs) have emerged as dominant foundational models in modern NLP. However, the understanding of their prediction processes and internal mechanisms, such as feed-forward networks (FFN) and multi-head self-attention (MHSA), remains largely unexplored. In this work, we probe LLMs from a human behavioral perspective, correlating values from LLMs with eye-tracking measures, which are widely recognized as meaningful indicators of human reading patterns. Our findings reveal that LLMs exhibit a similar prediction pattern with humans but distinct from that of Shallow Language Models (SLMs). Moreover, with the escalation of LLM layers from the middle layers, the correlation coefficients also increase in FFN and MHSA, indicating that the logits within FFN increasingly encapsulate word semantics suitable for predicting tokens from the vocabulary.

Keywords: Large Language Models, Interpretation and Understanding, Eye-Tracking, Human Behavioral

1. Introduction

Recent advancements in Large Language Models (LLMs) (Devlin et al., 2018; Radford et al., 2019; Touvron et al., 2023a,b) have showcased their superior capabilities in language understanding, generation as well as zero-shot transferring. Despite their remarkable successes, issues such as the generation of hallucinated (Rawte et al., 2023) and toxic outputs (Leong et al., 2023) have arisen, underscoring the importance of understanding the internal mechanisms and predictive behaviors of LLMs to develop models that are both powerful and reliable.

Research on LLM interpretation has emerged (Zhao et al., 2023; Wang et al., 2023), focusing on dissecting the components of *Feed-Forward Layers (FFN)* and *Multi-Head Self-Attention (MHSA)*. (Geva et al., 2022) highlighted the role of FFN in LLMs, demonstrating how tokens are promoted by utilizing logits in the late layers for word prediction from a vocabulary. (Bills et al., 2023) explored the activation of self-attention heads under varying prompts. Concurrently, cognition and psycholinguistic studies have documented various measures during human reading activities (Hollenstein et al., 2018, 2019; Cop et al., 2017; Luke and Christianson, 2018), closely paralleling the processes observed in language models (Hofmann et al., 2022). As depicted in Figure 1, the juxtaposition of *human reading patterns* and a *transformer block* illustrates the similarity in attention allocation—eye-tracking measurements for humans and FFN/MHSA values for LLMs—motivating our approach to probe LLMs from a human behavioral perspective.

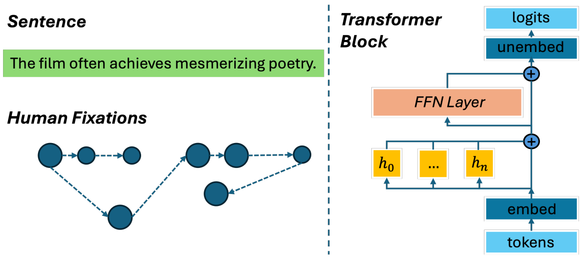

Figure 1: **Comparison of Human reading pattern and transformer block.** The left part shows the *fixation patterns* of a human reader over a given sentence, while the right part demonstrates a transformer block including *FFN layers and multi-head self-attention*. The blue dots mark fixations on the corresponding words above; a wider diameter represents a longer fixation duration.

Specifically, we investigate the **internal workings of FFN and MHSA** in LLMs, such as the GPT-2 model (Radford et al., 2019), by *correlating eye-tracking fixations with LLM values*. Our findings reveal that LLMs, particularly in their ***middle layers***, increasingly mirror human attention patterns, focusing more on essential words. However, in contrast to humans who prioritize *crucial content*, the ***upper layers*** of LLMs refine context understanding, indicating a divergence in focus on *less critical aspects*. This suggests that the outputs of FFN in the ***upper layers*** can facilitate predictions beyond just the final layers, encouraging methods for efficient semantic editing (Wang et al., 2023).

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ LREC-Coling 2024, pages 1–7
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Furthermore, our comparison of **prediction behaviors** between LLMs and Shallow Language Models (SLMs) reveals that *LLMs more closely resemble human predictive patterns*, where greater emphasis on significant words enhances the certainty of word predictions.

Our **contributions** are as follows:

- We conduct a detailed analysis of the internal mechanisms of FFN and MHSA in LLMs from a human behavioral perspective.
- We juxtapose the word prediction processes of LLMs and SLMs, reinforcing the evidence that LLMs more closely align with human attention patterns, focusing on crucial words to enhance prediction certainty.

2. Related Work

Human Behavior Measures: Studies in cognition and psycholinguistics have deployed simultaneous eye-tracking and electroencephalography during natural and task-specific reading to comprehend human reading processes. Noteworthy datasets in this context include ZuCo 1.0 (Hollenstein et al., 2018), ZuCo 2.0 (Hollenstein et al., 2019), GECO (Cop et al., 2017), and Provo (Luke and Christianson, 2018). However, to the best of our knowledge, there is a paucity of work utilizing these datasets to probe LLMs and their internal mechanisms.

Eye-movement Prediction: A shared task at ACL 2021 (Hollenstein et al., 2021) involved using language models for predicting eye-movement measures. In this shared task, models, including Boosting, MLP, and RoBERTa, displayed significant performance in this task. Besides, linguistic features proved crucial for achieving superior results (Bestgen, 2021). In this paper, we focus on employing eye-movement data for probing LLMs.

3. Preliminary

Large language models (LLMs) predominantly rely on the Transformer architecture (Vaswani et al., 2017), composed of Transformer blocks acting as layers denoted by $l = 1, 2..., L$. As shown in Figure 1, each Transformer block primarily consists of **multi-heads self-attention** and a **feed-forward network**. The motivation for the multi-head self-attention mechanism lies in its ability *to extract various aspects of the sequence, with its capacity deepening with the increase of layers*. Concurrently, the FFN serves to *output for the current layers and makes prediction over a vocabulary.*

More specifically, in layer l, the currently processed representation is denoted by X_i^l, and the output for FFN is computed as:

$$o_i^l = FFN^l\left(X_i^l\right), \qquad (1)$$

where o_i^l denotes the output for the current FFN.

An updated representation $\tilde{x}_i^l$, is then achieved by adding X_i^l and o_i^l. The updated representation, $\tilde{x}_i^l$, subsequently undergoes a self-attention process. Given the presence of multi-head self-attention in each layer, all the representations in each self-attention head are concatenated to serve as the input for the subsequent FFN layer, as illustrated below:

$$X_i^{l+1} = \mathrm{concatenate}\Big(\text{Attention}^l\left(\tilde{x}_i^l\right)\Big), \qquad (2)$$

In this work, we present empirical evidence understanding the function of multi-head self-attention and FFN layers by correlating their values with human behavioral data, eye-tracking measurements.

4. Eye-tracking Measurements

Human behavioral signals, such as **eye-tracking, fMRI, and EEG**, have been widely utilized in cognition and psycholinguistic studies. Among these signals, eye-tracking offers millisecond-precise recordings of gaze direction, illuminating the *focus of attention during reading and comprehension*. This process bears resemblance to the operations within a **transformer block**, as depicted in Figure 1. Thus, we employ ***eye-tracking data*** to uncover the internal mechanics of the transformer architecture.

Eye-movement Measures	**Abbrev.**	**Definition**
Gaze duration	GD	The sum of all fixations on the current word in the first-pass reading before the eye moves out of the word
Total reading time	TRT	The sum of all fixation durations on the current word, including regressions
First fixation duration	FFD	The duration of the first fixation on the prevailing word
Single fixation duration	SFD	The duration of the first and only fixation on the current word
Go-past time	GPT	The sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word

Table 1: **Definition of Five Eye-tracking Measures**: Gaze Duration (GD), Total Reading Time (TRT), First Fixation Duration (FFD), Single Fixation Duration (SFD), and Go-Past Time (GPT).

In our study, we establish correlations between metrics derived from *multi-head self-attention (MHSA), feed-forward neural (FFN) layers*, and **five specific eye-tracking measurements**: *Gaze Duration (GD), Total Reading Time (TRT), First Fixation Duration (FFD), Single Fixation Duration (SFD), and Go-Past Time (GPT).* Each of these metrics offers unique insights into the human reading process. For instance, Gaze Duration (GD) refers to the cumulative duration of all fixations on a given word during initial reading before moving to the next word, with *longer durations indicating the word's significance*. Similarly, Total Reading Time (TRT) encompasses all fixation durations on a word, including regressions, indicating that *readers may revisit a word multiple times to refine their understanding*. The detailed meanings of these eye-tracking measures can be found in Table 1.

By leveraging these interpretable eye-tracking metrics, we aim to probe LLMs by correlating their values with those observed in multi-head attention and FFN layers.

5. Experiments

5.1. Experimental Settings

Language Models: For our investigation, we utilized a pre-trained GPT-2 model (*base*) from HuggingFace, focusing on analyzing the internal mechanisms of FFN and multi-head self-attention mechanisms due to its **simplicity and general applicability**. We posit that our probing method is adaptable and can be extended to other, more advanced open-source LLMs such as LLaMA (Touvron et al., 2023a) and Qwen (Bai et al., 2023), among others. Additionally, we broaden our analysis to include **Shallow Language Models (SLMs)** like N-Gram language models (Pauls and Klein, 2011), Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTM) networks (Sherstinsky, 2020), and a recently enhanced RNN variant, the RWKV-V4 model (Peng et al., 2023), to conduct a comprehensive comparison of prediction probabilities. For the training of SLMs, we employ the WikiText-103 dataset.

Eye-tracking Data: For human behavioral data, we utilize the ZuCo 2.0 dataset (Hollenstein et al., 2019), which contains concurrent eye-tracking records captured during two types of reading activities: *natural reading (NR) and task-specific reading (TSR)*. This dataset is notably comprehensive, comprising 730 English sentences, split into 349 sentences read under normal conditions and 390 sentences read under a task-specific paradigm. Eye-tracking data from 18 participants were recorded during both NR and TSR activities. We conducted word prediction experiments using various language models on sentences from the ZuCo 2.0 dataset to then analyze the correlation patterns between human reading behaviors and language model predictions.

Correlation Metrics and Evaluation: Following previous studies (Eberle et al., 2022) on analyzing the prediction behavior of LLMs, we also employ three prevalent correlation metrics: Pearson (Freedman et al., 2007), Spearman (Caruso and Cliff, 1997), and Kendall (Abdi, 2007), to investigate the relationship between values derived from LLMs and human behavioral measures. Despite minor differences, we find these correlation metrics yield similar results. Among them, Spearman exhibits superior robustness when compared to Pearson and Kendall. Unless stated otherwise, experimental results are reported using Spearman analysis. *Given that larger fixations, as indicated by various eye-tracking measures, signify the importance of the current word,* ***a stronger correlation implies that LLMs also allocate more attention to the processed word.***

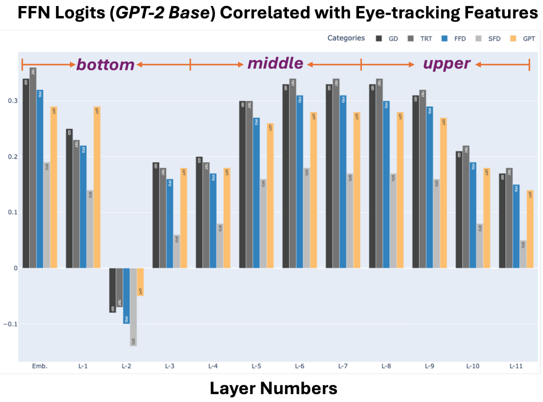

Figure 2: **FFN Correlation Values.** FFN values through layers in GPT-2 *base* Correlated with five different eye-tracking features in three groups: bottom, middle, and upper. (Significant at $p < 0.05$)

5.2. FFN Correlation Analysis

We examine the functions of the FFN within GPT-2. To elucidate our findings, we categorize the 12 layers of GPT-2 (*base*) into three groups: **bottom ($l_1 \rightarrow l_4$), middle ($l_5 \rightarrow l_8$), and upper ($l_9 \rightarrow l_{12}$).** As illustrated in Figure 2, the bottom most layers show a direct correlation between the embedding of input tokens and human reading fixations. This suggests that humans require more time to comprehend critical tokens that are also reflected in the embeddings of LLMs. This correlation diminishes as we ascend through the layers, with the topmost layer of the bottom group (Layer 3) indicating a divergence in processing tokens from human behavior; the FFN at this level begins to process tokens yet in a manner distinct from human reading patterns.

Progressing to the middle layers, the correlation coefficients initially increase and then stabilize, peaking at Layer 6. This pattern suggests that the FFN in these middle layers starts to show similar human fixation behaviors, indicating that the logits within FFN increasingly encapsulate word semantics suitable for predicting tokens from the vocabulary.

Intriguingly, in the upper layers, we observe a decline in correlation values. We hypothesize that at this stage, the LLM begins to incorporate less critical words within sentences into its considera-

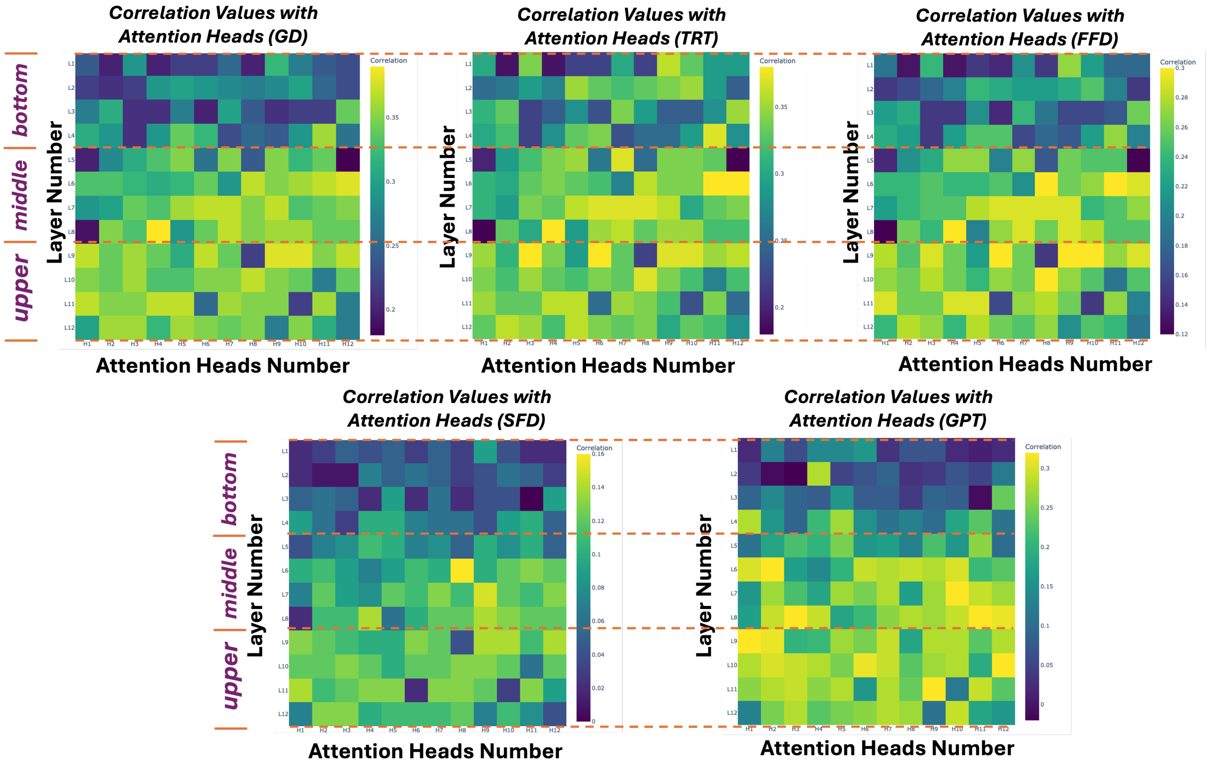

Figure 3: **Attention Heads Correlated Values with Eye-tracking Measurements through Layers Results.** ***Lighter and larger values*** signify stronger correlations.

tion, diverging from human intuition, which tends to focus on the most crucial aspects of the context and disregard less important information.

5.3. Multi-head Self-attention Correlation

Figure 3 presents heatmaps that illustrate the correlation between the values of 12 self-attention heads across 12 distinct layers and human behavioral data; where lighter and larger values signify stronger correlations. Similar to our FFN analysis, we categorized the 12 layers into three groups: **bottom, middle, and upper**. The bottom group exhibits a weaker correlation with human fixations, indicating that while self-attention mechanisms begin to process tokens at this stage, they do so differently from human behavior.

As we ascend through the middle and upper groups, we observe an increase in correlation across different layers and attention heads with human fixations. This pattern suggests that, in these layers, LLMs begin to align more closely with human patterns, especially in focusing on important contextual tokens. Notably, unlike in the FFN analysis, we did not observe a decrease in multi-head attention correlation values in the upper layers. This difference implies that the comprehension capabilities of LLMs are progressively refined up to the final layer, enabling more diverse and accurate word predictions compared to human reading patterns.

Furthermore, among the five eye-tracking measures analyzed, Gaze Duration (GD), Total Reading Time (TRT), First Fixation Duration (FFD), and Go-Past Time (GPT) demonstrate stronger correlations, whereas Single Fixation Duration (SFD) shows a weaker correlation. Given that SFD represents the first and only fixation on a current word—suggesting lesser importance—while GD, TRT, FFD, and GPT include regressions on significant words, this discrepancy explains why LLMs also prioritize these important words.

5.4. Prediction Probability Correlation

We further analyze word prediction probability behaviorals in LLMs and our investigation into the correlation of word prediction probabilities reveals distinct behaviors between Large Language Models (LLMs) and Shallow Language Models (SLMs). For this analysis, we employed two reading tasks: task-specific reading (TSR) and natural reading (NR). The TSR task encompassed 5335 words for prediction analysis, while the NR task included 5329 words. Our findings, detailed in Table 2, are divided into two parts: the upper section presents the correlation outcomes for the TSR task, and the lower section for the NR task.

Overall, SLMs exhibit a notable and consistent **negative correlation** in both the TSR and NR tasks. This trend suggests that SLMs tend to assign higher prediction probabilities with fewer fixations on critical words, thereby increasing the uncertainty of word predictions. In contrast, LLMs, exemplified by GPT-2, demonstrate a significant and **positive correlation** in both tasks. This positive correlation indicates that LLMs exhibit a prediction pattern akin to human behavior, where increased attention to crucial words leads to more confident predictions.

Though the aforementioned conclusions are consistent for both the TSR and NR tasks, it is noteworthy that the correlation values for the NR task are consistently higher than those for the TSR task. We hypothesize that during task-specific readings, humans are guided by specific clues to identify and concentrate on words that are pertinent to the task at hand. Consequently, our word prediction analysis across different LMs aligns more closely with the process in NR.

Model	**Eye-tracking Measures**				
	GD	TRT	FFD	SFD	GPT
			Task-specific Reading		
N-Gram	-0.26	-0.25	-0.23	-0.15	-0.23
RNN	-0.44	-0.43	-0.41	-0.28	-0.40
GRU	-0.46	-0.45	-0.43	-0.30	-0.43
LSTM	-0.42	-0.41	-0.39	-0.26	-0.39
RWKV	-0.39	-0.40	-0.40	-0.27	-0.33
GPT-2	0.23	0.21	0.20	0.12	0.28
			Natural Reading		
N-Gram	-0.33	-0.33	-0.31	-0.15	-0.29
RNN	-0.52	-0.51	-0.50	-0.26	-0.46
GRU	-0.54	-0.53	-0.52	-0.29	-0.48
LSTM	-0.52	-0.50	-0.49	-0.26	-0.46
RWKV	-0.39	-0.39	-0.38	-0.19	-0.28
GPT-2	0.33	0.30	0.30	0.14	0.37

Table 2: **Prediction Probability Correlation Results** using Spearman correlation metric. The numbers in blue mean the significant negative correlation, while the red represent the positive correlation. (Significant at $p < 0.05$)

6. Conclusion

In this work, we probe LLMs through human behavior, specifically employing eye-tracking measurements to dissect the internal workings of LLMs, including the feed-forward layers and multi-head attention. Our findings reveal a similarity between LLMs and humans on word prediction: both exhibit a tendency where heightened attention to pivotal words results in more confident predictions. Our analysis further delineates that feed-forward networks begin to align with human fixation patterns starting from the middle layers, leveraging upper layers to broaden the contextual understanding. Our probing approach stands out for its interpretability from human reading indicators and paves the way for the development of LLMs that are not only reliable but also imbued with a greater degree of trustworthiness.

7. Acknowledgements

This research was funded by the German Research Foundation DFG Transregio SFB 169: Crossmodal Learning: Adaptivity, Prediction, and Interaction.

8. Bibliographical References

Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yves Bestgen. 2021. Last at cmcl 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. *arXiv preprint arXiv:2104.13043*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *https://openai.com/research/language-models-can-explain-neurons-in-language-models*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

John C Caruso and Norman Cliff. 1997. Empirical size, coverage, and power of confidence intervals for spearman's rho. *Educational and psychological Measurement*, 57(4):637–654.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eye-tracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49:602–615.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:214.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78.

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. *Scientific data*, 5(1):1–13.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Chak Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4433–4449.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Honzaernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*.

Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. 2023. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

9. Ethical Considerations

The eye-tracking data employed in this study, derived from the ZuCo 2.0 dataset, are publicly accessible and adhere to established ethical protocols.