HIERARCHICAL EPISODIC MEMORY IN LLMS VIA MULTI-SCALE EVENT ORGANIZATION

Martin A Benfeghoul[1], **Haitham Bou-Ammar**[1,2], **Jun Wang**[2] **& Zafeirios Fountas**[1]
[1]Noah's Ark Lab, Huawei, London, UK
martin.antoine.benfeghoul@h-partners.com
{haitham.ammar,zafeirios.fountas}@huawei.com
[2]University College London, London, UK
jun.wang@ucl.ac.uk

ABSTRACT

A major limitation of contemporary large language models (LLMs) is their significant performance degradation when processing long contexts, primarily due to self-attention dilution and context window limitations. Recent work on retrieval-augmented LLMs has shown that integrating formation and retrieval of human-inspired episodic memory (a form of associative memory) into Transformers, via an architecture termed EM-LLM, enables pre-trained models to process up to 10M tokens while consistently outperforming their full-context versions using only a fraction of the computational resources. A crucial feature of EM-LLM is the segmentation of the model's KV-cache into human-like events based on token-level surprise. However, this approach overlooks the hierarchical nature of human episodic memory, which exhibits nested timescale organization across multiple levels of abstraction. Here, we introduce two novel head-level event segmentation methods that leverage the inherent hierarchical processing in Transformer layers, combining similarity-based boundary detection with coordinated event hierarchies. Our experiments suggest that these structures are not only likely to improve retrieval performance but also show patterns consistent with the nested event hierarchies observed in human cognition, providing both practical advances in LLM capabilities and insights into memory organization across artificial and biological systems.

1 INTRODUCTION

Modern pre-trained large language models (LLMs) based on the Transformer architecture (Vaswani et al., 2017) rely on their context window to provide domain-specific, contextual information at inference time. However, such models struggle when the context necessary for accurate inference gets too long (Liu et al., 2024). Text lengths beyond the maximum context window during training (Kazemnejad et al., 2024), as well as attention dilution (Tworkowski et al., 2023; Ye et al., 2024) are common culprits for such decreases in performance.

Recent works have tackled these issues by developing two main retrieval-based approaches. The first approach, retrieval-augmented generation (RAG: Lewis et al. 2020; Gao et al. 2024), enhances the context by retrieving relevant information from the larger body of text. The second approach involves retrieving previously computed key-value (KV) pairs within the corresponding attention head (Wu et al., 2022; Tworkowski et al., 2023; Bertsch et al., 2023). Such methods have since been extended by grouping KV pairs into separate segments at token-level and then retrieved as continuous blocks of tokens in the attention heads (Xiao et al., 2024; Fountas et al., 2025).

One such approach, EM-LLM (Fountas et al., 2025), achieves state-of-the-art performance by dynamically segmenting KV blocks using surprise (measured as token-level prior surprisal), inspired by human episodic memory [1] and event cognition. However, because this approach relies on next-token prediction, it is limited to token-level event segmentation. While this successfully mirrors

[1]As a key associative memory system, episodic memory binds contextual details over time into events.

how the human brain segments continuous experience into discrete episodic events (Clewett et al., 2019; Zacks, 2020) at points of high surprise (Zacks et al., 2007; 2011; Roseboom et al., 2019; Sinclair et al., 2021; Fountas et al., 2022; 2025), it lacks the hierarchical, nested-timescale structure observed in human event segmentation (Baldassano et al., 2017). Such structure is crucial for generative models' performance in other domains (Zakharov et al., 2022b;a; Saxena et al., 2021).

The layers and attention heads of Transformers naturally organize information hierarchically: they are known to focus on distinct linguistic features and semantics (Clark et al., 2019; Voita et al., 2019), with attention sparsity increasing in higher layers (Zhang et al., 2023). This hierarchical organization is reflected in memory access patterns as well— Fountas et al. (2025) found that each layer in EM-LLM retrieved different events from the others. These insights suggest that Transformers learn to process information in a focused, hierarchical manner, motivating the need for memory structures that preserve this organization. We therefore propose two novel methods for head-level event segmentation to this end.

An event is retrieved based on the similarity of its representative key with the current query. Thus, higher similarity between event keys is likely to mean higher similarity with the current query once retrieved, and hence higher attention scores. This intuition is supported by Fountas et al. (2025), who found that both humans and their surprise-based segmentation method produce events with significantly higher within-event key similarity compared to fixed and random segmentation. Building on these insights, our proposed segmentation methods directly measure the similarity between a new key and the current event's keys, ending the event when this similarity falls below a threshold.

Our experiments demonstrate that our head-level event segmentation methods closely align with both surprise-based segmentation and human-perceived events. Moreover, the hierarchical structure we observe across layers and heads mirrors theories of human event cognition. Overall, our results motivate our methods promising for improving KV-Retrieval performance, and provide a computational framework for studying parallels between human memory processes and LLMs.

2 METHODS

2.1 SURPRISE-BASED SEGMENTATION

EM-LLM's definition of surprise acts as a measure of prediction error at the token level. Fountas et al. (2025) demonstrated EM-LLM's capacity to segment events similarly to humans using surprise, improving both event modularity (within-event key similarity) and task performance when compared to its fixed segmentation counterpart. We will therefore look to compare our own methods to this baseline surprise-based segmentation method. In Bayesian terms, surprise is quantified by the surprisal (negative log-likelihood) of observing the current, ground-truth token given the previous tokens in an autoregressive model. A token x_t is considered to fall outside of the current event if its surprise value exceeds a threshold T:

$$-\log P(x_t|x_1,\ldots,x_{t-1};\theta) > T_{Sur} \quad \text{with} \quad T_{Sur} = \mu_{t-\tau:t} + \gamma\sigma_{t-\tau:t} \tag{1}$$

where $\mu_{t-\tau:t}$, $\sigma^2_{t-\tau:t}$ are the mean and variance of surprise over a window τ, and γ a scaling factor.

2.2 GAUSSIAN EVENTS

For key-based, head-level event segmentation, we model events as coherent clusters in key space. Specifically, we introduce a head-specific Gaussian representation of single events based on the event keys of the attention head. We therefore obtain a multivariate Gaussian distribution parametrised by the mean ($\overline{\mathbf{k}}^h$) and covariance (Σ^h) of the corresponding keys for each attention head h. We can then evaluate the probability of a new key corresponding to such an event.

In order to enable online detection of events, we implement efficient updates to the event's Gaussian representation in each head h using the multivariate Gaussian update rule. For each new key $\mathbf{k}^h_t$, we evaluate its probability of belonging to the current event. If this is too low, we end the current event and start a new one with this key; otherwise, we update the current event's parameters using:

$$\overline{\mathbf{k}}^h_t = \frac{(n^h_t - 1)\overline{\mathbf{k}}^h_{t-1} + \mathbf{k}^h_t}{n^h_t}\ , \quad \Sigma^h_t = \frac{(n^h_t - 1)\Sigma^h_{t-1} + (\mathbf{k}^h_t - \overline{\mathbf{k}}^h_{t-1})(\mathbf{k}^h_t - \overline{\mathbf{k}}^h_t)}{n^h_t} \tag{2}$$

where n_t^h is the number of keys contained within this event. In practice, we use a diagonal Gaussian to avoid having a singular covariance matrix when n_t^h is smaller than the head dimension d_k.

2.3 K-SIMILARITY

Updating a multivariate Gaussian and calculating a probability for every head in the model, for every new key in a sequence, is computationally very expensive. Alternatively, our second proposed method leverages the highly-optimised matrix multiplication capabilities of modern deep learning libraries to compute a similar segmentation method based on key similarity. We will refer to this method as K-Similarity.

This method takes the dot product of a new key with the mean key of the current event, as a measure for the similarity of such a key with that event. We then apply a similar threshold on this signal as in Eq. 1 but only over the event keys, to take into account the mean and variance within the current event. Combining notation from equations 1 and 2, we define K-Similarity as:

$$(\bar{\mathbf{k}}_{t-1}^h)^T \mathbf{k}_t^h \; > \; T_{Sim} \qquad \text{with} \qquad T_{Sim} = \mu_{t_0:t} + \gamma \sigma_{t_0:t} \tag{3}$$

$$\mu_{t_0:t} = \frac{1}{t - t_0} \sum_{i=t_0+1}^{t} (\bar{\mathbf{k}}_{i-1}^h)^T \mathbf{k}_i^h \tag{4}$$

where t_0 is the start of the current event, $\mu_{t_0:t}$ and $\sigma^2_{t_0:t}$ are the mean and variance of the dot product for each key in that event (Eq. 4), and γ is a scaling factor. We overload notation for μ and σ but they are separate from those described in Eq. 1.

2.4 EXPERIMENTS

We begin by assessing whether our proposed segmentation methods can effectively replace surprise-based segmentation at the sequence level. Success at this level would suggest comparable performance when applied to individual heads. To evaluate sequence-level performance, we aggregate signals across all attention heads by averaging the Gaussian probabilities and key similarity dot products, then apply thresholding to these averaged values. Subsequently, we analyse the emergence of hierarchical patterns when such methods are applied independently to individual attention heads, enabling head-specific event segmentation.

For consistency, we will look to initially replicate the analysis in Fountas et al. (2025) for our comparison with EM-LLMs' surprise, as well as human-perceived events. We therefore evaluate on the human-annotated podcasts (two podcasts, $5 - 7k$ tokens long) based on human data obtained from previous studies (Kumar et al., 2023; Michelmann et al., 2021; Lositsky et al., 2016), as well as the long-context benchmarks PG-19 (Rae et al., 2020) and Long-Bench (Bai et al., 2023) in order to confirm our results on larger datasets.

We naturally also adopt their metrics for a direct comparison. Namely, we use Wasserstein distance (Panaretos & Zemel, 2019) to compare the relative positions of events between human annotations, surprise-based segmentation, and our own proposed methods. We use modularity (Newman & Girvan, 2004) to measure the impact of our methods on the similarity of keys within events. The threshold parameter for our own methods is set such that the number of sequence-level events matches that of surprise-based segmentations for comparison with humans, as well as random and fixed methods, as described in Fountas et al. (2025).

3 RESULTS

3.1 SEQUENCE-LEVEL SEGMENTATION

The correlation between event boundary positions of Gaussian and K-Similarity segmentations at sequence-level with that of humans is shown in Table 1. Although surprise is still closest to humans, both Gaussian and K-Similarity methods achieve relatively similar scores, especially when compared to random and fixed segmentations. As surprise is closest to humans, we use this method as the baseline for comparison on longer benchmarks, where human annotations of event positions are unavailable (Table 2). Note that we limit this part of the analysis to K-Similarity only due to the

Table 1: Measuring Wasserstein distance with humans (lower = better) over the human-annotated podcasts dataset using LLaMA-3-8B-Instruct.

Podcast	Surprise	K-Similarity	Gaussian	Random	Fixed
Monkey	0.02	0.02	0.05	0.11	0.22
Tunnel	0.01	0.04	0.04	0.10	0.22

Table 2: Measuring Wasserstein distance with surprise (lower = better) over longer texts and benchmarks using LLaMA-3-8B-Instruct.

Dataset	K-Similarity	Random	Fixed
PG-19 (test+val)	0.03	0.10	0.24
Long-Bench	0.04	0.13	0.31

significantly higher computational cost of the Gaussian method. The results in these longer benchmarks are consistent with those observed in the podcasts dataset, with K-Similarity achieving significantly smaller Wasserstein distance with surprise than both random and fixed segmentation methods. Our analysis reveals that the Gaussian method achieves the highest modularity score among

Table 3: Measuring Modularity of segmented events (higher = better) over the human-annotated podcasts dataset (ratio with fixed segmentation) using LLaMA-3-8B-Instruct.

Podcast	Surprise	K-Similarity	Gaussian	Human	Random	(Fixed)
Monkey	1.32	1.49	2.04	1.32	1.05	1.00
Tunnel	1.20	1.19	1.30	1.45	1.01	1.00

all approaches (Tab. 3), followed by humans and K-Similarity, while surprise-based segmentation shows slightly lower modularity. This superior modularity in our key-based methods is expected, as they explicitly segment events based on key similarity. Notably, both methods accomplish this segmentation online, with K-Similarity offering particularly efficient computation.

Overall, both proposed methods replicate the key characteristics of EM-LLM's surprise-based segmentation, indicating their potential to enhance downstream task performance. The Gaussian-based method demonstrates superior modularity (up to $2.04\times$ fixed segmentation), while K-Similarity achieves stronger correlation with both human segmentation patterns and surprise-based approaches while reducing computational complexity by eliminating expensive covariance calculations.

3.2 Head-level Segmentation

When segmenting events at head-level, we observe a hierarchical structure (indicated primarily by the variation in average event counts per layer, see Fig. 1 Left) where later layers form more numerous, shorter events than earlier layers, creating nested timescales across the model. Also, most layers show a non-negligible standard deviation in this number across attention heads. This further motivates the potential of head-level segmentations in segmenting more meaningful events to be later retrieved for each head. Interestingly, we note that these events appear to move closer to humans and surprise as we move up the model layers, with an especially big improvement up to layer 5. These results are further confirmed on the longer summarisation tasks of the Long-Bench dataset (Fig. 2). However, we observe greater variation in later layers for both plots. This is further accentuated in certain tasks (see Appendix A.1), with GovReport and QMSum closely matching Figure 1, while MultiNews and SAMSum show more variation in later layers.

4 Discussion

Our results show that both new approaches for event segmentation in attention heads are promising alternatives to token-level surprise-based segmentation methods for LLMs using KV-Retrieval. The

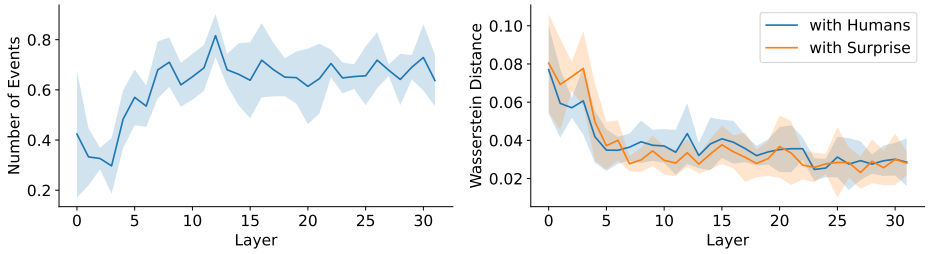

Figure 1: Visualisation of head-level K-Similarity segmentation properties averaged over the podcasts dataset, using LLaMA-3-8B-Instruct. Left: Average number of events found in the sequence (normalised) per head for each layer of the model. Right: Average Wasserstein distance between K-Similarity vs. humans and surprise per head, for each layer of the model. Standard deviation is measured across the heads of each layer in both figures.

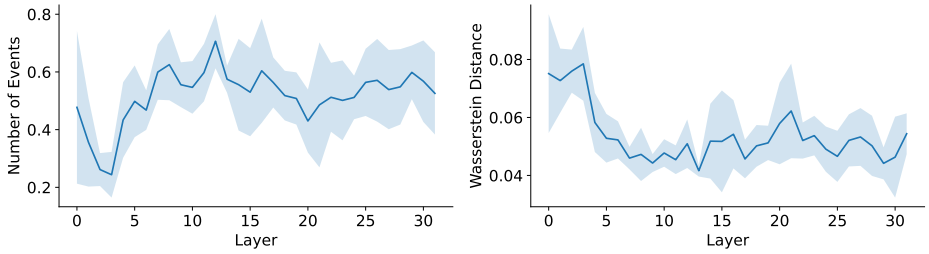

Figure 2: Same metrics and visualisations as Fig. 1 for Long-Bench's summarisation tasks (GovReport, QMSum, MultiNews, SAMSum). Wasserstein distance is measured against surprise only here as human event annotations are unavailable. See Appendix A.1 for individual task plots.

large differences across layers and heads in the number of events suggest that head-level segmentation captures functionally meaningful variations in temporal structure. This aligns with Fountas et al. (2025), who found that EM-LLM improved performance over RAG, even when both methods retrieved from the same surprise-based events. They attributed a large part of this improvement to EM-LLM's layer-wise retrieval approach, compared to RAG's single retrieval step at the input level. Our findings suggest that further tailoring retrieved events to specific attention heads could enhance this advantage by leveraging the distinct temporal patterns each head learns to recognize.

The human brain is believed to segment continuous experience into discrete episodic events (Clewett et al., 2019; Zacks, 2020) at points of high surprise (Zacks et al., 2007; 2011; Roseboom et al., 2019; Sinclair et al., 2021; Fountas et al., 2022; 2025), organizing them into a hierarchical, nested-timescale structure (Baldassano et al., 2017). While layer-wise retrieval allows EM-LLM to maintain some form of hierarchy across layers, our head-level segmentation approach offers parallels to this biological organization. The substantial variation in event counts between layers, combined with high variance between attention heads within layers, reveals naturally emerging nested timescales – though their organization may differ from biological systems due to architectural features like residual connections. Furthermore, the increased number of events in later layers, resulting in shorter events, aligns with Zhang et al. (2023)'s observations of increased sparsity in attention scores in these layers. Although the precise mapping between these model-learned hierarchies and human cognitive structures remains unclear and warrants investigation, we find that event positions identified in later layers show the strongest correlation with human-perceived sequence-level events.

In conclusion, our work demonstrates that head-level event segmentation in LLMs offers a promising direction for improving KV-Retrieval performance. The observed hierarchical structure across layers and heads aligns with theories of human event cognition, particularly regarding nested timescales. While the precise relationship between model-detected and human-perceived events requires further investigation, our findings suggest that later layers capture sequence-level events that correspond best to human intuitions.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-range transformers with unlimited length input. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lJWUJWLCJo.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

David Clewett, Sarah DuBrow, and Lila Davachi. Transcending time in the brain: How event memories are constructed from experience. *Hippocampus*, 29(3):162–183, 2019.

Zafeirios Fountas, Anastasia Sylaidi, Kyriacos Nikiforou, Anil K. Seth, Murray Shanahan, and Warrick Roseboom. A Predictive Processing Model of Episodic Memory and Time Perception. *Neural Computation*, 34(7):1501–1544, 06 2022. ISSN 0899-7667. doi: 10.1162/neco_a_01514. URL https://doi.org/10.1162/neco_a_01514.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BI2int5SAC.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Manoj Kumar, Ariel Goldstein, Sebastian Michelmann, Jeffrey M Zacks, Uri Hasson, and Kenneth A Norman. Bayesian surprise predicts human event segmentation in story listening. *Cognitive science*, 47(10):e13343, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Olga Lositsky, Janice Chen, Daniel Toker, Christopher J Honey, Michael Shvartsman, Jordan L Poppenk, Uri Hasson, and Kenneth A Norman. Neural pattern change during encoding of a narrative predicts retrospective duration estimates. *elife*, 5:e16070, 2016.

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.

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):405–431, 2019. ISSN 2326-831X. doi: https://doi.org/10.1146/annurev-statistics-030718-104938. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-030718-104938.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Warrick Roseboom, Zafeirios Fountas, Kyriacos Nikiforou, David Bhowmik, Murray Shanahan, and Anil K Seth. Activity in perceptual classification networks as a basis for human subjective time perception. *Nature communications*, 10(1):267, 2019.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=fU7-so5RRhW.

Alyssa H. Sinclair, Grace M. Manalili, Iva K. Brunec, R. Alison Adcock, and Morgan D. Barense. Prediction errors disrupt hippocampal representations and update episodic memories. *Proceedings of the National Academy of Sciences*, 118(51):e2117625118, 2021. doi: 10.1073/pnas.2117625118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2117625118.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=s1FjXzJ0jy.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TrjbxzRcnf-.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. In *Advances in Neural Information Processing Systems*, 2024. To appear.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

Jeffrey M Zacks. Event perception and memory. *Annual review of psychology*, 71:165–191, 2020.

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.

Jeffrey M Zacks, Christopher A Kurby, Michelle L Eisenberg, and Nayiri Haroutunian. Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of cognitive neuroscience*, 23(12):4057–4066, 2011.

Alexey Zakharov, Qinghai Guo, and Zafeirios Fountas. Long-horizon video prediction using a dynamic latent hierarchy. *arXiv preprint arXiv:2212.14376*, 2022a.

Alexey Zakharov, Qinghai Guo, and Zafeirios Fountas. Variational predictive routing with nested subjective timescales. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=JxFgJbZ-wft.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=RkRrPp7GKO.

A Appendix

A.1 Further Results

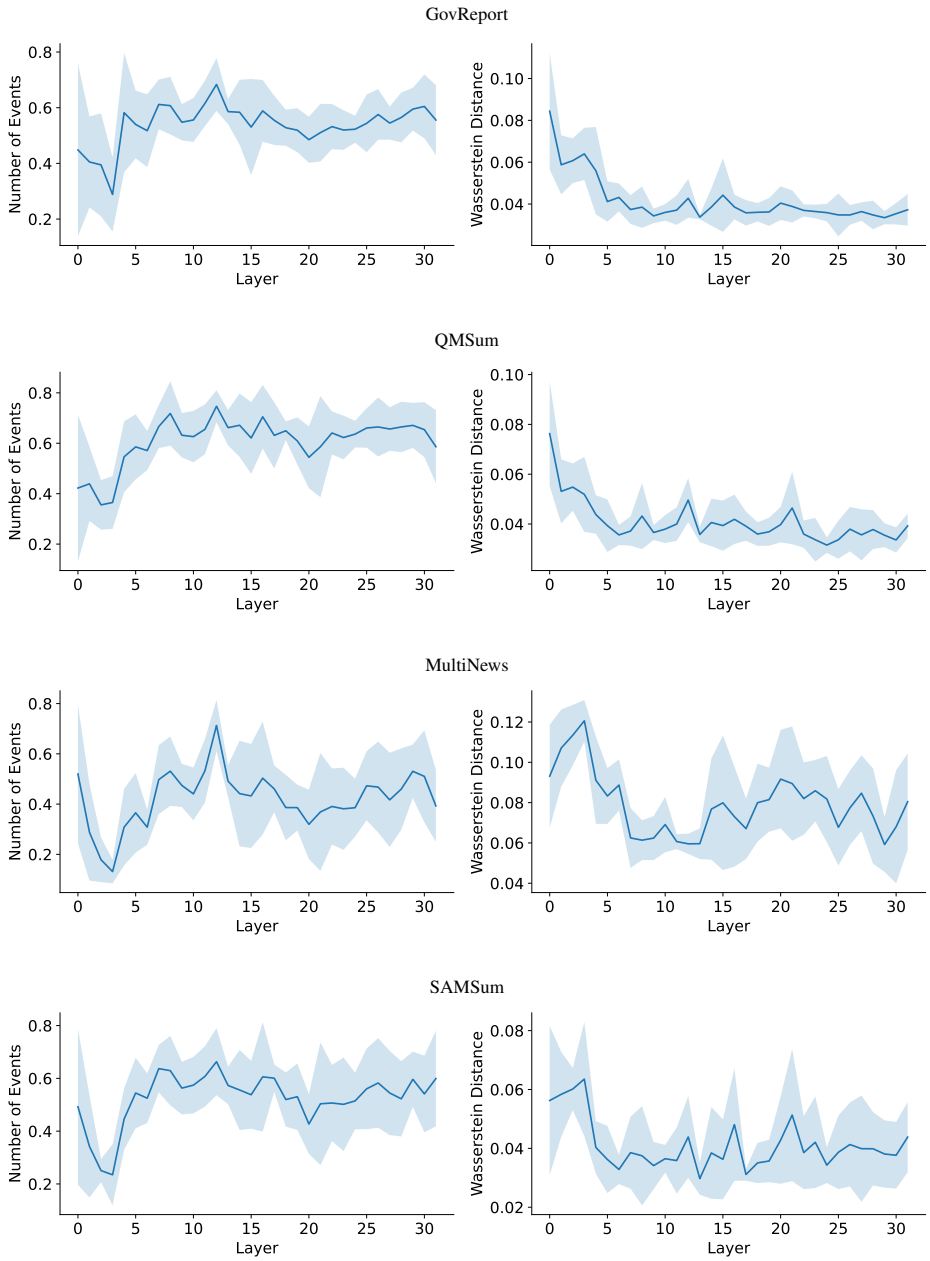

Figure 3: Visualisation of head-level K-Similarity segmentation properties averaged over examples from Long-Bench's summarisation tasks (200 examples each), using LLaMA-3-8B-Instruct. For each task, the left plot shows the average number of events found in the sequence (normalised) per head for each layer of the model, while the right plot shows the average Wasserstein distance between K-Similarity vs. (sequence-level) surprise per head. Standard deviation is measured across the heads of each layer in all figures.