Neuroscience Implications And Comparisons Within Transformer Model Architecture

Chris Vaisnor
Graduate Student - Artificial Intelligence M.S.
cvaisno1@jh.edu

The Johns Hopkins University
Baltimore, MD

Abstract—**This paper showcases recent findings within neuroscience literature and implicates biological utility to the transformer model. A basic overview of the model's main components and self-attention heads are provided. One parameter and two components are proposed within the architecture that follow attention and memory research done in human studies. These are Memory Decay, Exposure Count, and Context Look-Up, respectively. This is an effort to improve the structure of the transformer as well as discuss components that could be necessary to its application in Artificial General Intelligence (AGI).**

I. Introduction

A transformer is a type of deep neural network that incorporates abundant use of self-attention heads and feed forward networks to process and forecast a sequential style of input and output. This model was first proposed in December of 2017 from researchers with the Google Brain team. The original model and corresponding research paper [1] has become the playbook for designing any modern natural language processor (NLP), as well as specific cases using computer vision (CV). *This paper assumes general knowledge of deep neural networks and familiarity with attention-heads used in neural-machine translation.*

Among relevant research, there have not been neurological ties to the architecture of the model. It is the hypothesis of this paper that comparing biological hardware systems and processes to the transformer self-attention mechanism could lead to findings that improve the specific and general performance of the model.

It appears to be a coincidence that the best performing model shares similarities with our main cognitive function. As a matter of experience, what you pay attention to is who you are. Humans shine this attention spotlight on different aspects of sensory input, whether voluntary or involuntary. It is how we interact with the world and how we decide on what is important.

Attention is not just a singular act, but a multifaceted hierarchy within neuronal processing.

> The effectiveness of attention (relies) on much more than whether the spotlight is focused or not. It is also critical when, where and how long the spotlight is wielded. These three aspects of attention that build on its selectivity are known as expectation, directionality, and sustainability [2].

The self-attention mechanism within the transformer model has been aptly named. The transformer decides what parts of the input sequence to attend to, with an effectively infinite time-horizon. The goal of this paper is to review the competencies of human neurological function with respect to top-down attention modulation and provide ideas for integrating those competencies within the transformer model.

II. Brief Overview of Transformer Architecture

For a transformer based system, it can be better to think about the inputs and outputs in terms of tensors with varying dimensionality. (A reminder of tensor rank: zero (0) is a scalar, one (1) is a vector, two (2) is a matrix, three (3) is a cube, etc...) These tensors, which traditionally represent word tokens (vectors), are passed through layers in the network.

In the general model (Fig. 1 - left), the left box represents the encoder and the right box represents the decoder. In the original paper [1], the researchers stacked the encoder and decoder layers six (6) times.

Because there is no recurrent structure, there needs to be a way for the model to see location information for each token. Therefore, a positional tensor must be

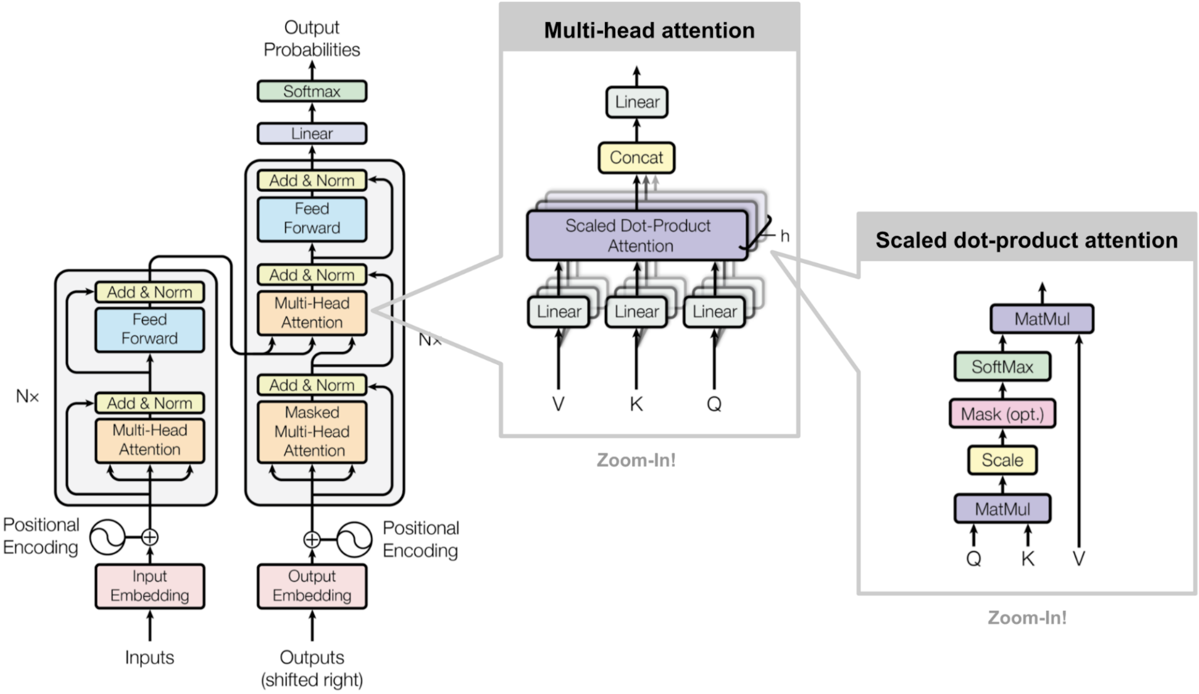


Fig. 1. The Original Transformer - Expanded View
[3]

added to each word tensor before it is passed into the first layer. This positional tenor is a relative position based on a sin and cosine function. For even position tensors a sine wave is added, and odd positions use a cosine wave. The position of the tensor i, location dimension pos, and model dimension d_{model} are passed to these functions.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

In the multi-headed self-attention layer, these tensors are split into the Query (Q), Key (K), and Value (V) inputs and are multiplied against each other in various orders and combinations. Each combination outputs a new tensor which can be thought of as the attention score. (Fig. 2)

The attention score matrix that results from the multiplication can be thought of as a self-referential attention process. In other words, the input is being multiplied against itself and creating the attention matrix. This happens in multiple parallel layers with variations in initialization weights in an attempt to find different interpretations of the same input. In the original paper [1], eight (8) Scaled Dot-Product attention heads were used per encoding layer. (Fig. 1 - middle)

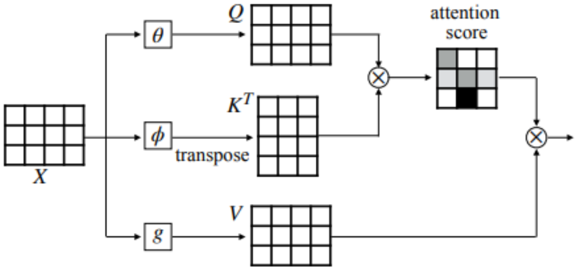


Fig. 2. Dot-Product Self-Attention
[4]

On the decoder side (Fig. 1 - left), each layer has two attention heads that receive different input. The first decoder layer receives the target as input to its masked attention head. This does exactly the same as a regular attention head except tensors beyond the current iteration are hidden.

> We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on

the known outputs at positions less than i [1].

The second attention head of the decoder layer receives the encoder output as its Key (K) and Value (V) inputs, with the masked attention head output taking the place of the Query (Q). All subsequent decoder layers receive the encoder output as well as input from the decoder before it. (Fig. 3)

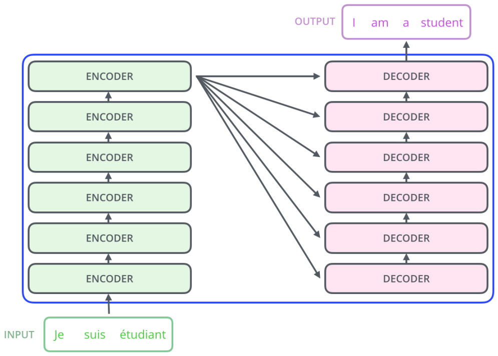


Fig. 3. Encoder-Decoder Relationship
[5]

There are other feed-forward and normalization layers within each encoder and decoder layer, but for the purposes of this paper they are not going to be examined. When the model is being used for inference instead of training, the thought is that the encoder and decoder weights are sufficiently trained to be able to compute an correct interpretation of the target without having seen it.

III. Attention and Memory In Neuroscience

The source of our cognitive functioning, our brain and subsystems, must distill the most important information. When this process is done actively, this is considered top-down modulation.

> Top-down modulatory functions dynamically modulate neuronal excitability both in the presence of stimuli, i.e., during selective encoding of items to be remembered and selective retrieval of a memorandum, as well as in the absence of external stimuli, i.e., in expectation of items to encode or ignore and during maintenance of items during a temporal delay [7].

While deciding on what to focus on, we are constantly accessing and retrieving memory states that we encoded in distant brain regions.

This is in contrast to bottom-up attention which generally comes from environmental stimuli that is passively observed. Under normal waking consciousness, we discount less-than-novel stimulation in order to direct our goals via top-down processes. If we encounter a new stimulus, our bottom-up attention throws a novelty response and top-down attention takes over. This is why humans can pay attention to a podcast and drive a car at the same time. However, if there is a significant bottom-up event, our top-down attention is pulled to the new stimulus and we lose focus on the other task. [8]

> (Patterns of activity) provide bias signals to other brain structures whose net effect is to guide the flow of activity along neural pathways that establish the proper mappings between inputs, internal states, and outputs needed to perform a given task.” Thus, the cognitive control needed to enact our goals is manifested by higher-order representations in the prefrontal cortex that result in the top-down modulation of neural activity (...) [2].

It’s clear that there is a hierarchy of neuronal structures that play off each other for specific tasks. If a task uses physical movement, that region is activated. If a task requires heavy use of language, that region is activated. We are constantly accessing different regions via control from the prefrontal cortex.

Findings within the last 10 years also show the utility of ignoring stimuli. In research done by the Gazzaley Lab at the University of California, San Francisco, they found the ability to ignore information as important as the ability to decide on what is useful.

> This experiment revealed that focus was not the primary determinant of high-level working memory performance; rather, memory depended more on effectively ignoring distractions [2].

Human creativity also relies on these distant connections across the brain.

> The recurrent theme that epitomizes the creative process is not generating something brand-new out of the blue. Rather, a creative spark occurs when unexpected associations among existing elements are suddenly forged—a sort of cognitive alchemy [9].

As different neural components are accessed, the prefrontal cortex can find stability among the noise and see associations between distant ideas; and memory holds a large portion of the structure together. Memory and its various types also play a key role in the foundation of cognitive processes. Intellectual work would not be possible without a memory frame to reference.

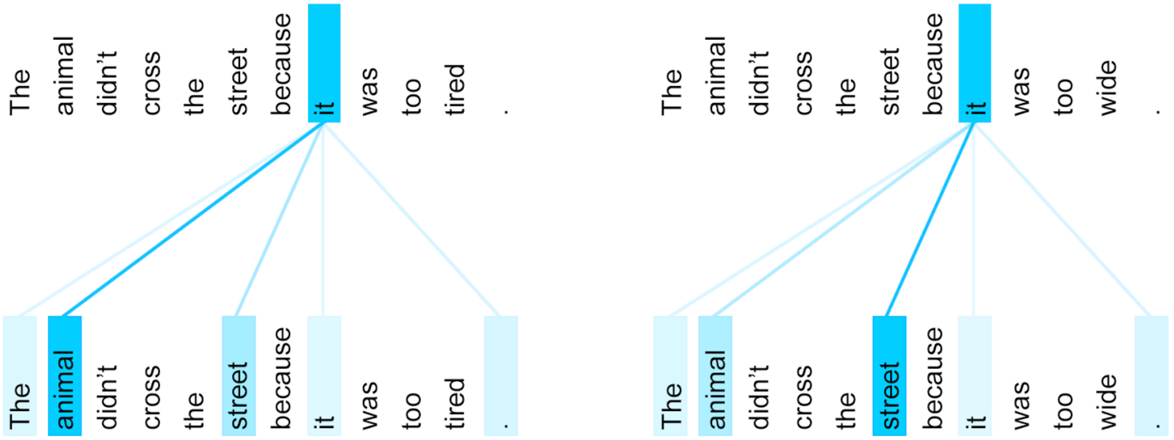


Fig. 4. Token Relationship Map
[6]

The output of the encoder layer acts as a memory structure. The attention head weights are complex enough in dimension to be fine tuned and attend to specific tokens. Is it possible to do better by using our findings of neuronal attention processes?

IV. Where can things be improved?

Some of humanity's greatest accomplishments have come from impressions we have found in nature. The neural network structure is a (basic) attempt to mimic our biological neuronal processing. The transformer model is a breakthrough in computational processes and reflecting on our own biology is a productive step as we continue to develop our intelligent machines.

While impressive, these transformer models are far from general intelligence agents. They are able to distinguish between the importance of specific words, but not due to the word's actual meaning. At their core, these models are sophisticated language mimics. Brief examples are provided in Fig. 5 and Fig. 6 using the OpenAI playground where you can interact with a variety of transformer models.

In Fig. 5, the use of past tense could have influenced the model to pick a previous President's name, even if the question does not follow coherent logic. A more intelligent response would have been to clarify the initial question, or to comment that it's impossible to know who the president will be in the year 2100.

Because the models lack context and tangible word understanding, they can be fooled with incoherent logic even when the sentence is grammatically correct. The sentence could look similar to possible text it has seen before, but changing a single word can change the interpretation.

In Fig. 6, the model is sure the answer is a bookmark, rather than responding that turning a page of the "banana" I'm reading is not possible. If the word was switched to "book", it would be sound logic. Because there is no context implied within the actual words, the model follows a standard structure for the other 18 words in the sentence.

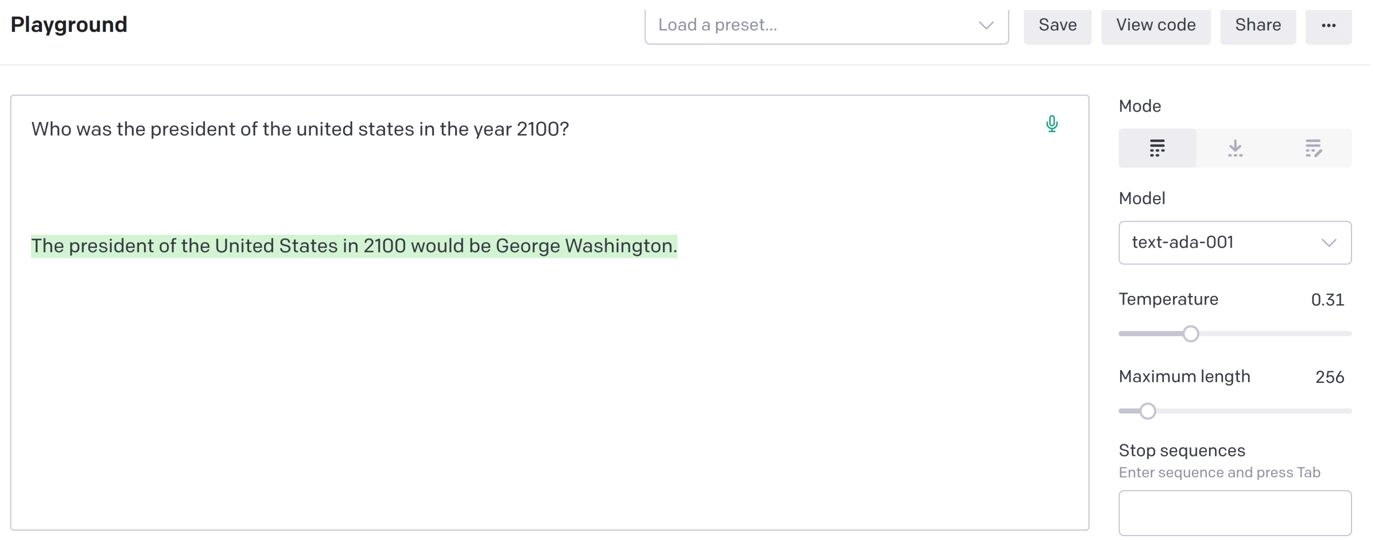
Using attention and memory research from neuroscience concepts outlined above, I have outlined three potential factors that seek to improve the transformer model.

1) Memory Decay
2) Exposure Count
3) Context Look-Up

A. *Memory Decay*

In research done by Dr. Scott Small at the University of Columbia, Alzheimer's disease has provided information on how cognitive flexibility works. In all animals, key molecules within cortical memory structures are extremely similar. The ability to forget has been precluded within intellectual processing among biological life.

> By testing different computer algorithms, computer scientists have learned that adding more memory—the equivalent of adding more dendritic spines—will not improve pattern recognition of faces or of anything else. Instead, the more effective way to artificially create



Who was the president of the united states in the year 2100?

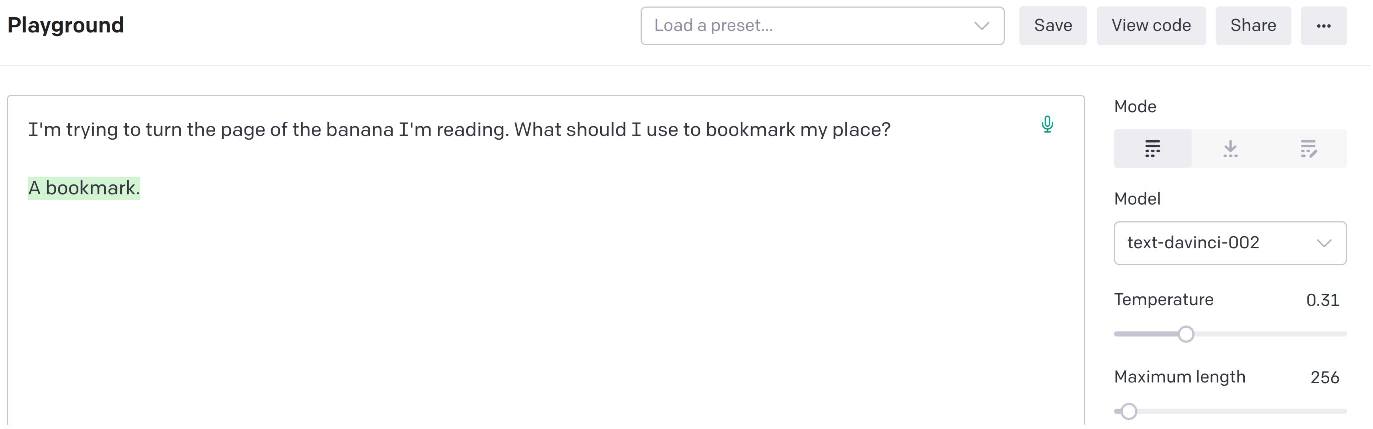
The president of the United States in 2100 would be George Washington.

Model: text-ada-001

Temperature 0.31

Maximum length 256

Fig. 5. OpenAI Playground with text-ada-001
https://beta.openai.com/playground?model=text-ada-001



I'm trying to turn the page of the banana I'm reading. What should I use to bookmark my place?

A bookmark.

Model: text-davinci-002

Temperature 0.31

Maximum length 256

Fig. 6. OpenAI Playground with text-davinci-002
https://beta.openai.com/playground?model=text-davinci-002

human computational flexibility is to force the algorithm to have more forgetting [9].

Currently, the transformer model can operate its attention method on the entire sequence of input. Due to memory and processing constraints, most models have a maximum input length of around 2000 tokens. Allowing for a parameter that discounts tokens located further apart could potentially increase the model's cognitive flexibility.

This parameter should be tuned dependent on different goals. If the attention matrix shows a token with a high attention score located at the beginning of a long input, this parameter could be used to slightly devalue that score. This would prevent the model from getting stuck on any specific long-distanced token.

B. Exposure Count

In regards to human attention, repeated exposure to a stimulus is the basis of familiarity. There are other subtle findings such as the Mere-exposure Effect, also known as the Familiarity Principle that distinguish our interactions and preferences for stimuli we have seen before. [10] Giving the model the ability to keep track of similar tokens could provide a foundation for further improvements.

This could be an additional component next to the encoder and decoder layers that updates with training. This could count tokens and then later be referenced depending on the desired context. Cosine similarity could be used to measure potentially similar tensors within the initial attention-head inputs. As the training corpus is iterated through, an exposure counter could be useful for the developers as well. After training, semantic analysis could even be done based on the count of specific words. This could give additional insight to test if the model is biased towards using specific tokens.

C. Context Look-Up

In a similar way to keeping track of exposure to tokens, tokens with large attention values could be marked as valuable. This follows neuroscience findings when semantic information is encoded with emotional information.

> If factual information is processed and coded by the central hubs in the cortex, the amygdala can be considered a subcortical central hub that processes and codes emotional information [9].

The attention head helps the model decide what to look at, but it does not help the model with the tangible meaning. Researchers assume that with enough training data, the transformed tokens inherit meaning from their location within the text; and this is likely true to an extent. It is a brute force way to catch meaningful information.

If there was a second component that encoded explicit information about known high-attention tokens, the model's overall performance could improve. Humans have developed additional methods for storing attention and emotion states for future decision making.

> (Choose) words that are precise and specific when describing what we feel. Accurately distinguishing among interoceptive sensations is associated with making sounder decisions, acting less impulsively, and planning ahead more successfully—perhaps because it gives us a clearer sense of what we need and what we want [11].

Being highly descriptive is useful, and the large corpora that these models are trained on is not guaranteed to be high quality.

Being more specific when interacting with large transformer models already greatly increases accuracy. As recently as October 2022, research done by the University of Tokyo in conjunction with the Google Brain team found that simply adding the phrase, "Let's think step by step", into their natural language prompts increased accuracy up to 300 percent [12].

This Context Look-Up component could be a hashmap with key-value pairs being the token and its closest dictionary definition. The model could then see a nested sentence where that specific token is replaced or supplemented with its definition. This would fix the earlier example (Fig. 6) using a banana and bookmark. If the model had access to the explicit information about the "banana" token, it would have been more equipped to provide a realistic answer.

If the key-value map is still insufficient, then a more sophisticated component could allow read-access to a database with greater information and examples. The database would act as the memory framework and the user could set an attention score threshold for when the model calls a query to the database. This has the added benefit of not increasing training complexity but a sufficient database would need to be built.

It is the opinion of this paper that this Context Look-Up component would prove the most useful. This creates a type of metamemory for the model,

> Metamemory is defined as how well our subjective sense of our memory ability matches an objective measure of it [9].

It would not only have training weights as memory parameters, but higher quality nested input strings as well.

V. Conclusion and Looking Forward

Neuroscience has a lot to offer machine learning engineers in the future. The human brain is the greatest reference point we have for continuing the development of deep neural networks. There are plenty of other issues that need to be discussed such as prompt-injection attacks and multi-month training times. There is some consensus among NLP experts that the current models we have are already over-scaled,

> DeepMind found that all super-large models are "significantly undertrained." They're unnecessarily big [13].

There are few meaningful circumstances where quantity is better than quality. As important as AI will inevitably be to human civilization, we should not get lost in powerful hardware and forget the refinement of the models running on it. The powerful hardware used for these training sessions are hitting their own constraints as well. Training GPT-3 in 2020 cost between four (4) and twelve (12) million dollars and took weeks depending on the configuration [14]. Other larger models can take a magnitude longer. Working and refining what we have could be the next step in transformer model development.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Gazzaley and L. D. Rosen, *The distracted mind: Ancient brains in a high-tech world*. Mit Press, 2016.

[3] L. Weng, "The original transformer," June 2018. https://lilianweng.github.io/posts/2018-06-24-attention/.

[4] P. Singh, "Multi-head self-attention in nlp," May 2020. https://blogs.oracle.com/ai-and-datascience/post/multi-head-self-attention-in-nlp.

[5] J. Alammer, "The illustrated transformer," June 2018. https://jalammar.github.io/illustrated-transformer/.

[6] J. Uszkoreit, "Attending to a single word," August 2017. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html.

[7] A. Gazzaley and A. C. Nobre, "Top-down modulation: bridging selective attention and working memory," *Trends in cognitive sciences*, vol. 16, no. 2, pp. 129–135, 2012.

[8] S. Harris and A. Gazzaley, "226: The price of distraction," November 2020. https://www.samharris.org/podcasts/making-sense-episodes/226-price-distraction.

[9] S. A. Small, *Forgetting: The Benefits of Not Remembering*. Crown, 2021.

[10] T. D. Lab, "Why do we prefer things that we are familiar with?," 2020. https://thedecisionlab.com/biases/mere-exposure-effect.

[11] A. M. Paul, *The extended mind: The power of thinking outside the brain*. Eamon Dolan Books, 2021.

[12] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[13] A. Romero, "Ultra-large ai models are over," Oct 2022. https://albertoromgar.medium.com/ultra-large-ai-models-are-over-c7b5fc007d0a.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.