Memory in humans and deep language models: Linking hypotheses for model augmentation

Omri Raccah
Intel Labs
New York University
or409@nyu.edu

Phoebe Chen
New York University
hc2896@nyu.edu

Ted L. Willke
Intel Labs
ted.willke@intel.com

David Poeppel
New York University
dp101@nyu.edu

Vy A. Vo
Intel Labs
vy.vo@intel.com

Abstract

The computational complexity of the self-attention mechanism in Transformer models significantly limits their ability to generalize over long temporal durations. Memory-augmentation, or the explicit storing of past information in external memory for subsequent predictions, has become a constructive avenue for mitigating this limitation. We argue that memory-augmented Transformers can benefit substantially from considering insights from the memory literature in humans. We detail an approach for integrating evidence from the human memory system through the specification of cross-domain linking hypotheses. We then provide an empirical demonstration to evaluate the use of surprisal as a linking hypothesis, and further identify the limitations of this approach to inform future research.

1 Introduction

Transformer model architectures [1] have become an indispensable tool in several domains, such as natural language processing [2], image processing [3] and reinforcement learning [4, 5]. A widely acknowledged scaling limitation of the self-attention mechanism is that its computational complexity scales quadratically with the size of the attention window. This limits the model's ability to capture long-range dependencies in data such as books, scientific articles, or code. Several efficient Transformers have been proposed to address this principal limitation [6]. A subset of these focus on augmenting the network with an external memory, which we henceforth refer to as memory-augmented Transformers [7–10]. Notably, knowledge in artificial neural networks (ANNs) is thought to be implicitly stored in the parameters of a pre-trained model, requiring ever-larger networks to store more facts [11, 12]. In memory-augmented Transformers, however, information is more explicitly stored in an external memory and retrieved when making predictions. This may increase the capacity of the network to represent knowledge about the world, in addition to helping it capture information over long temporal durations. Unlike temporal convolutions, which require a pre-specified width, external memories can be stored for arbitrary durations and retrieved when relevant.

This property is also true of human memory, which demonstrates the remarkable ability to generalize over an immense amount of information in written documents and over events in one's life. The rich literature on memory in psychology and neuroscience presents ample opportunities for augmenting ANNs with biologically-inspired memory. Here, we aim to lay some groundwork for understanding memory across fields, and describe some practical considerations for effectively integrating findings from cognitive neuroscience into Transformers, with a particular focus on language models (LMs). However, note that several of these considerations can be applied to other models and domains in AI.

Preprint. Under review.

Finally, we provide an empirical demonstration of evaluating a memory augmentation strategy for GPT-2 [13] using human behavioral data and identify its limitations and strengths to inform future research.

2 Considerations for biologically-inspired memory augmentation

2.1 Memory-augmentation from a cognitive lens

We argue that specifying appropriate linking hypotheses across domains will not only facilitate novel biologically-inspired approaches, but will also provide a way to empirically evaluate different hypotheses. In cognitive neuroscience, a linking hypothesis is a proposition for a formal mapping between a neurobiological state and a psychological state [14, 15], such as the firing of a single neuron leading to a visual percept. A central aim of biologically-inspired AI is to formulate linking hypotheses between a component in an AI system and a well-defined aspect of cognition. Strong linking hypotheses should lead to a formal and quantifiable mapping between a representation in an AI system and some neurobiological/psychological data, as has been demonstrated in some cases for computer vision [16–18] and natural language [19–22]. These linking hypotheses must be specified at the correct level of analysis [15, 23], e.g., a modification of the equations to perform similarity search on a database in a retrieval-augmented system should map to research on the biological mechanisms of memory retrieval. In our view, proper accounts would be best derived through decomposing the problem into computational subroutines appropriate for comparison across domains. Many AI systems already assume a linking hypothesis between ANNs and human cognition without explicitly stating them as hypotheses or evaluating them. Here, we briefly explore some of these hypotheses in memory-augmented Transformers and propose possible mappings to findings in the human literature.

We divide memory-augmented Transformers into two general types. A *static* memory stores information in a corpus of fixed size and content (e.g., a Wikipedia knowledge base), which it learns to retrieve from during training [24, 25, 9]. The contents of a static memory do not get modified, although they can be encoded in different formats such as raw text or embeddings [25, 26]. *Dynamic* memory mechanisms store new information as inputs that are being processed by the model. Training the network involves learning both storage and retrieval policies. For example, new information may be remembered or forgotten on the basis of input properties or model activations. Furthermore, inputs may be transformed in some manner (e.g., through compression) before being stored in the external memory [7]. Both static and dynamic memory-augmented Transformers have shown significant improvements over non-augmented models when making predictions over long texts [7, 8, 27, 28].

These augmentation strategies do not map cleanly to the types of memory commonly delineated in cognitive theories of human memory [29]. That said, classical memory taxonomies are often the source of AI inspiration, with papers citing work on short- vs. long-term memory or episodic memory [30, 31, 10, 11]. In our view, a static memory could be like human semantic memory if it uses a knowledge base, or it could be a fairly direct analog of episodic memory if it stores previously seen examples [24]. Instead, our proposed division focuses on the subprocesses thought to be involved in human long-term memory: encoding, consolidation, and retrieval [32]. Different strategies for memory augmentation will therefore pursue different implementations of each subprocess, and can draw direct inspiration from studies of that specific subprocess. Current work on memory-augmented Transformers has already proposed separate mechanisms for each subprocess, although there is often no direct link to human data. For example, there is a growing literature on retrieval augmentation [8, 9, 24, 25, 33–35] that proposes similarity search as the retrieval mechanism. Other work has proposed specific encoding policies which determine what to store and forget, either by exploiting the attention weights [7] or learning which memories to forget [36].

2.2 Incorporating insights from human memory via policy modifications

Here we discuss some findings from the human memory literature to demonstrate how they may be used to inform policy modifications in memory-augmented Transformers. Lexical properties (e.g., written-frequency, word length, animacy, etc.) serve as strong predictors of subsequent memory for individual words and lists [37–39]. Furthermore, humans have been shown to have the remarkable ability to remember whether they have seen an image from up to 10,000 images after only a single exposure [40]. The properties that determine the memorability of an image are thought to be multifaceted, including high-level properties such as emotional valence [41] and overall semantic meaning [42, 43]. If some property is directly computable from the inputs, it can be efficiently used as a biologically-plausible *encoding policy* in memory-augmented models. Recent work in cognitive neuroscience has also been focused on uncovering the process by which humans segment continuous experience into composite events in memory, known as event segmentation [44–47]. This evidence can also inform encoding policies for model augmentation, as studies have shown preferential encoding at event boundaries. Furthermore, this area of research can be leveraged to inform *storage policies*, which delineate how sequential information with ordered constituents is structured or formatted in memory. Lastly, *retrieval policies*, or the manner by which information is read from an existing memory store, can take practical influence from human memory. For example, items that share a temporal or semantic context during encoding are retrieved sequentially with relation to one another [48, 49]. These examples provide theoretically and empirically motivated hypotheses for memory-augmentation. Next, we demonstrate the evaluation of a specific linking hypothesis.

3 Evaluating a candidate linking hypothesis for memory augmentation

Surprisal The loss function of an LM estimates the negative log likelihood of an upcoming word given its context. In information theoretic terms, this is known as *surprisal*. Some have proposed that next-word prediction is a fundamental computational process that occurs during human language processing [50–52], and have shown evidence that LM-estimated surprise predicts behavioral [53–55] and neural data [51]. Surprise (or unsigned prediction error) is also theorized to play a critical role in memory and learning, and experimental evidence supports this notion [56–58]. Word surprisal in particular may predict human memory during natural language comprehension [59, 60]. Since surprise is a readily available quantity in LMs, we test its feasibility as a linking hypothesis by examining human behavioral data in a memory experiment. If model-based surprisal can predict human memory, it could be a practical and effective memory encoding policy for augmented models.

Dataset of human recall behavior We used a public dataset collected by Michelmann et al. from two groups of participants [61]. The first group ("story-exposure"; N = 50) listened to a naturally spoken story containing 965 words. Then participants completed a cloze task [62] similar to an autoregressive LM objective, in which they were given 10 words from the story and asked to predict the final word. This task was administered in order for every word in the story, starting with the third word, limiting the context for words in the beginning of the story (Appendix A). The second group ("no exposure"; N = 50) completed the same cloze task but had no exposure to the story before completing the cloze task. The memory effect is the difference in performance across groups. For a full account of the methods, see Appendix A.

For each word tested in the story, cosine similarity was computed between the GloVe embeddings [63, 61] for the responded word and the correct answer, and averaged across participants. In contrast to a binary scoring approach (correct vs. incorrect), this allows partial credit to be assigned for semantically similar responses to the correct answer (Appendix A). Replicating the findings in Michelmann et al. [61], we found that the story-exposure group significantly outperformed the no-exposure group in guessing the correct words ($p < 0.001$; one-tailed test; Figure 1).

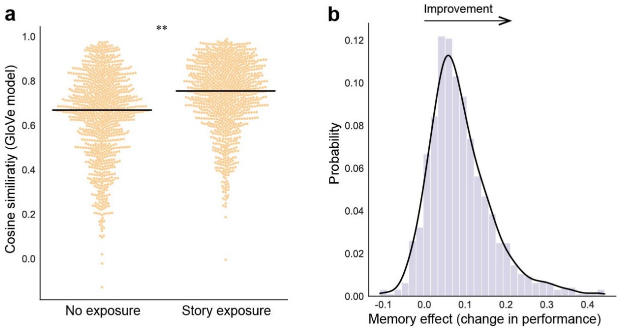

Figure 1: Behavioral results. (A) Cloze performance increases as a function of story-exposure across individual words. Black lines indicate the mean. (B) Histogram of memory improvement across words (signed difference between the story-exposure and no-exposure groups.)

3.1 Model-based word surprisal is related to human memory for spoken narratives

We next tested the effect of word surprisal on cloze performance. We used GPT-2 to estimate surprisal for each of the 1033 story tokens and combined sub-tokens for each word. We found that word surprisal shows a robust inverse correlation in both the no-exposure ($R^2 = 0.61$; $p < 0.001$) and the story-exposure group ($R^2 = 0.55$; $p < 0.001$; Figure 2A). This indicates that surprising words predict lower performance in the cloze task, regardless of prior experience with the story.

We find that GPT-2 estimated surprise is positively correlated with the effect of memory on cloze task performance, i.e., the signed difference between the story-exposure and no-exposure groups ($R^2 = 0.17$, $p < 0.001$), shown in Figure 2B. This effect was consistent for larger models with better perplexity, with only marginal differences in the overall correlation (GPT-2 medium: $R^2 = 0.13$, $p < 0.001$; large: $R^2 = 0.13$, $p < 0.001$; XL: $R^2 = 0.11$, $p < 0.001$).

However, it is possible that other properties that co-vary with surprise explain this memory effect. In particular, we evaluate how word written-frequency and distinctiveness affect the relationship between LM-based surprise and memory performance (Appendix Figure B.1). Measures of distinctiveness describe the uniqueness of a word's semantic associations, and have been shown to strongly drive lexical memorability [64]. We define distinctiveness as the average GloVe dissimilarity between a word and all other words in the story. Note that this represents story-specific distinctiveness, which differs from the measure in [64] that quantifies average dissimilarity between a word and all other words in a large corpus. Word frequency, i.e. the unigram probability of a given word, quantifies the overall exposure to a word in a large corpus of text [65]. We fit these features along with word surprisal as predictors in a multiple regression model to evaluate their overall and individual impact on memory performance. Importantly, this allows us to quantify the role of context in memory, while taking into account lexical features such as distinctiveness and word frequency. As expected, we found that the overall model significantly predicts memory performance, with a higher R^2 than word surprisal alone ($R^2 = .24$, $p < 0.001$). Each normalized β coefficient shows a significant contribution in predicting memory (distinctiveness $\beta = 0.011$, $p < 0.001$; word surprisal $\beta = 0.012$, $p < 0.001$; written-frequency $\beta = -0.02$, $p < 0.001$, Appendix Figure B.2). These findings suggest that word surprisal in GPT-2 predicts human memory performance for narratives, indicating its candidacy as an encoding policy.

To further understand the effect of context, we next examined how varying the context length given to GPT-2 affected its ability to predict memory performance. Because of the power-law scaling of language model performance as a function of context length [66], we selected roughly logarithmically spaced lengths. We found the correlation between surprisal and memory effect improves with longer windows, and plateaus at around 600-token length (Appendix Figure B.4). The correlations at all context lengths were significant using permutation tests ($p < 0.001$).

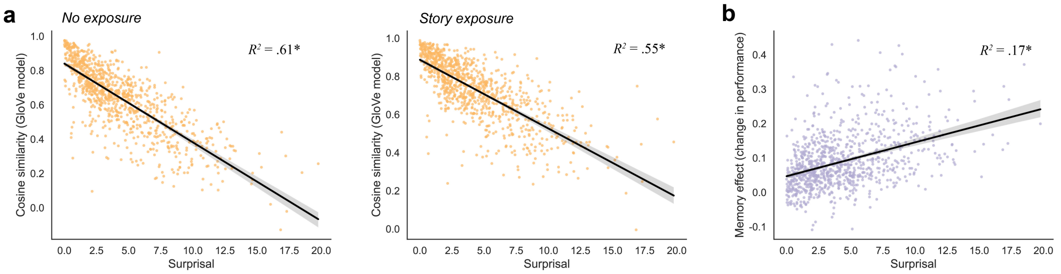

Figure 2: Predicting human behavior with GPT-2 surprise. (A) Model-based surprisal strongly predicts cloze performance, regardless of story exposure. (B) Surprisal is significantly correlated with memory performance for individual words in a spoken story.

3.2 Testing the ability of attention weights to predict human memory

Prior work has proposed that attention weights could serve as a memory encoding mechanism [7]. Extensive research on the role of attention in human memory [67–69] suggests this may be a feasible linking hypothesis, although see [70]. In our dataset, we tested whether memory performance was correlated with attention weights from several of the 12 layers, and found either no relation or an extremely weak one for layer 11 (Appendix Figure B.3).

4 Discussion and Future Directions

In this article, we detailed an approach for augmenting LMs on the basis of the human memory literature. We also provided an empirical demonstration for validating candidate mechanisms through a principled comparison with human behavioral data. In particular, we tested the hypothesis that word surprisal in GPT-2 would predict human memory performance in a narratives dataset [61], a prediction borne from the human memory literature more broadly [56–60]. We found that surprisal significantly correlated with human memory in this task, indicating its viability as a candidate encoding policy for future architectures. For example, surprisal could determine whether some context or word is written to an external memory. A similar encoding policy has been attempted in lifelong learning setups with mixed success [71, 72], suggesting that understanding the interaction of surprisal with other variables affecting memory may improve the policy [73].

Several other limitations should be identified in our empirical approach. First, the story that was used only consisted of 1033 tokens, which only exceeds the input size of GPT-2 by a small margin. Future work would benefit from using texts which significantly exceed 1024 tokens to better address temporal generalization limits in Transformer models. In addition, evaluating model performance on naturalistic free-recall data would provide a more ecologically valid benchmark for comparison [74]. Future work should also consider a more comprehensive evaluation of surprisal-related measures which can be generated using Transformer outputs or parameter states. In this work, we also define two types of memory-augmented Transformers, static and dynamic, which have direct consequences for successful modification. Research in this area would benefit by continuing to develop a taxonomy of model types which integrates explicit memory mechanisms to facilitate further progress.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[3] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM Computing Surveys*, 54(10s):200:1–200:41, September 2022.

[4] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A Survey of Reinforcement Learning Informed by Natural Language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. arXiv, June 2019.

[5] Luckeciano C. Melo. Transformers are Meta-Reinforcement Learners. In *Proceedings of the 39th International Conference on Machine Learning*. arXiv, June 2022.

[6] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A Survey. September 2020.

[7] 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*, September 2019.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37 Th International Conference on Machine Learning*, page 10, Vienna, Austria, 2020.

[9] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Augmenting Transformers with KNN-Based Composite Memory for Dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99, March 2021.

[10] Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. Adaptive Semiparametric Language Models. *arXiv:2102.02557 [cs]*, 9:362–373, February 2021.

[11] Aida Nematzadeh, Sebastian Ruder, and Dani Yogatama. On Memory in Human and Artificial Language Processing Systems. In *"Bridging AI and Cognitive Science" at International Conference on Learning Representations*, 2020.

[12] Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24.

[14] Davida Y Teller. Linking Propositions. *Vision Research*, 24(10):1233–1246, 1984.

[15] David Poeppel. The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, 29(1-2):34–55, March 2012.

[16] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, June 2016. Number: 1 Publisher: Nature Publishing Group.

[17] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J. Majaj, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 1147, pages 12805–12816. Curran Associates Inc., Red Hook, NY, USA, December 2019.

[18] Daniel LK Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. Publisher: Nature Publishing Group.

[19] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, August 2022. Publisher: Proceedings of the National Academy of Sciences.

[20] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. Publisher: Proceedings of the National Academy of Sciences.

[21] Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior, June 2020. arXiv:2006.01912 [cs].

[22] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014. Publisher: Public Library of Science San Francisco, USA.

[23] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, 1982.

[24] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. *arXiv:1911.00172 [cs]*, February 2020.

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[26] Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. Entities as Experts: Sparse Memory Access with Entity Supervision. *arXiv:2004.07202 [cs]*, October 2020.

[27] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv:1901.02860 [cs, stat]*, January 2019.

[28] Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing Transformers. *arXiv:2203.08913 [cs]*, March 2022.

[29] Endel Tulving. Episodic and semantic memory. In *Organization of Memory*, pages xiii, 423–xiii, 423. Academic Press, Oxford, England, 1972.

[30] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. *arXiv:1410.5401 [cs]*, October 2014.

[31] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, October 2016.

[32] Dale Purves, George J. Augustine, David Fitzpatrick, William C. Hall, James O. McNamara, and S. Mark Williams, editors. Neuroscience, 3rd ed. Sinauer Associates, Sunderland, MA, US, 2004.

[33] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-Augmented Diffusion Models. *arXiv:2204.11824*, April 2022.

[34] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2206–2240. PMLR, June 2022.

[35] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. KNN-Diffusion: Image Generation via Large-Scale Retrieval. *arXiv:2204.02849*, April 2022.

[36] Sainbayar Sukhbaatar, Da Ju, Spencer Poff, Stephen Roller, Arthur Szlam, Jason Weston, and Angela Fan. Not All Memories are Created Equal: Learning to Forget by Expiring. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9902–9912. PMLR, July 2021.

[37] C. Hulme, S. Roodenrys, R. Schweickert, G. D. Brown, M. Martin, and G. Stuart. Word-frequency effects on short-term memory tasks: evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 23(5):1217–1232, September 1997.

[38] Earl Y. Popp and Michael J. Serra. The animacy advantage for free-recall performance is not attributable to greater mental arousal. *Memory (Hove, England)*, 26(1):89–95, January 2018.

[39] Ada Aka, Tung D. Phan, and Michael J. Kahana. Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47:765–784, 2021. Place: US Publisher: American Psychological Association.

[40] Lionel Standing. Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, 25(2):207–222, May 1973.

[41] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and Predicting Image Memorability at a Large Scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[42] Wilma A. Bainbridge. Memorability: How what we see influences what we remember. In *Psychology of Learning and Motivation*, volume 70, pages 1–27. Elsevier, 2019.

[43] Nicole C. Rust and Vahid Mehrpour. Understanding image memorability. *Trends in Cognitive Sciences*, 24(7):557–568, July 2020.

[44] Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001.

[45] Zhilin Wang, Anna Jafarpour, and Maarten Sap. Uncovering Surprising Event Boundaries in Narratives. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 1–12, Seattle, United States, July 2022. Association for Computational Linguistics.

[46] Omri Raccah, Keith B Doelling, Lila Davachi, and David Poeppel. Acoustic features drive event segmentation in speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2022.

[47] Manoj Kumar, Ariel Goldstein, Sebastian Michelmann, Jeffrey M Zacks, Uri Hasson, and Kenneth A Norman. Bayesian surprise predicts human event segmentation in story listening. *PsyArXiv*, Sep 2022.

[48] Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of mathematical psychology*, 46(3):269–299, 2002.

[49] Jeremy R Manning and Michael J Kahana. Interpreting semantic clustering effects in free recall. *Memory*, 20(5):511–517, 2012.

[50] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021.

[51] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022.

[52] Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, February 2020.

[53] Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah, January 2018. Association for Computational Linguistics.

[54] Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *CoRR*, abs/2009.03954, 2020.

[55] Cassandra L. Jacobs and Arya D. McCarthy. The human unlikeness of neural language models in next-word prediction. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, page 115, Seattle, USA, July 2020. Association for Computational Linguistics.

[56] Alyssa H. Sinclair and Morgan D. Barense. Surprise and destabilize: Prediction error influences episodic memory reconsolidation. *Learning & Memory*, 25(8):369–381, August 2018.

[57] Meadhbh I. Foster and Mark T. Keane. The Role of Surprise in Learning: Different Surprising Outcomes Affect Memorability Differentially. *Topics in Cognitive Science*, 11(1):75–87, 2019.

[58] James W. Antony, Thomas H. Hartshorne, Ken Pomeroy, Todd M. Gureckis, Uri Hasson, Samuel D. McDougle, and Kenneth A. Norman. Behavioral, Physiological, and Neural Signatures of Surprise during Naturalistic Sports Viewing. *Neuron*, November 2020.

[59] Richard Futrell, Edward Gibson, and Roger P. Levy. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3):e12814, 2020.

[60] Katja I. Haeuser and Jutta Kray. Effects of prediction error on episodic memory retrieval: Evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, 0(0):1–17, May 2021.

[61] Sebastian Michelmann, Amy R. Price, Bobbi Aubrey, Camilla K. Strauss, Werner K. Doyle, Daniel Friedman, Patricia C. Dugan, Orrin Devinsky, Sasha Devore, Adeen Flinker, Uri Hasson, and Kenneth A. Norman. Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature Communications*, 12(1):5394, September 2021.

[62] Wilson L. Taylor. “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433, September 1953.

[63] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[64] Greta Tuckute, Kyle Mahowald, Phillip Isola, Aude Oliva, Edward Gibson, and Evelina Fedorenko. Intrinsically memorable words have unique associations with their meanings. *PsyArXiv*, February 2018.

[65] David A. Balota, Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459, August 2007.

[66] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020.

[67] Marvin M Chun and Nicholas B Turk-Browne. Interactions between attention and memory. *Current opinion in neurobiology*, 17(2):177–184, 2007.

[68] Mariam Aly and Nicholas B Turk-Browne. Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, 113(4):E420–E429, 2016.

[69] Mariam Aly and Nicholas B Turk-Browne. How hippocampal memory shapes, and is shaped by, attention. In *The hippocampus from cells to systems*, pages 369–403. Springer, 2017.

[70] Grace W. Lindsay. Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, 14, 2020.

[71] Tiago Ramalho and Marta Garnelo. Adaptive posterior learning: Few-shot learning with a surprise-based memory module. In *International Conference on Learning Representations*, page 14, 2019.

[72] Cyprien de Masson d’Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic Memory in Lifelong Language Learning. *arXiv:1906.01076 [cs, stat]*, June 2019.

[73] Nina Rouhani, Kenneth A. Norman, and Yael Niv. Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(9):1430–1443, September 2018.

[74] Liberty S. Hamilton and Alexander G. Huth. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, June 2020.

[75] Lynn J. Lohnas and Michael J. Kahana. Parametric effects of word frequency effect in memory for mixed frequency lists. *Journal of experimental psychology. Learning, memory, and cognition*, 39(6):1943–1946, November 2013.

[76] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

A Supplemental Materials and Methods

Dataset To evaluate model parameters with respect to human memory performance, we used a previously published behavioral dataset [61] that was made publicly available through a Creative Commons 4.0 license. This dataset contains data collected online in which an experimental group of participants listened to a spoken story before performing a behavioral task. All participants provided informed consent in accordance with the local Institutional Review Board.

The behavioral experiment was run online (Amazon's Mechanical MTurk) across two groups of volunteer participants [61]. In total, 100 participants were collected. We used anonymized data from these participants in the current study ("replication" dataset in [61], which is the only shared dataset that provides participant-level results).

Story materials Participants listened to a humorous story which is 7 minutes and 30 seconds in duration and contained 965 words (recorded live as part of the "The Moth" storytelling event in New York City; "Pieman" by Jim O' Grady). The only offensive content in the story is the use of occasional swear words. The transcript for this story was provided as part of the public dataset release.

Experimental procedures All participants took part in a cloze task [62], which provided participants with text from the story and asked them to guess the next word. The cloze task is widely used in the psychological literature to estimate the probability of the next word given some context, making this behavioral design particularly suitable for comparison with probabilistic predictions computed from language models. In the current paradigm, participants saw 10 words from the transcribed story and asked to guess the word that followed (continuing for every word in the story). The task starts with the third word in the story, limiting the context for words in the beginning of the story [61]. In order to capture memory for the story, participants were split into two experimental groups. In the first group (N = 50), participants listened to the story once before performing the cloze task ("story-exposure group"), while the second group (N = 50) did not listen to the story before performing the cloze task ("no-exposure group"). The signed difference in performance between the story-exposure group and the no-exposure group represented the improvement in performance as a function of recalling the story from memory. The paradigm and experimental procedures are described in greater detail in [61].

Behavioral data analysis As acquired from [61], cloze task performance was computed using GloVe vector embeddings (pre-trained on Common Crawl data [63]). Specifically, participants' word predictions were scored by computing cosine-similarity between the vector embeddings for the predicted word and the correct word. The average semantic similarity among participants for a particular word represents the group-level performance for that word. We compared these word similarity scores across the story-exposure and no-exposure groups to evaluate differences in performance as a function of story recollection. Notably, this approach provides a more continuous measure of performance as opposed to a binary dependent measure (correct or incorrect) as used in the main text of Michelmann et al. ([61]; but see Michelmann Supplementary Method 2 and Supplementary Note 3). In other words, this measure of performance can be thought to provide a more sensitive measure than binary scoring, in that some credit can be assigned in the case that a participant predicts a semantically similar, but not identical, word to the correct answer.

Word surprisal as negative-log likelihood We computed word surprisal for each word in the story using GPT-2 [13], with the GPT-2 tokenizer and model from Huggingface [2]. For each of the 1033 sub-word tokens in the story, we included all previous tokens as preceding context (up to the maximum 1024). This ensured that we maximized contextual information for each token. We then take the cross entropy loss for the last token.

The surprisal for each word in the story is represented by the negative log likelihood generated by the loss function:

$$S_{model}\left(x_i\right) = -logP_{model}\left(x_i \,|x_{j<i}\right)$$

To compute the overall negative log likelihood of a word, we simply sum the negative log likelihood values of the component sub-tokens. As such, a higher value represents a more surprising word given its context according to the model. We refer to the inverse log likelihood at the word-level as word surprisal.

Factoring in word frequency and distinctiveness Past research has shown that both word frequency [75] and distinctiveness [64] have a significant impact on word retention. We sought to understand to what extent these factors along with word surprisal predict memory performance in the current dataset. We acquired the word written-frequency for each word in the story using the Hyperspace Analogue to Language (HAL) frequency norms [65]. Word-distinctiveness, or how few unique word associations a word possesses, was computed using GloVe word embeddings [76]. In particular, we computed the cosine-similarity between each word in our story and all of the other words in the story. Taking one minus the average of these similarity values provided us with a story-level distinctiveness value for each word, such that higher values represent more distinctive words. We fit word written-frequency and distinctiveness along with word surprisal as predictors in a general linear model (GLM) to predict memory performance. Notably, words for which written-frequency was not available [65], or for which we were not able to compute distinctiveness, were excluded from this analysis (19 words excluded in total)

Statistical testing Throughout this work, statistical analysis was applied using nonparametric permutation tests. Across comparisons, we implemented 10,000 permutations to ensure a reliable estimate of the null distribution. Significance was evaluated at $p < 0.05$. Note, we indicate the use of a one-tailed test when an effect is evaluated in a specific direction, otherwise a two-tailed statistic is reported.

B Supplemental Results

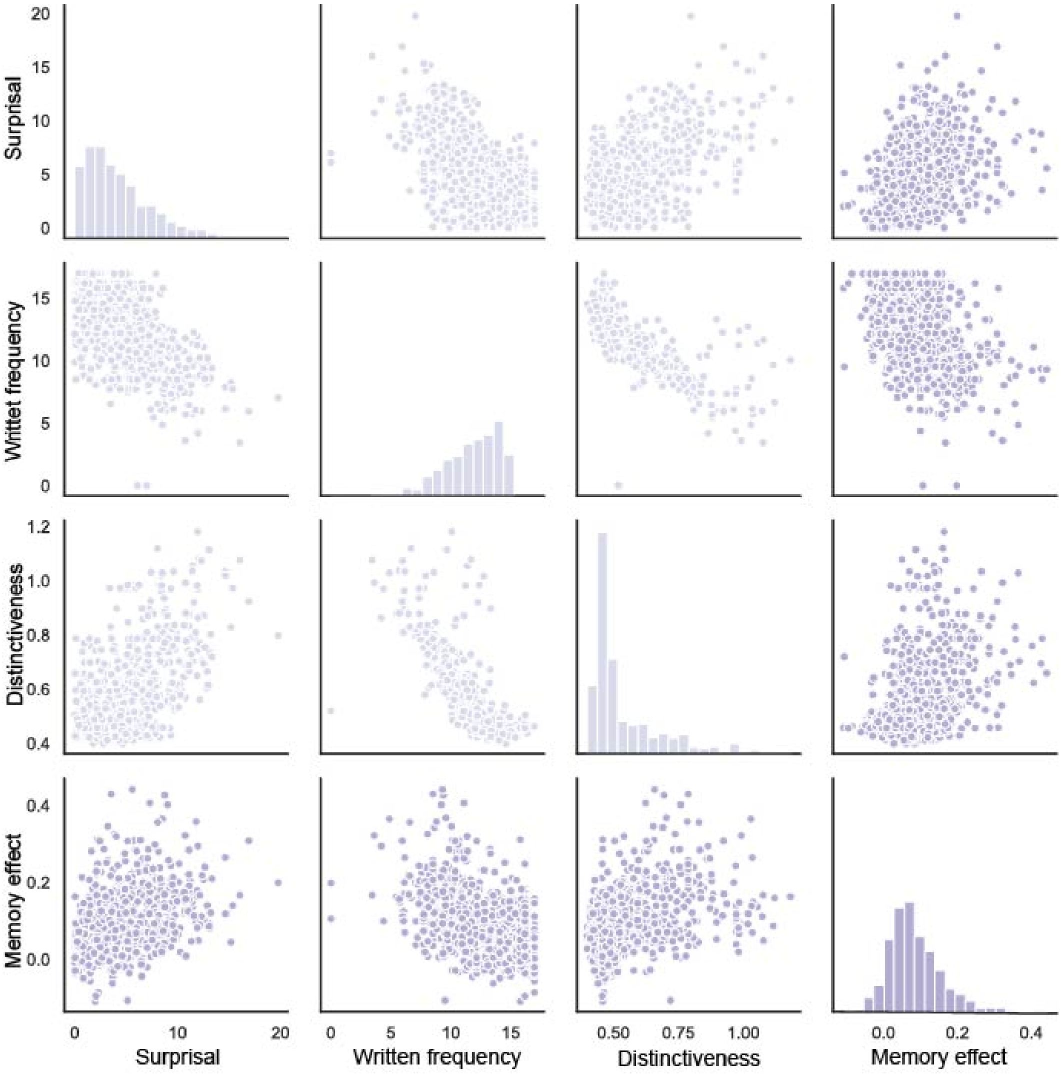

Figure B.1: Relationships across predictors and memory effect (cosine similarity) in multiple regression analysis.

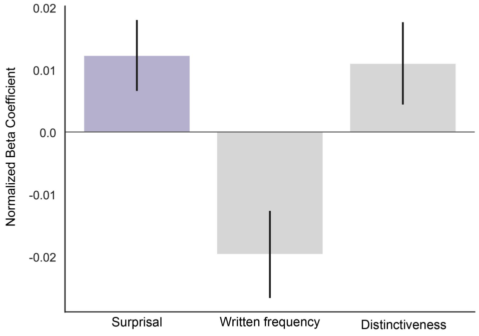

Figure B.2: Normalized beta coefficients for predictors in multiple regression analysis. Error bars represent 95% bootstrapped confidence intervals (10,000 bootstrap iterations).

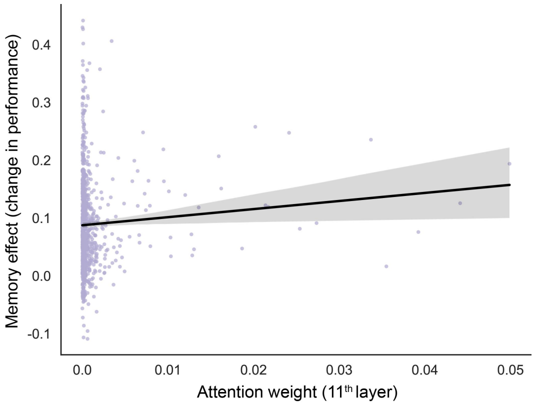

Figure B.3: Attention weights for 11th layer in GPT-2 versus memory effect ($R^2 = 005$; $p = 0.036$). We did not find a significant effect in the 1st, 6th, and 12th layers ($p > 0.05$).

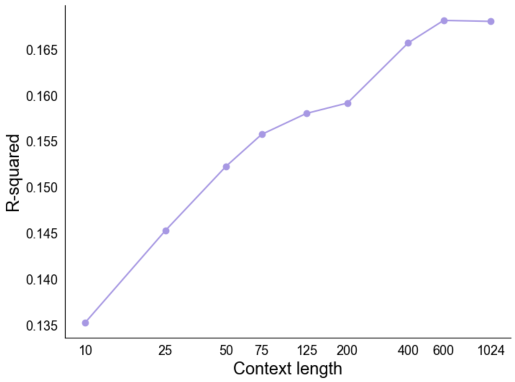

Figure B.4: R-squared for Pearson correlation between the memory effect and GPT-2 surprisal as a function of input window sizes. Permutation tests indicate all correlations are significant ($p < 0.001$).