Attention Redistribution During Event Segmentation In Large Language Model

Yuyang Xu
Department of Psychology
New York University
New York, NY 10012
yx2227@nyu.edu

Qingqing Yang
Department of Psychology
The Ohio State University
Columbus, OH 43210
yang.6118@osu.edu

Abstract

Human beings perceive a continuous string of experiences by segregating the experience into discrete events. Recently, it has been proven that a large language model can segregate events similarly to humans, even though the model is not specifically trained to do so. In this research, we used naturalistic stimuli like stories to explore the underlying changes in the attention mechanisms when large language model performs event segmentation. We discovered a redistribution of attention outputs toward words that play different roles in structuring an event. We found that the model enhances attention directed toward words indicative of potential changes in elements like time, space, objects, and goals in a continuous narrative. The model also reduces attention directed toward other kinds of words not indicative of such change. Our results provide better insights into the underlying processes of the high-level cognitive features in large language models and in the human brain.

1 Introduction

The stimulus received by the human sensory systems is continuous. Yet, when processing a long string of continuous incoming stimuli, humans often take the stimulus and perceive them as discrete events (Zacks et al., 2007). The ability to segment continuous stimuli into events is crucial to various cognitive processes (Jafarpour et al., 2022; Bangert et al., 2020; DuBrow and Davachi, 2016). Recently, it has been demonstrated that large language models (LLMs) like GPT-3 can segment events similarly to humans (Michelmann et al., 2023). Although the large language models are initially trained for next-word prediction, the ability of event segmentation emerges after extensive training naturally (Radford et al., 2019). This effect can be viewed as a convergent evolution; the challenge of language processing might force the human brains and LLMs to converge on similar properties (Waldrop, 2024). However, the exact mechanisms behind this capability's development remain an open question. Specifically, it is unclear how the changes in the inner function evolve during training.

In this work, we aim to provide additional insights into this question by looking into the attention mechanism of the LLM. We used the zero-shot prompting technique to compare the LLM attention activation pattern when it conducts event segmentation task versus not on various stories. The attention activation patterns are extracted at each layer and attention head (Vig and Belinkov, 2019). To examine how the model directs attention towards different event structures, we segregated the attention activities based on different grammatical features (nouns, verbs, determiners, etc.).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Stories used

Story name	Number of events	Length (number of words)
Secret Life of Walter Mitty	6	1145
Story 1	8	817
Story 2	7	672
Story 3	8	673
Story 4	9	510
Story 5	9	663

2 Methods

2.1 Experimental procedures

An overview of our methods was shown in Figure 1 in supplementary material.

2.2 Text materials

We used six stories that involve continuous narration of physical and verbal interactions between the characters, as shown in Table 1. Our repository of stories consisted of one hand-written story and five GPT-generated stories. We used "Secret Life of Walter Mitty" for the hand-written story (Thurber, 1939). The natural structure of the story can be divided into separate, discrete events so that it was feasible for the model to perform the event segmentation task. We also used GPT-4 to generate five stories (Achiam et al., 2023). Each one of the stories included descriptions of sequential, real-life events that were structurally similar to the Walter Mitty story.

2.3 Prompts

To generate the story, we used the following prompt: "*Generate a story about a day in the life of college student Sarah. The story should be able to be divided into independent, discrete events. Shuffle the sequence of different events. Do not denote the event boundaries.*". When generating other stories, "*a day in the life of college student Sarah*" was replaced by other information and characters. Here, we specified that the story needs to be able to be divided into separate events instead of consisting of one continuous, inseparable event. In this way, it will be meaningful for the model to attempt event segmentation. Moreover, we asked GPT-4 to shuffle the events. By presenting the events out of their logical, linear sequence, the story ensured that the model did not rely solely on temporal cues to segment the events. Instead, it must rely on other indicators, such as changes in context, actions, or dialogue, to determine where one event ended and the other began.

To instruct the model to perform the event segmentation task, we leveraged the zero-shot prompting technique (Michelmann et al., 2023). The user role content prompt follows: "*An event is an ongoing coherent situation. The following story needs to be copied and segmented into events. Copy the following story word-for-word and start a new line whenever one event ends and another begins. This is the story:*". Then, the story will be inputted into the model. Additionally, we prompted to renew the instruction: "*This is a word-for-word copy of the same story that is segmented into events:.*" The system role content was set as "*You are a person listening to a continuous story, which can be divided into distinct events.*".

In contrast, in the non-segmentation condition, we revised the prompt: "*Copy the following story word-for-word. This is the story:*". Then the story would also be inputted, and the model was prompted to renew the instruction: "*This is a word-for-word copy of the same story:.*" The system role content was set as "*You are a person listening to a continuous story.*".

2.4 Model and grammatical feature extraction

The model we used to conduct event segmentation was `Llama-3-8B-Instruct` (Dubey et al., 2024). Stories were tokenized, and we used spaCy to extract grammatical features from the stories (Honnibal and Montani, 2017). Namely, we identified part-of-speech (POS) tags using spaCy. The POS tags T identified across all 6 stories were: T = {ADJ, ADP, ADV, AUX, DET, INTJ, NOUN, NUM, PART,

PRON, PROPN, PUNCT, SCONJ, VERB}. Corresponding linguistic feature for each POS tag is shown in Table 2.

We extracted the attention output of each layer and each attention head, aggregated across each word token w_i in each POS tag. For a single POS tag T_j and layer k, the average attention output across all attention heads was given by:

$$\overline{A}_{j,k} = \frac{\sum_{w_i \in T_j} \sum_{h=1}^{H_k} S_{i,j,k,h}}{H_k} \tag{1}$$

where: $\overline{A}_{j,k}$ was the average attention output across attention heads for tag T_j in layer k. $S_{i,j,k,h}$ represents the attention output for word token w_i in tag T_j at layer k and head h. H_k was the number of attention heads in layer k. $S_{i,j,k,h}$ was formally defined as:

$$S_{i,j,k,h} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

where: Q was the query matrix. K was the key matrix. V was the value matrix. d_k was the dimensionality of the key vectors. softmax$(\cdot)$ was the softmax function applied row-wise (Vaswani, 2017).

We also calculated the attention output across all layers for each tag, which was formally defined as:

$$\text{Category Attention} = \sum_{k=1}^{L} \frac{\overline{A}_{j,k}}{L} \tag{3}$$

where: L was the total number of layers.

3 Results

3.1 Attention output across layers and attention heads

The attention towards per token identified in each POS tag at each layer and attention head was calculated with the formula for $S_{i,j,k,h}$. To compare the attention output across layers and attention heads across event segmentation conditions, a pairwise t-test was performed on flattened vectors containing 1024 measurements. The results showed that there were significant differences in attention scores for all the POS tags identified from stories ($p < 0.001$), as shown in Table 3 in the appendix. Therefore, the overall pattern of attention outputs for all the POS tags were sensitive to the event segmentation task. Most of the POS tags gained more attention in the event segmentation task, except that INTJ, NUM, and PROPN gained less attention attribution in event segmentation compared to the non-segmentation condition. The average attention outputs across all layers and attention heads for each POS tag per token were shown in Figure 30 in the appendix.

3.2 Attention pattern across layers

To assess whether the general layer-wise attention pattern was influenced by the event segmentation task, for each POS tag, we computed the pairwise t-test on attention scores among layers when averaging across attention heads. The t-test results were shown in Table 2. At the layer level, ADP, ADV, AUX, DET, PUNCT, SCONJ, and VERB gained significantly higher attention scores in the event segmentation task ($p < 0.05$). However, INTJ and PROPN gained lower attention scores in the event segmentation task ($p < 0.05$). There was no significant difference in the layer pattern of attention outputs between the two segmentation conditions for POS tag ADJ, NOUN, NUM, PART, and PRON ($p > 0.05$).

Additionally, to investigate whether the attention output of a specific layer was consistently sensitive to certain POS tags across stories, we performed a non-parametric Wilcoxon signed-rank test on each layer's averaged attention score between 2 segmentation conditions. None of the 32 layers exhibited significant differences for most of the POS tags. However, the model significantly and consistently attributed more attention to NOUN in layer 19 in the segmentation task compared to the non-segmentation condition ($W = 0$, $p = 0.0312$, $n = 6$). Besides, PUNCT gained significantly and consistently less attention in the event segmentation task in layers 2 and 11, while this category also

Table 2: T-test results for each POS across layers

POS Tag	Corresponding linguistic feature	t-value	df	p-value
ADJ	Adjective	1.99	31	0.0556
ADP	Adposition	5.78	31	0.0000
ADV	Adverb	6.32	31	0.0000
AUX	Auxiliary verb	4.89	31	0.0000
DET	Determiner	3.23	31	0.0029
INTJ	Interjection	−2.49	31	0.0181
NOUN	Noun	1.18	31	0.2488
NUM	Number	−2.00	31	0.0549
PART	Particle	1.54	31	0.1341
PRON	Pronoun	1.87	31	0.0712
PROPN	Proper noun	−2.65	31	0.0125
PUNCT	Punctuation	2.27	31	0.0306
SCONJ	Subordinating conjunction	4.98	31	0.0000
VERB	Verb	5.59	31	0.0000

consistently gained more attention in the event segmentation task across stories in layers 4, 14-15, 18-19, 21, 23-25 and 29 ($W = 0$, $p = 0.0312$, $n = 6$).

4 Discussion

In the sections above, we demonstrated that attention directed toward words like interjections, numbers, and proper nouns experienced reduction when LLM performs event segmentation. On the contrary, the attention directed toward other words like adverbs, verbs, and adpositions experienced a significant increase. Words like verbs, adverbs, and adpositions are essential in establishing relationships between words and constructing meanings in a sentence. For example, adpositions are crucial in indicating contextual information like the spatial and temporal relationships in narratives (Huddleston and Pullum, 2005). These categories of words are highly informative for discovering changes in elements like time, space, objects, and goals in a continuous narrative due to their semantic and grammatical roles in sentences (Ursini, 2011; Payne et al., 2010). Previous studies have shown that changes in these elements like time and space, are significantly correlated with the presence of event boundary detected by human (Michelmann et al., 2021). Therefore, it is possible that LLM redistributed more attention on words that indicate changes in the previously mentioned elements in the input text, contributing to the emerging feature of event segmentation. Although the initial training goal of LLM is not to perform event segmentation, the way it segments continuous events converges with the approach that human takes when doing the same task. In conclusion, our results provide an enhanced understanding of an naturally emerging cognitive feature, which is crucial in LLMs as well as in human high-level cognition. Moreover, previous research has demonstrated that LLMs can be considered a model organism for investigating language processing in the human brain (Tuckute et al., 2024; Goldstein et al., 2022; Kumar et al., 2022). Therefore, our results can potentially provide additional insights into underlying mechanism for continuous event processing in human, due to the naturally convergent property of the human brain and LLMs.

5 Limitations

Our task does not have relatively sufficient variations among story structures in terms of the total number of events and the total number of words. It would be ideal to examine our results further using stories that have drastically different total numbers of words and total number of events. Moreover, due to time and computing resources constraints, we only ran our task on one variation of the Llama 3 models. Testing with other language models across architectures and model sizes will be a promising research direction.

Acknowledgments and Disclosure of Funding

We thank Prof. Sebastian Michelmann at New York University and Prof. Hsin-Hung Li at The Ohio State University for their valuable inputs. We ran the model on New York University High Performance Computing Cluster (NYU HPC). We have not received any funding for this study. We declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ashley S Bangert, Christopher A Kurby, Allyson S Hughes, and Omar Carrasco. Crossing event boundaries changes prospective perceptions of temporal length and proximity. *Attention, Perception, & Psychophysics*, 82:1459–1472, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sarah DuBrow and Lila Davachi. Temporal binding within and across events. *Neurobiology of learning and memory*, 134:107–114, 2016.

Ariel Goldstein, Eric Ham, Samuel A Nastase, Zaid Zada, Avigail Grinstein-Dabus, Bobbi Aubrey, Mariano Schain, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*, pages 2022–07, 2022.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

Rodney Huddleston and Geoffrey Pullum. Introduction to english grammar. *Zeitschrift für Anglistik und Amerikanistik*, 53(2):195–197, 2005.

Anna Jafarpour, Elizabeth A Buffalo, Robert T Knight, and Anne GE Collins. Event segmentation reveals working memory forgetting rate. *Iscience*, 25(3), 2022.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, pages 2022–06, 2022.

Sebastian Michelmann, Amy R Price, Bobbi Aubrey, Camilla K Strauss, Werner K Doyle, Daniel Friedman, Patricia C Dugan, Orrin Devinsky, Sasha Devore, Adeen Flinker, et al. Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature communications*, 12(1): 5394, 2021.

Sebastian Michelmann, Manoj Kumar, Kenneth A Norman, and Mariya Toneva. Large language models can segment narrative events similarly to humans. *arXiv preprint arXiv:2301.10297*, 2023.

John Payne, Rodney Huddleston, and Geoffrey K. Pullum. The distribution and category status of adjectives and adverbs. *Word Structure*, 3(1):31–81, 2010.

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

James Thurber. "the secret life of walter mitty"(1939). 1939.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 2024.

Francesco-Alessio Ursini. *The Language Of Space: The Acquisition And Interpretation of Spatial Adpositions In English*. PhD thesis, Macquarie University Printing services, 2011.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

MM Waldrop. Can chatgpt help researchers understand how the human brain handles language? *Proceedings of the National Academy of Sciences of the United States of America*, 121(25):e2410196121–e2410196121, 2024.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

A Appendix / supplemental material

A.1 The experimental process

The experimental process was shown in Figure 1.

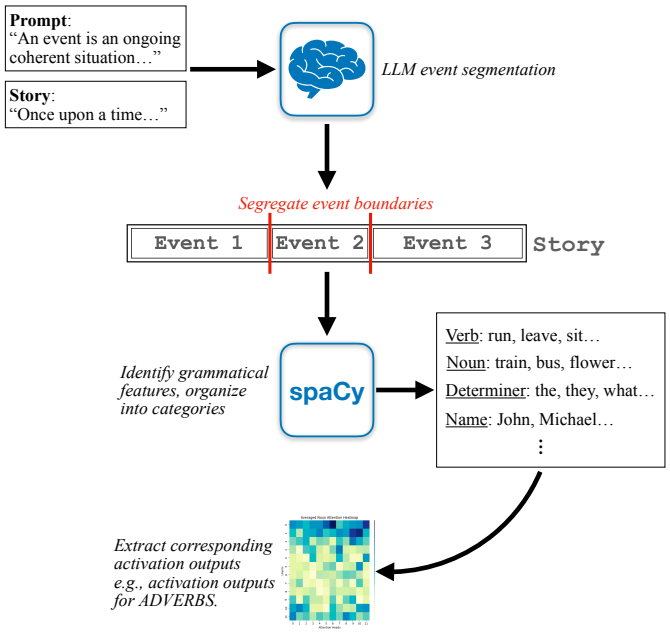

Figure 1: The experimental process

A.2 Code and datasets availability

All of the code and datasets required to reproduce our study are uploaded and can be accessed using the following link: `https://anonymous.4open.science/r/attention_event-4A53/README.md`

A.3 Attention distribution for each POS tag

The attention scores of each noun token (NOUN) across all the layers and attention heads were shown in Figure 2, broken out by layer (vertical axis) and head (horizontal axis). The attention scores of nouns at each layer were also shown in Figure 3. The attention was attributed more to nouns in middle layers, such as 9, 10, 14, and 15, and several attention heads are sensitive to nouns, such as layer 10 head 3, layer 14 head 4, layer 15 head 14, layer 16 head 27, and layer 17 head 9.

The same analysis and plots were performed and shown for all the POS tags. For adjective (ADJ) tokens, Figure 4 and Figure 5 indicated that the middle layers generally attributed more attention to them, such as layers 10, 12, 14, and 15. The layer 18 attention head 14 was specifically sensitive to adjective tokens.

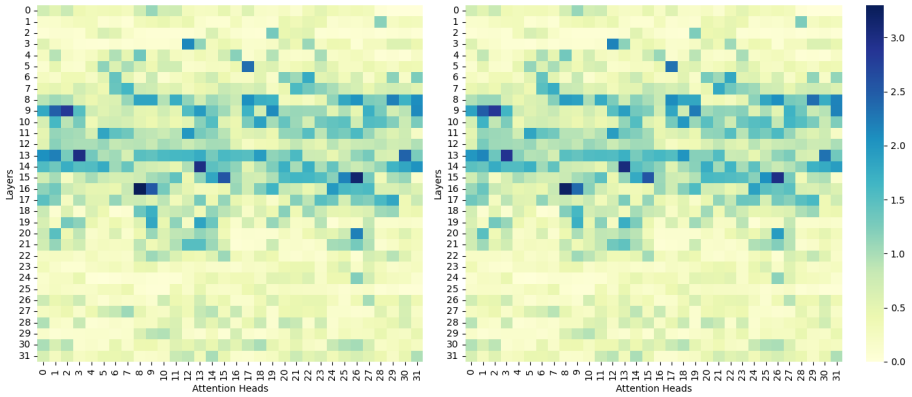

Figure 2: Heatmap of NOUN attention score in segmentation (left) and non-segmentation (right) task

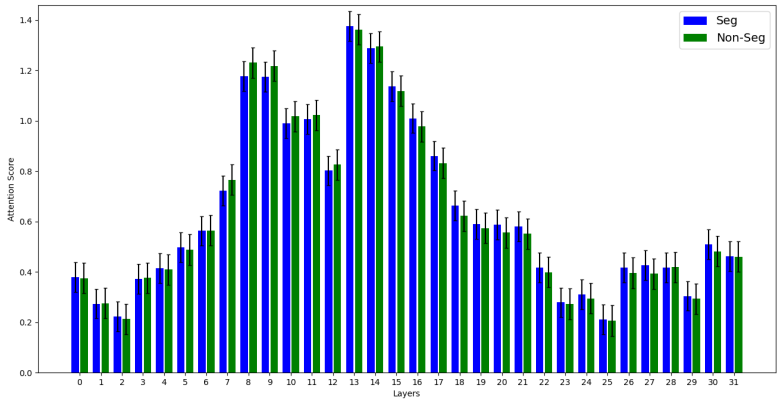

Figure 3: Layer average of NOUN attention score in two segmentation conditions

For adposition (ADP) tokens, Figure 6 and Figure 7 indicated that the middle layers generally attributed more attention to them, such as layers 12 and 14. Layer 13, head 2, and Layer 21, head 21 were sensitive to the adposition.

For adverb (ADV) tokens, Figure 8 and Figure 9 indicated that the middle layers generally attributed more attention to them, such as layers 14-15. Layer 13 head 2, layer 14 head 3, and Layer 18 head 14 were specifically sensitive to adverb tokens.

For auxiliary verb (AUX) tokens, Figure 10 and Figure 11 indicated that late middle layer 31 attention head 18 was specifically sensitive to auxiliary verb tokens.

For determiner (DET) tokens, Figure 12 and Figure 13 indicated that the early layer 1 and middle layers 12 and 14 generally attributed more attention to them. Layer 1, head 9, 11, and 16, and Layer 25, head 24 were specifically sensitive to determiner tokens.

For interjection (INTJ) tokens, Figure 14 and Figure 15 indicated that the early layer 1 head 10, 12, and 16, late layers 23 head 3 and layer 26 head 6 were specifically sensitive to interjection tokens.

For number (NUM) tokens, Figure 16 and Figure 17 indicated that the early layer 1 and middle layers 10, 14, 15, and 16 generally attributed more attention to them. The layer 1 head 11, layer 10 head 28, layer 14 head 2, 13, and 31, layer 21 head 26, layer 24 head 15, layer 29 head 29, and layer 30 head 17 were specifically sensitive to number tokens.

For particle (PART) tokens, Figure 18 and Figure 19 indicated that layer 1 generally attributed more attention to them, and late middle layer 31 attention head 18 were specifically sensitive to particle tokens, similar to Auxiliary verb (AUX).

For pronoun (PRON) tokens, Figure 20 and Figure 21 indicated that the early layer 1 and middle layers 14 and 15 generally attributed more attention to them. Layer 1, heads 9 and 16; layer 14, head 3; layer 17, head 2; layer 21, heads 16 and 24; layer 25, head 24; and Layer 31, head 18, were specifically sensitive to pronoun tokens.

For proper noun (PROPN) tokens, Figure 22 and Figure 23 indicated that the middle layers generally attributed more attention to them, such as layers 9, 10, 12, 14, and 15. The layer 17 attention head 14 was specifically sensitive to proper noun tokens.

For punctuation (PUNCT) tokens, Figure 24 and Figure 25 indicated that layer 1 generally attributed more attention to them. The layer 1 attention heads 29, 30, and 31 were specifically sensitive to punctuation tokens, similar to particle (PART) and auxiliary verbs (AUX).

For subordinating conjunction (SCONJ) tokens, Figure 26 and Figure 27 indicated that the middle layers generally attributed more attention to them, such as layers 9-12 and 14-16. The layer 13 attention heads 1 and 2, layer 21 head 21, and layer 29 head 26 were specifically sensitive to subordinating conjunction tokens.

For verb (VERB) tokens, Figure 28 and Figure 29 indicated that the middle layers generally attributed more attention to them, such as layers 9-12 and 14-16. The layer 1 head 11, layer 11 head 15, layer 14 head 14, layer 20 head 14, and layer 21 head 2 were specifically sensitive to verb tokens.

A.4 Attention score across layers and attention heads

Here we exhibited the pairwise t-test results in Table 3 for each POS tag when comparing the attention score pattern across layer and attention heads in two segmentation conditions.

Table 3: T-test for each POS across layers and attention heads

POS Tag	Corresponding linguistic feature	t-value	df	p-value
ADJ	Adjective	5.22	1023	0.0000
ADP	Adposition	10.16	1023	0.0000
ADV	Adverb	16.01	1023	0.0000
AUX	Auxiliary verb	7.64	1023	0.0000
DET	Determiner	6.81	1023	0.0000
INTJ	Interjection	-4.65	1023	0.0000
NOUN	Noun	3.56	1023	0.0000
NUM	Number	-4.68	1023	0.0000
PART	Particle	3.32	1023	0.0000
PRON	Pronoun	3.94	1023	0.0000
PROPN	Proper noun	-8.17	1023	0.0000
PUNCT	Punctuation	5.75	1023	0.0000
SCONJ	Subordinating conjunction	8.65	1023	0.0000
VERB	Verb	11.99	1023	0.0000

A.5 Attention score across all POS tags

The averaged attention score for each POS tag, across stories, layers, and attention heads, in two segmentation conditions is shown in Figure 30. The pairwise t-test was conducted to compare the overall attention scores across all POS tags between two segmentation conditions and investigate whether event segmentation increased or decreased the overall attention attributed to per token. The results indicate that the difference between the two conditions was not statistically significant ($t(13) = 1.96, p = 0.0721$).

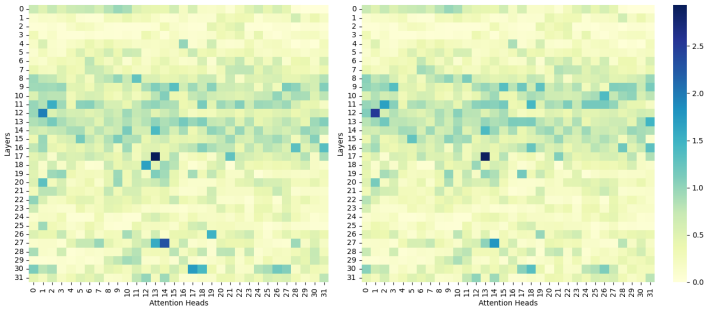

Figure 4: Heatmap of ADJ attention score in segmentation (left) and non-segmentation (right) task

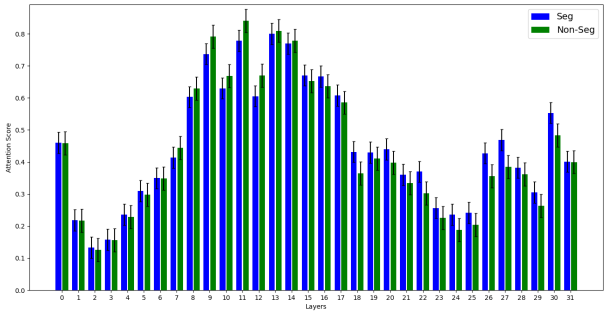

Figure 5: Layer average of ADJ attention score in two segmentation conditions

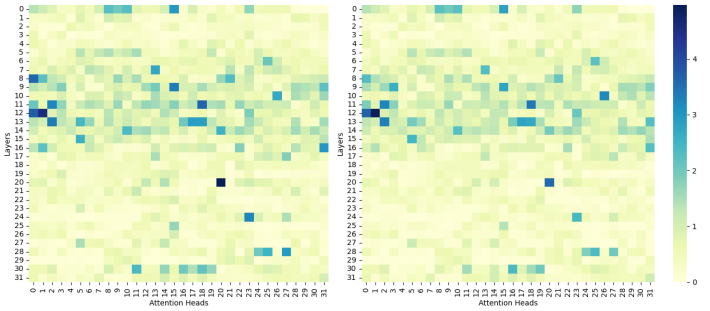

Figure 6: Heatmap of ADP attention score in segmentation (left) and non-segmentation (right) task

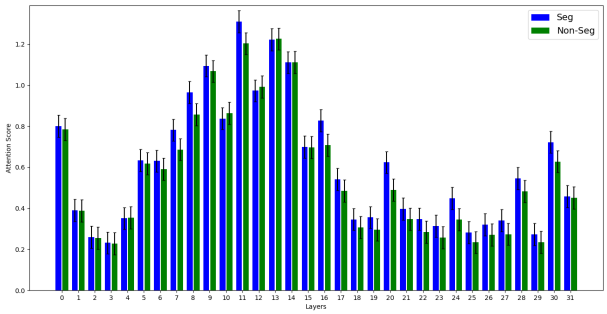

Figure 7: Layer average of ADP attention score in two segmentation conditions

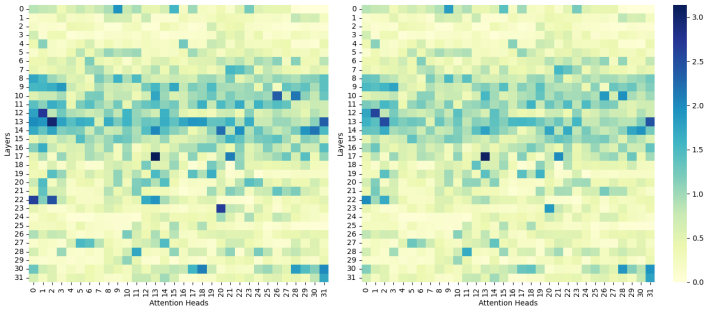

Figure 8: Heatmap of ADV attention score in segmentation (left) and non-segmentation (right) task

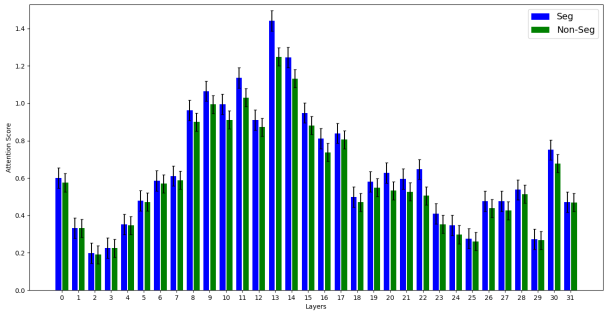

Figure 9: Layer average of ADV attention score in two segmentation conditions

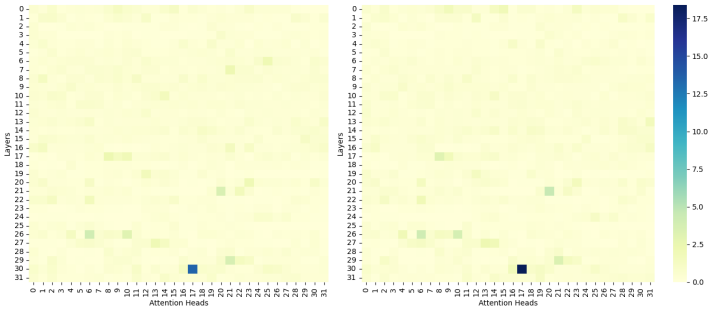

Figure 10: Heatmap of AUX attention score in segmentation (left) and non-segmentation (right) task

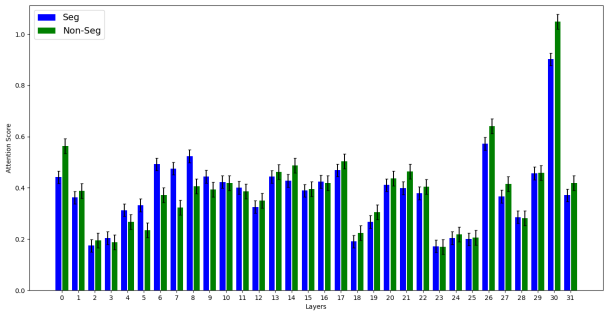

Figure 11: Layer average of AUX attention score in two segmentation conditions

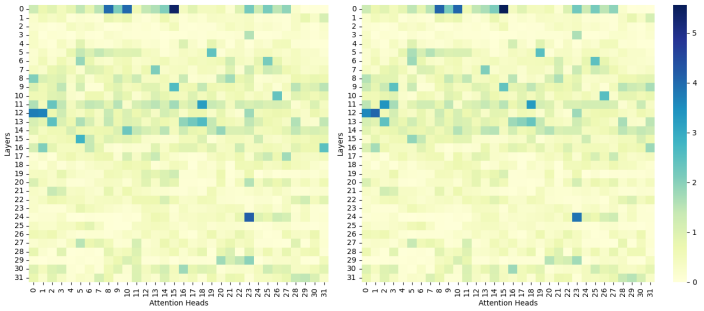

Figure 12: Heatmap of DET attention score in segmentation (left) and non-segmentation (right) task

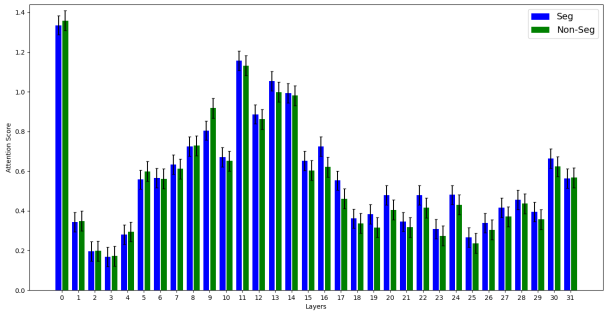

Figure 13: Layer average of DET attention score in two segmentation conditions

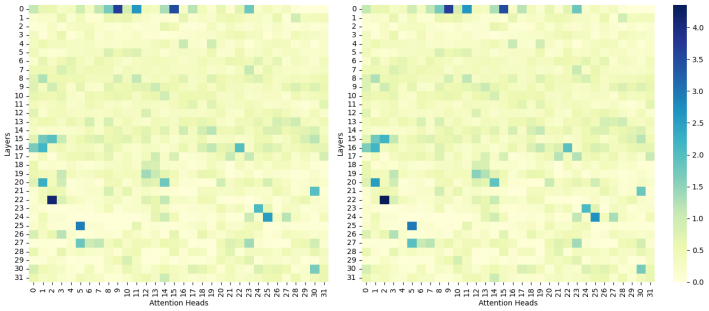

Figure 14: Heatmap of INTJ attention score in segmentation (left) and non-segmentation (right) task

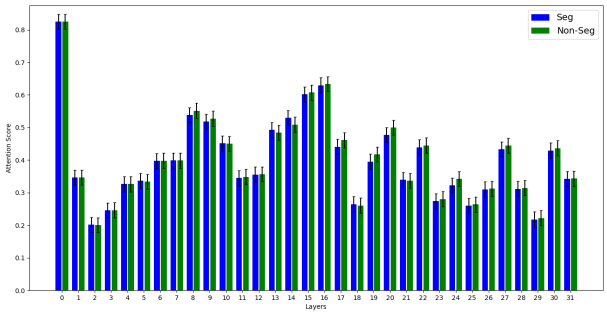

Figure 15: Layer average of INTJ attention score in two segmentation conditions

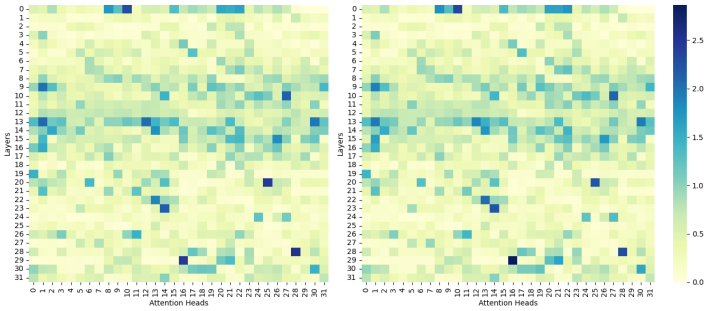

Figure 16: Heatmap of NUM attention score in segmentation (left) and non-segmentation (right) task

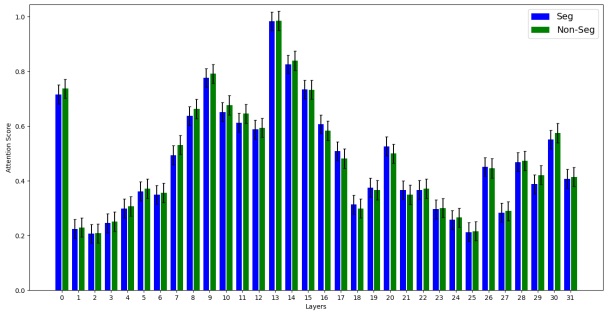

Figure 17: Layer average of NUM attention score in two segmentation conditions

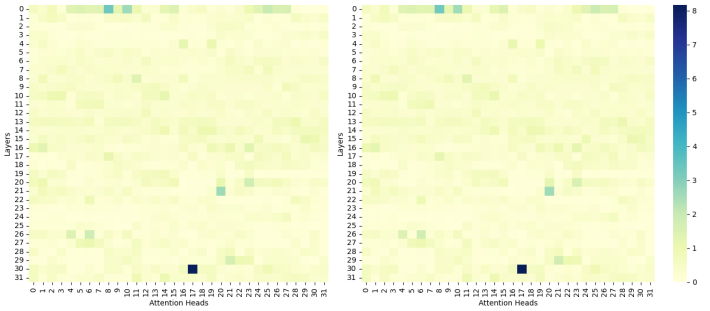

Figure 18: Heatmap of PART attention score in segmentation (left) and non-segmentation (right) task

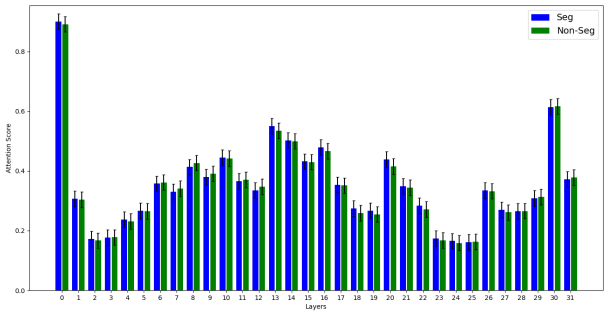

Figure 19: Layer average of PART attention score in two segmentation conditions

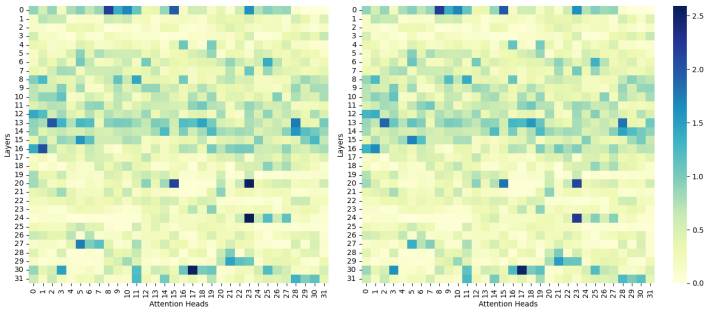

Figure 20: Heatmap of PRON attention score in segmentation (left) and non-segmentation (right) task

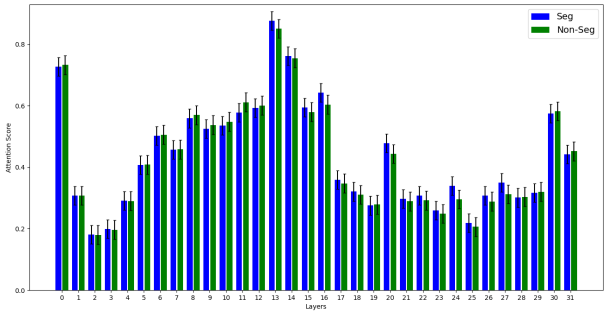

Figure 21: Layer average of PRON attention score in two segmentation conditions

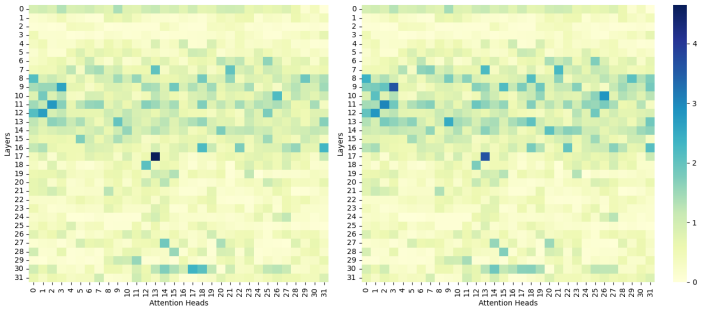

Figure 22: Heatmap of PROPN attention score in segmentation (left) and non-segmentation (right) task

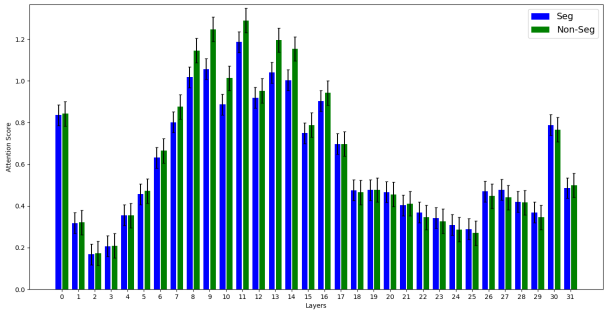

Figure 23: Layer average of PROPN attention score in two segmentation conditions

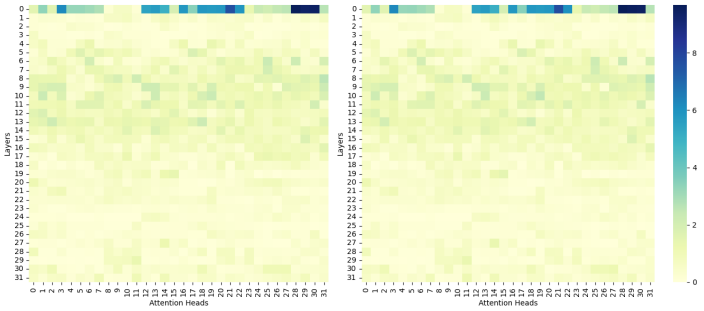

Figure 24: Heatmap of PUNCT attention score in segmentation (left) and non-segmentation (right) task

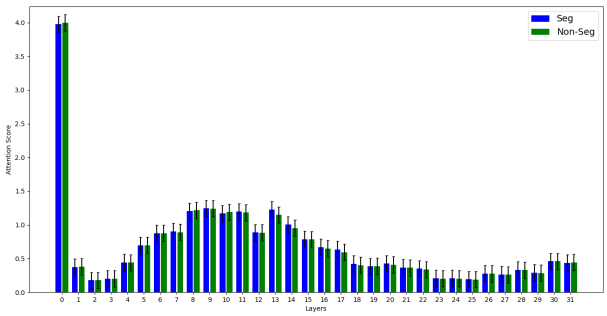

Figure 25: Layer average of PUNCT attention score in two segmentation conditions

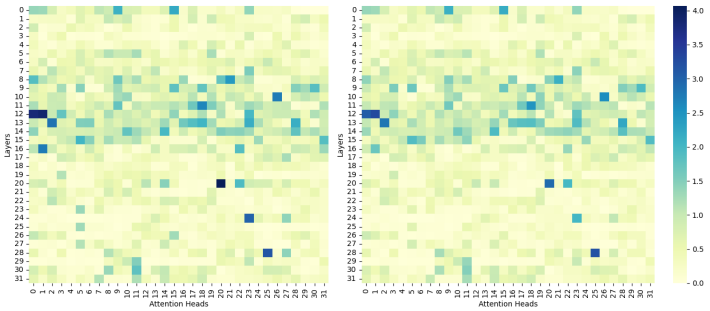

Figure 26: Heatmap of SCONJ attention score in segmentation (left) and non-segmentation (right) task

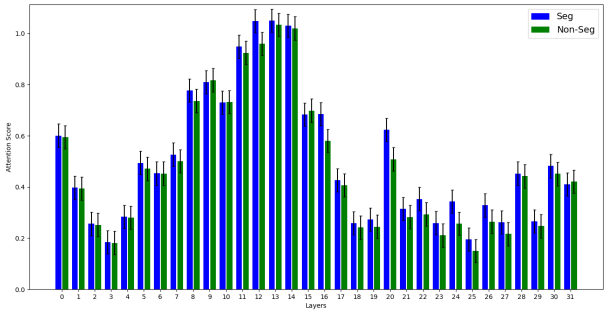

Figure 27: Layer average of SCONJ attention score in two segmentation conditions

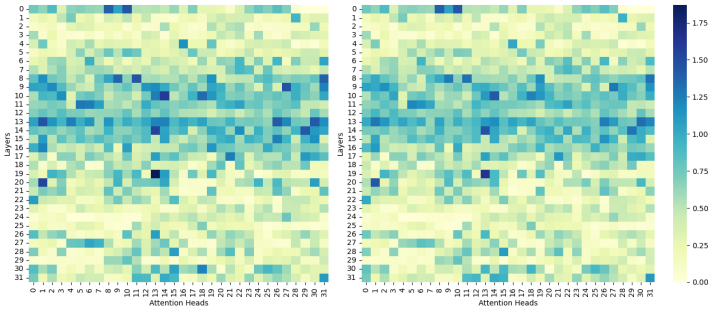

Figure 28: Heatmap of VERB attention score in segmentation (left) and non-segmentation (right) task

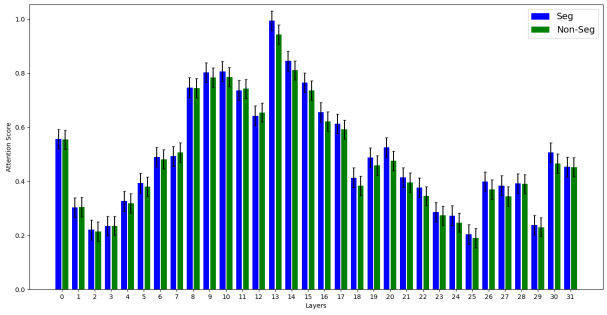

Figure 29: Layer average of VERB attention score in two segmentation conditions

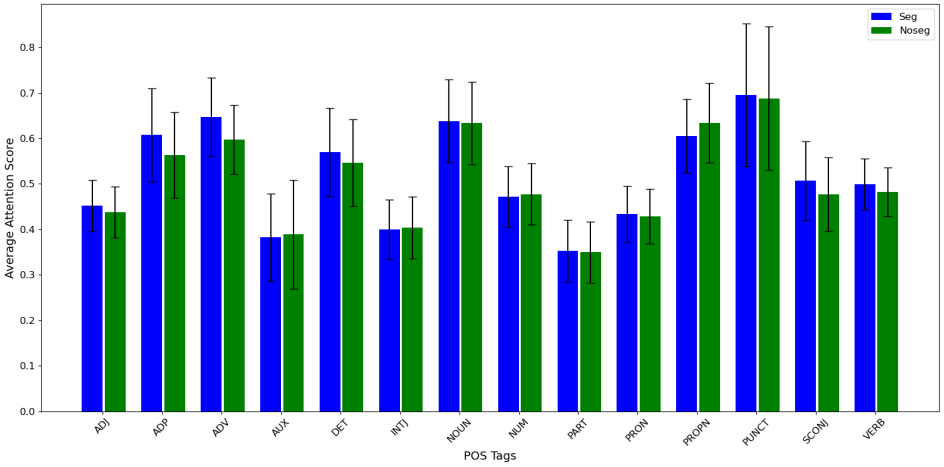

Figure 30: POS tag attention.