Probing Language Models from A Human Behavioral Perspective

Xintong Wang[♠] and **Xiaoyu Li**[♡] and **Xingshan Li**[♢] and **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 process and internal mechanisms, such as feed-forward networks and multi-head self-attention, remains largely unexplored. In this study, we probe LLMs from a human behavioral perspective, correlating values from LLMs with eye-tracking measures, which are widely recognized as meaningful indicators of reading patterns. Our findings reveal that LLMs exhibit a prediction pattern distinct from that of RNN-based LMs. Moreover, with the escalation of FFN layers, the capacity for memorization and linguistic knowledge encoding also surges until it peaks, subsequently pivoting to focus on comprehension capacity. The functions of self-attention are distributed across multiple heads. Lastly, we scrutinize the gate mechanisms, finding that they control the flow of information, with some gates promoting, while others eliminating information.

1 Introduction

Modern Large Language Models (LLMs) (Devlin et al., 2018; Radford et al., 2019; Touvron et al., 2023) have demonstrated remarkable success as foundational models in generalization. These models, built upon transformers (Vaswani et al., 2017), primarily consist of two core components: the feed-forward network and multi-head self-attention. How do LLMs construct word prediction and what are the internal functions of their core components? We approach this question from the perspective of human behavior.

Cognition and psycholinguistic studies often record measures while humans engage in natural or task-specific reading (Hollenstein et al., 2018, 2019; Cop et al., 2017; Luke and Christianson, 2018). These well-defined measures align closely with the processes of language models (LM) (Hofmann et al., 2022). Our investigation commences with the examination of word prediction in different LMs, from statistical N-Gram LM (Pauls and Klein, 2011) to RNN-based LMs (Mikolov et al., 2010), and finally, LLMs like RKVW (Peng et al., 2023) and GPT-2 (Radford et al., 2019). Notably, RNN-based LMs exhibit a pattern most akin to human word prediction. Words that pose difficulty for humans also challenge these models. Interestingly, LLMs like GPT-2 demonstrate a distinct pattern with a positive correlation to human behavioral data. The GPT-2 model boasts powerful capabilities, with an increase in the probability of predicting challenging words, compared to other LMs.

We probe each layer to comprehend the functions of FFN and multi-head self-attention. FFN layers appear to serve as memory units that encode linguistic knowledge toward word prediction, with their ability increasing with the layer count until peaking around layer 8. Beyond this point, the FFN emphasizes improving comprehension. On the other hand, multi-head self-attention is found to be distributed. These attention heads work to promote crucial information, their ability also peaking around layer 8, thus reinforcing the function of FFN.

In addition to probing LLMs, we correlate hidden states, memory cells and gates with human behavioral data, enabling an understanding of the function of memory units and gate mechanisms in RNN-based models. Our findings indicate that the hidden memory and memory cells are utilized in semantic processing and comprehension. The gates in GRU and LSTM operate to control information, promoting or eliminating.

In conclusion, we delve into the construction process of word prediction in LMs, and the mechanisms of gates, FFN, and multi-head self-attention. Our findings illuminate the internal workflow of LLMs, contributing to future work on their interpretability, control, and efficient training.

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 LMs for predicting eye-movement measures. Several models, including Boosting, MLP, and RoBERTa, displayed significant performance in this task. Linguistic features proved crucial for achieving superior results (Bestgen, 2021). In this paper, we focus on employing behavioral data for probing LLMs.

3 Language Models and Word Prediction

Modern language models are predominantly transformer-based models designed to predict the subsequent token given a context. The primary components of these LMs comprise Feed-Forward Networks (FFN) and Multi-Head Self-Attention mechanisms. In this section, we delineate LMs from the standpoint of word prediction, while underscoring key mechanisms such as the gate and self-attention. We commence with a statistical model, the N-Gram LM (Pauls and Klein, 2011), followed by RNN-based LMs (Mikolov et al., 2010), and subsequently, the self-attention-based LM, GPT-2 (Radford et al., 2019).

Given a sequence of input tokens, denoted as $\mathbf{w} =< w_1, w_2, ..., w_n >$, an n-gram represents a sequence of n words. Upon applying the Markov assumption, which approximates the context history for the current predicted token to $n-1$, the conditional probability for the current token can be expressed as: $P(w_n \mid w_{1:n-1}) \approx P(w_n \mid w_{n-N+1;n-1})$, where N signifies the parameter of the n-gram, and n indexes the predicted token. To calculate the n-gram probability, we estimate the parameters of an n-gram model by counting the co-occurrences, denoted as $C(w_{n-N+1;n-1})$, within the corpus. The **N-Gram LM**, being a statistical LM, is straightforward to implement. As the value of N increases, the precision of the N-Gram LM also escalates. However, this brings about challenges related to storage.

Prior to the emergence of Transformer-based language models, **RNN-based language models** were the prevalent choice for numerous tasks due to their superior ability to model temporal information. The vanilla RNN retains all previous sequence context in the hidden state as follows:

$$\mathbf{H}_t = \phi\left(\mathbf{X}_t\mathbf{W}_{xh} + \mathbf{H}_{t-1}\mathbf{W}_{hh} + \mathbf{b}_h\right) \quad (1)$$

where $X_t = E_t \cdot w_t \in \mathcal{R}^{|\mathcal{V}|\times d}$ represents the word embeddings, $\mathcal{V}$ is the vocabulary, and d is the dimension. However, as the length of a sequence expands, the hidden state struggles to retain all historical information, leading to gradient vanishing.

Motivated by human's ability to selectively remember and forget information, GRU and LSTM models were proposed, incorporating gate mechanisms. The computation of these gates is reliant on the current input token and the previous hidden state. To streamline notation, we use G_t to denote different gates as shown in the following equation:

$$\mathbf{G}_t = \sigma\left(\mathbf{X}_t\mathbf{W}_x + \mathbf{H}_{t-1}\mathbf{W}_h + \mathbf{b}\right) \quad (2)$$

GRU has two gates: reset gate and update gate, while LSTM includes forget gate, input gate, and output gate. The update gate in GRU is applied to the candidate hidden state. The final hidden state further synchronizes the current input with the preceding context, as depicted in the equation.

$$\begin{aligned}\tilde{\mathbf{H}}_t &= \tanh\left(\mathbf{X}_t\mathbf{W}_{xh} + \left(\mathbf{R}_t \odot \mathbf{H}_{t-1}\right)\mathbf{W}_{hh} + \mathbf{b}_h\right)\\ \mathbf{H}_t &= \mathbf{Z}_t \odot \mathbf{H}_{t-1} + \left(1 - \mathbf{Z}_t\right) \odot \tilde{\mathbf{H}}_t\end{aligned} \quad (3)$$

In LSTM, an extra memory unit, termed the memory cell, is introduced alongside the hidden states. The forget gate and input gate are directly applied to these memory units. Furthermore, the final hidden states are computed by multiplying the output gate with the current memory cell, as demonstrated in the subsequent equation:

$$\begin{aligned}\mathbf{C}_t &= \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t\\ \mathbf{H}_t &= \mathbf{O}_t \odot \tanh\left(\mathbf{C}_t\right)\end{aligned} \quad (4)$$

Finally, the output probability distribution is derived from the hidden state computed by various models:

$$\mathbf{y}_t = \mathrm{softmax}\left(\mathbf{H}_t\right) \quad (5)$$

Overall, the gate mechanisms significantly enhance the efficiency of RNN-based LMs, enabling the

model to selectively focus or forget, ensuring crucial information is smoothly transmitted to subsequent sequences. *In this study, we elucidate the clear functions of these engineered operations by correlating them with human behavioral data.*

Large language models (LLMs) predominantly rely on the Transformer architecture, composed of Transformer blocks acting as layers denoted by $l = 1, 2..., L$. Each Transformer block primarily consists of a multi-head 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:

$$o_i^l = FFN^l\left(X_i^l\right) \tag{6}$$

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:

$$X_i^{l+1} = \text{concatenate}\left(\text{Attention}^l\left(\tilde{x}_i^l\right)\right) \tag{7}$$

In this work, we present empirical evidence supporting the function of multi-head self-attention and FFN layers by correlating their representations with human behavioral data. Intriguingly, we provide substantiation of why FFN can be manipulated and promote concepts and timing of such manipulation as proposed in (Geva et al., 2022).

4 Experiments

4.1 Correlation Metrics

We 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 LMs 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. Interestingly, we discovered that applying a `log10` transformation to the raw values from LMs, considerably enhances the robustness of the results. Unless stated otherwise, experimental results are reported using Spearman analysis **without** `log10` normalization.

4.2 Datasets

We employ WikiText-103 (Merity et al., 2016) for training RNN LM, GRU LM, and LSTM LM. The WikiText-103 is a tokenized corpus encompassing 28,475, 60, and 60 articles in the training, validation, and test sets respectively. To address the sparsity issue encountered when training an N-Gram LM, we extract 4G data from the original Wikipedia corpus[1]. For human behavioral data, we utilize the ZuCo 2.0 dataset, which comprises concurrent eye-tracking and electroencephalograph records during natural reading (NR) and task-specific reading (TSR). More precisely, ZuCo 2.0 includes 730 English sentences, of which 349 are under a normal reading paradigm and 390 under a task-specific paradigm. These are coupled with eye-tracking and EEG data recorded during the NR and TSR process from 18 participants.

4.3 Language Model Implementations

In our experiments, we trained an N-Gram Language Model using a Wikipedia corpus comprised of 4GB, equivalent to 40 million sentences, with the parameter N set to three. We observed that as the data scale reaches 4GB, the precision of the N-Gram model achieves stability.

Subsequently, we implemented two-layer RNN-based Language Models using the WikiText-103 training set, based on the PyTorch open-source project [2]. The dimensions for both the embedding and hidden state were set to 200. The initial learning rates were configured as 2.0 for the RNN and 20 for both GRU and LSTM, with weight decay. The batch size was set at 20, gradient clipping at 0.25, and a dropout rate of 0.2.

For Large Language Models (LLMs), we employed a pre-trained GPT-2 model from Huggingface[3] to analyze the internal workings of the FFN and multi-head self-attention. To present a comparative analysis of LLMs, we incorporated the recently proposed pre-trained RWKV-V4 model [4],

[1] https://dumps.wikimedia.org/wikidatawiki/

[2] https://github.com/pytorch/examples/tree/main/word_language_model

[3] https://huggingface.co/gpt2

[4] https://github.com/BlinkDL/RWKV-LM/tree/main/RWKV-v4

an RNN model delivering Transformer-level LLM performance, specifically designed to model temporal information.

4.4 Prediction Probability Correlation

In our experiment, we establish the correlation between the predicted word probability derived from various LMs and five distinct eye-tracking measures, namely GD, TRT, FFD, SFD, and GPT, all obtained from the ZuCo 2.0 dataset. The definitions for these eye-tracking measures can be found in Table 1. We consider these five measures collectively for a comprehensive probe of LMs.

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 for Eye-tracking Measures

The TSR task consists of 5335 words for prediction, while the NR task contains 5329 words. Table 2 illustrates our results: the upper presents the correlation outcomes in the TSR task, and the lower does the same for the NR task. In general, all LMs, except GPT-2, demonstrate noticeable and robust negative correlations in both the TSR and NR tasks. This can be interpreted as a higher reading time for humans correlating with increased prediction difficulty for the LMs. The correlation in NR tasks outperforms that in TSR tasks, potentially due to the LMs' training being more akin to the NR task structure.

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 (Significant at $p < 0.05$)

Specifically, the N-Gram model shows correlation with human behavioral data, likely due to its shared frequency-based feature with human linguistic knowledge. Notably, RNN-based LMs exhibit an exceptional correlation of over 0.5 with human behavioral data, suggesting a significant similarity with human patterns. The GRU model delivers superior results compared to RNN and LSTM.

Interestingly, RWKV LLM indicates negative correlations while GPT-2 shows positive. Given the powerful capabilities of GPT-2, we observe an increase in the probability of hard-predicted words compared to other LMs. This discrepancy might arise from *LLMs confidently predicting words that humans take longer to process, indicating the diverse prediction capacity of LLMs, where the probability of easy-to-predict words decreases, while that of hard-to-predict words increases*.

4.5 Probing Gate Mechanism

We then probe the functions of memory units - hidden states and memory cells, as well as the gate mechanism in GRU and LSTM. As depicted in Table 3, the first layer's hidden states of RNN exhibit a robust positive correlation with human behavioral data, while the second layer does not show significant correlation patterns. *It appears that the first layer's hidden states in the RNN process raw linguistic information of sequences* that directly map onto human behavioral patterns in both TSR and NR tasks. On the other hand, *the deeper hidden states of RNN are engaged in comprehension in the TSR task*, requiring more reasoning ability. The correlation results for the *RNN's deeper layer remain positive in NR, illustrating further syntactic processing and comprehension abilities*.

RNN Model	Eye-tracking Measures				
	GD	TRT	FFD	SFD	GPT
Task-specific Reading					
H-L1	0.53	0.51	0.51	0.41	0.49
H-L2	−0.09	−0.11	−0.09	−0.04	−0.04
Natural Reading					
H-L1	0.56	0.54	0.55	0.35	0.52
H-L2	0.42	0.41	0.4	0.23	0.41

Table 3: RNN Hidden State from Different Layers Correlation Results (Significant at $p < 0.05$)

In both GRU and LSTM models shown in Tables 4 and 5, we observed no discernible correlation between hidden states, memory cells, and human behavioral data. However, the gate mechanisms exhibited a strong correlation, leading us to find

that *in GRU and LSTM, memory units primarily respond to comprehension synthesizing, while the gates handle contextual processing and information flow.*

GRU Model	Eye-tracking Measures				
	GD	TRT	FFD	SFD	GPT
Task-specific Reading					
H-L1	0.07	0.06	0.05	0.03	0.15
H-L2	−0.02	−0.01	−0.01	0.00	−0.11
Reset Gate	−0.43	−0.42	−0.41	−0.32	−0.41
Update Gate	−0.46	−0.45	−0.45	−0.34	−0.43
Candidate Hidden State	−0.47	−0.45	−0.44	−0.33	−0.46
Natural Reading					
H-L1	0.15	0.14	0.15	0.08	0.24
H-L2	−0.09	−0.09	−0.09	−0.04	−0.20
Reset Gate	−0.44	−0.42	−0.42	−0.25	−0.41
Update Gate	−0.48	−0.46	−0.46	−0.27	−0.43
Candidate Hidden State	−0.56	−0.54	−0.54	−0.33	−0.53

Table 4: GRU States from Different Layers Correlation Results (Significant at $p < 0.05$)

LSTM Model	Eye-tracking Measures				
	GD	TRT	FFD	SFD	GPT
Task-specific Reading					
H-L1	0.01	0.0	0.0	0.0	0.12
H-L2	0.01	0.0	−0.01	−0.01	0.1
C-L1	−0.18	−0.23	−0.2	−0.09	−0.04
C-L2	−0.01	−0.01	−0.01	−0.01	−0.09
Input Gate	−0.48	−0.46	−0.46	−0.37	−0.45
Forget Gate	0.29	0.28	0.28	0.24	0.31
Candidate Cell State	−0.38	−0.36	−0.35	−0.26	−0.39
Output Gate	0.46	0.46	0.44	0.34	0.43
Natural Reading					
H-L1	0.08	0.06	0.08	0.05	0.2
H-L2	0.08	0.08	0.08	0.02	0.19
C-L1	−0.01	−0.09	−0.01	0.1	0.13
C-L2	−0.04	−0.03	−0.04	0.0	−0.14
Input Gate	−0.5	−0.48	−0.48	−0.31	−0.45
Forget Gate	0.31	0.29	0.3	0.2	0.32
Candidate Cell State	0.45	0.43	0.43	0.24	0.45
Output Gate	0.52	0.51	0.51	0.32	0.47

Table 5: LSTM States from Different Layers Correlation Results (Significant at $p < 0.05$)

In the case of both TSR and NR tasks within the GRU model, the reset gate, update gate, and candidate hidden states all exhibited negative correlation results. This indicates that the more time human readers spent on current words, the smaller the gate values and the other way around. In the LSTM model, the input gate demonstrated a negative correlation, while both the forget gate and output gate showed positive correlations. An interesting anomaly observed was the reverse correlation of the candidate cell state in LSTM between TSR and NR tasks, which warrants further investigation.

In conclusion, our analysis underscores the importance of the gate mechanism in handling contextual processing and information flow in GRU and LSTM. The functionalities of the gates can be viewed as promotion and elimination of tokens. Certain gates promote tokens, whereas others are tasked with eliminating them.

4.6 FFN and Multi-head Self-attention

Finally, we scrutinize the roles of FFN and multi-head self-attention. As delineated in Figure 1, we establish a correlation between FFN output at each layer and the corresponding human behavioral data. It was noted that the embedding of input tokens displayed a direct correlation with human reading times. With escalating layers, the correlation coefficients initially rise, only to witness a minor dip later. This suggests that *the proficiency in processing syntactic and semantic elements gradually amplifies until the pinnacle at layer 8. Post layer 8, the FFN primarily concentrates on enhancing comprehension skills*. Each output of FFN exhibits potential in predicting words over a vocabulary, thereby supplying empirical substantiation for the work (Geva et al., 2022), where FFN is used to promote tokens. Moreover, FFN output succeeding layer 8 appears more apt for word generation, while the concluding layers seem well-suited for tasks involving reasoning and comprehension.

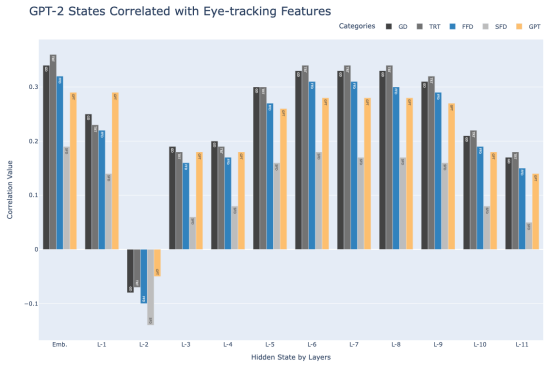


Figure 1: FFN through Layers in GPT-2 Correlated with Eye-tracking Features (Significant at $p < 0.05$)

Figure 2 offers heatmaps delineating the correlation between 12 self-attention heads values across 12 distinct layers and the human behavioral data; lighter and larger values indicate stronger correlations. Initially, multi-head self-attention was devised to model disparate facets of the input sequence. *Intriguingly, we observe that the functions are allocated across varying attention heads*. As the layers amplify, the correlation coefficients show an organic augmentation, corroborating our findings in the probe of FFN layers. Of note, a minority of attention heads in higher layers display smaller values. This could be attributed to the subjective bias inherent in eye-tracking experiments. The

Zuco 2.0 dataset is a moderately-sized compilation involving 18 participants. Increasing the participant pool could potentially alleviate the impact of individual biases.

5 Conclusion

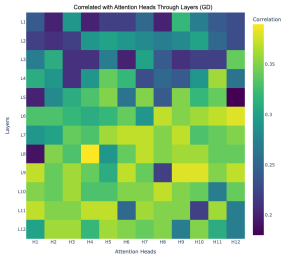
Understanding the process of word prediction and the internal workings of LLMs is crucial for explainability and for further strengthening powerfulness and eliminating the hardness of LLMs. Our study probes LLMs from a human behavior perspective using eye-tracking measures. We found that RNN-based models had the most similar pattern as humans for word prediction, unlike LLMs. Further, we discovered that in RNN-based LMs, hidden states and memory cells serve as comprehensive units, with gates directing information flow. Some gates are prompting information, while others are eliminating. Analysis of FFN and self-attention reveals that FFN directly maps to the prediction with a peak and then enhances comprehension. The functions of self-attention are distributed across multiple heads.

Limitations

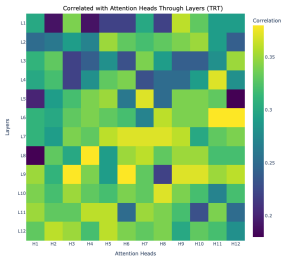
Our work primarily concentrates on probing Large Language Models (LLMs) using human behavioral data. During the interpretation of the multi-head self-attention mechanism, we encountered individual bias. Future research could harness a larger volume of eye-tracking data to counteract this bias. Additionally, including EEG data could allow us to probe LLMs from a brain activation standpoint.

In terms of LLMs, we selected RWKV and GPT-2 as our baselines for understanding. RWKV, a recently proposed RNN-style model, warrants further exploration to comprehend the internal mechanisms that enable it to achieve performance comparable to the GPT model. Additionally, with the release of the LLaMA model, which boasts an even larger parameter set, future work could leverage the LLaMA model to elucidate the emergent phenomena of LLMs using human behavioral data.

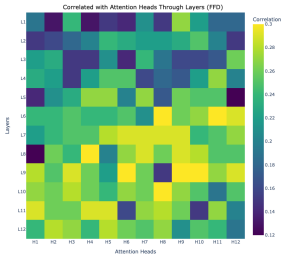
Lastly, we employed various eye-tracking measures collectively to probe LLMs. In future endeavors, we aim to examine these measures individually to better leverage the definition of each measure, thereby enhancing our understanding of the reasoning line.



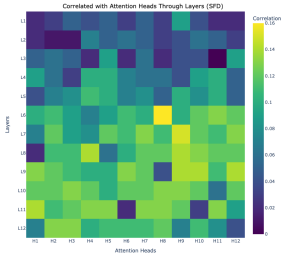
(a) Attention Heads Correlated Results (GD)



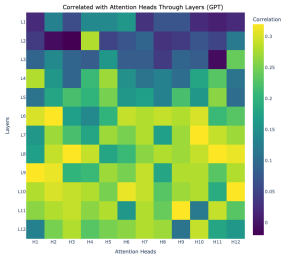
(b) Attention Heads Correlated Results (TRT)



(c) Attention Heads Correlated Results (FFD)



(d) Attention Heads Correlated Results (SFD)



(e) Attention Heads Correlated Results (GPT)

Figure 2: Attention Heads through Layers Correlated Results (Significant at $p < 0.05$)

Acknowledgements

We thank Markus Hofmann, Ralph Radach, and Liang Ding for helpful feedback and construction suggestions. This research was funded by the German Research Foundation DFG Transregio SFB 169: Crossmodal Learning: Adaptivity, Prediction and Interaction.

References

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

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*.

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.

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.

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*.

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.

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*.

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.

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

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.