To model human linguistic prediction, make LLMs less superhuman

Byung-Doh Oh and Tal Linzen
New York University
{oh.b, linzen}@nyu.edu

Abstract

When people listen to or read a sentence, they actively make predictions about upcoming words: words that are less predictable are generally read more slowly than predictable ones. The success of large language models (LLMs), which, like humans, make predictions about upcoming words, has motivated exploring the use of these models as cognitive models of human linguistic prediction. Surprisingly, in the last few years, as language models have become better at predicting the next word, their ability to predict human reading behavior has declined. This is because LLMs are able to predict upcoming words much better than people can, leading them to predict lower processing difficulty in reading than observed in human experiments; in other words, mainstream LLMs are 'superhuman' as models of language comprehension. In this position paper, we argue that LLMs' superhumanness is primarily driven by two factors: compared to humans, LLMs have much stronger long-term memory for facts and training examples, and they have much better short-term memory for previous words in the text. We advocate for creating models that have human-like long-term and short-term memory, and outline some possible directions for achieving this goal. Finally, we argue that currently available human data is insufficient to measure progress towards this goal, and outline human experiments that can address this gap.

Introduction

Human language comprehension is a rapid and efficient process. A key reason for this is the highly predictive nature of this process: we do not wait for the next word to start constructing the meaning of the sentence, but rather do so based on our prediction of upcoming words. As a consequence, readers experience a slowdown in reading and show heightened neural activity when they encounter unpredictable words (Ehrlich and Rayner, 1981; Kutas and Hillyard, 1984; Smith and Levy, 2013).

For a fully quantitative understanding of this process, we need to estimate how predictable a word in a particular context is to readers; for example, given the partial sentence *I purchased a* ______ , how much more predictable is *banana* compared to *cassowary*? Traditionally, predictability was estimated through the cloze task (Taylor, 1953), where subjects are asked to provide a completion given an incomplete sentence; words that are frequently produced by subjects in a particular context are deemed to be more predictable in that context. But this method is not a very practical way to estimate predictability: to estimate a difference in predictability between two low-predictability words, which are likely to be produced only rarely in a cloze experiment, enormous samples from many millions of participants are required to obtain reliable estimates.

An alternative to the cloze task relies on conditional probabilities derived from *language models*, computational systems that define probability distributions over sequences of words. Early studies used *n*-gram language models, which estimate the probability of a word in a context (typically one to four preceding words) based on the number of times it occurred in this context in a corpus. While predictability estimates derived in this way can reliably predict human word-by-word reading (McDonald and Shillcock, 2003; Boston et al., 2008; Fossum and Levy, 2012; Smith and Levy, 2013; Shain, 2019), *n*-gram language models are too impoverished to model human linguistic prediction: the context window they consider is far shorter than the context length to which humans are sensitive (Fitzsimmons and Drieghe, 2013; Brothers et al., 2020), and the ability of these models to generalize to context that were not seen in the training corpus is very limited.

Neural network language models (Jozefowicz et al., 2016; Gulordava et al., 2018; Radford et al., 2019) address these issues to a large extent, as the probabilities they produce are determined by longer sequences of words, and they can generalize better to new contexts. These models were quickly adopted by cognitive scientists, who showed that predictability estimates derived from these models are often superior to those from *n*-gram models (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Merkx and Frank, 2021).

A consistent finding from these early studies was that language models that predict the next word more accurately — i.e. place higher probability on the word that in fact occurred next in the sentence — also yield conditional probabilities that align more closely with human

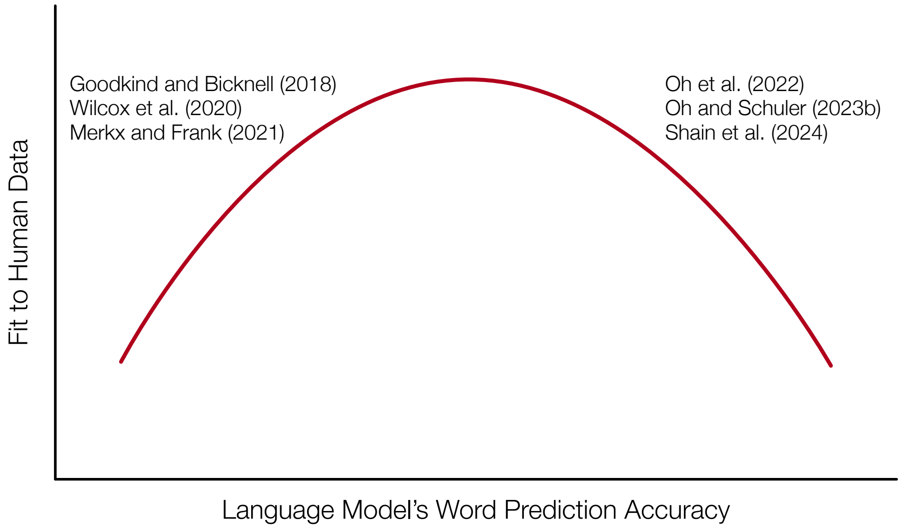
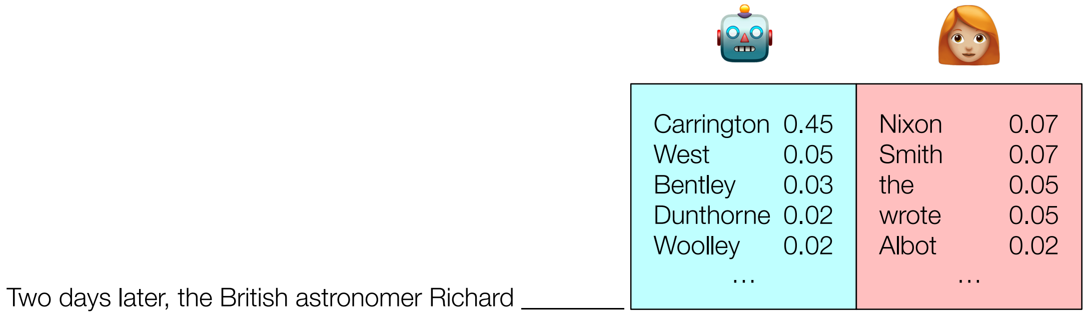


Figure 1: **The relationship between a language model's word prediction accuracy and fit to human data reversed.** Studies published in 2018-2021 report that models with higher prediction accuracy provide a better fit to human data. However, this relationship reversed in studies published after 2022, where more accurate language models provide a poorer fit.

reading times. Intriguingly, this relationship completely reversed in studies published more recently (from around 2022 on). These studies — which used more powerful language models based on the Transformer neural network architecture (Vaswani et al., 2017) that were trained on a much larger amount of text — showed that these models find the words they encounter to be more predictable than humans do, resulting in a divergence between the models' predictions and human reading times (Figure 1; Oh et al., 2022; Oh and Schuler, 2023b; Shain et al., 2024). This suggests that these newer models — which we will call "mainstream large language models (LLMs)" throughout this position paper — make predictions that average human readers readily cannot, making them 'superhuman' as models of linguistic prediction.

In this position paper, we review the extant studies and argue that two factors underpin this superhumanness of mainstream LLMs. The first is their long-term memory of massive amounts of training data: LLMs remember too much text compared to the average human reader. The second is their near-perfect short-term memory of previous words in the text: unlike humans, mainstream Transformer-based LLMs are not subject to limitations in working memory and can faithfully recall specific words that occurred many pages before the current word, if that is helpful to predict upcoming words. We argue that for language models to serve as cognitive models of linguistic prediction, efforts toward developing more human-like language models are necessary, and that further human experiments are required to support these preliminary hypotheses and benchmark research progress on this front.



Two days later, the British astronomer Richard ________

🤖		👩	
Carrington	0.45	Nixon	0.07
West	0.05	Smith	0.07
Bentley	0.03	the	0.05
Dunthorne	0.02	wrote	0.05
Woolley	0.02	Albot	0.02
…		…	

Figure 2: **Mainstream LLMs have much more factual knowledge than average humans.** Top five words with highest probabilities from the Llama 3 LLM (blue; Grattafiori et al., 2024) and proportions of top five frequent completions collected from 41 human subjects (red; Luke and Christianson, 2018). The LLM places high probabilities on names of astronomers, including the correct continuation *Carrington*.

Superhuman long-term memory

A core reason that LLMs exceed the ability of human readers to predict the next word lies in their propensity to remember examples from the training data much more faithfully than humans can. An illustrative example of this issue comes from text that contains factual knowledge about the world. Consider the partial sentence *Elvis Presley was born in the city of* _____. How predictable the upcoming word is for the reader will crucially depend on whether they already know this fact about the birthplace of Elvis; a reader who knows this fact is more likely to correctly predict *Tupelo* and experience less processing difficulty upon encountering it. Indeed, the fact that readers bring their background knowledge to bear on the comprehension process has been emphasized in theories of reading comprehension (Kintsch, 1998; van Dijk and Kintsch, 1983) and studied as a source of individual difference in language comprehension (Smith et al., 2021). But many readers quite plausibly will have never heard about the birthplace of Elvis Presley, or, if they have, may have completely forgotten it after having been exposed to it at some point (see Figure 2 for a related example).

Given the central role of background knowledge in next-word prediction, we expect a cognitive model of linguistic prediction to make predictions that are consistent with the reader's knowledge. But through their training, mainstream LLMs come to embody too much knowledge for modeling human readers. This can be attributed to three reasons. First, the data typically used to train mainstream LLMs is text from the internet, which contains a high proportion of material that conveys factual information, such as Wikipedia articles. Second, the amount of data used to train mainstream LLMs is often orders of magnitudes larger than what humans are exposed to (Hart and Risley, 1995; Wilcox et al., 2025). The typical English-speaking child will have experienced at most 100 million words by the age of 12, while mainstream LLMs like Llama 3 (Grattafiori et al., 2024) are trained on as many as 15 *trillion* words. Finally, and perhaps most fundamentally, mainstream LLMs do not easily forget the text that they encounter, unlike humans who naturally do so over time. Mainstream LLMs have been shown to store long sequences of words they have encountered such that it is possible to extract them verbatim by prompting the model (McCoy et al., 2023; Carlini et al., 2023; Merrill et al., 2024). To return to our running example, the properties of the training data make it more likely for mainstream LLMs to be exposed to the birthplace of Elvis Presley, and the learning behavior of these models makes it likely that this piece of knowledge will be retained. As a consequence, mainstream LLMs become much more likely than human readers to correctly predict *Tupelo* given the example above.

Recent studies provide preliminary correlational evidence for the conjecture that the mismatch between LLM and human linguistic prediction is in part due to the LLMs' superior long-term memory. A regression analysis of reading times showed that language models were much less surprised than human readers by proper names like *Tupelo*, which can only be predicted with accurate factual knowledge (Oh and Schuler, 2023b). This discrepancy was more severe for larger, more powerful models, suggesting that improving the next-word prediction accuracy of language models will not make them more human-like. Because proper nouns only constituted a small subset of the reading time datasets studied by Oh and Schuler (2023b), these results should at present be regarded as preliminary; this hypothesis should be tested in a more targeted way in future experiments.

Another, more indirect source of evidence for this hypothesis comes from the finding that language models trained on smaller amounts of data than typical mainstream LLMs yield probabilities that are better aligned to human reading times (Oh and Schuler, 2023a). Although limiting the amount of training data could affect language model probabilities in many different ways, one of them is by preventing the exposure of factual information like the birthplace of Elvis that could eventually be memorized.

We hypothesize that the long-term memory issue goes beyond factual knowledge-based prediction: whenever there is a mismatch in the familiarity with the text between humans and models, there will likely be a corresponding mismatch in processing. Another potential mismatch could come from multiword expressions like the idiom "spill the beans," the later words of which are very easily predicted given the first ones by LLMs (Rambelli et al., 2023); we hypothesize that mainstream LLMs, given their vast training data and memorization capabilities, are able to store a much larger range of multiword expressions from different dialects and registers of the language than humans can, and consequently can predict upcoming words in these expressions with greater accuracy. An extreme case of this issue is illustrated by famous documents like the Declaration of Independence, which are memorized and predicted accurately by mainstream LLMs (McCoy et al., 2023; Merrill et al., 2024), but are not likely to be memorized by most humans. Crucially, the amount of exposure required for mainstream LLMs to become familiar with these examples is much lower than that required for humans: in some cases, LLMs can memorize a training example after only a handful of exposures (Tirumala et al., 2022; Lesci et al., 2024). This may explain the greater misprediction of reading behavior on low-frequency words (Oh et al., 2024), some of which could be memorized as part of multiword expressions.

In summary, a crucial factor that limits the ability of LLMs to serve as models of prediction in human language comprehension is their excessive knowledge of text compared to the average human reader. This is due to the difference in the quantity and quality of language exposure and the ability to remember what was seen. As a consequence, mainstream LLMs embody massive amounts of information that enable them to predict upcoming words that are unpredictable to humans with limited knowledge.

Superhuman short-term memory

Whether it is spoken or written, linguistic input is transient during language processing. From this impermanent input, the goal of the human comprehender is to build a mental representation of its meaning (Bransford and Franks, 1971; Jarvella, 1971). During this process, humans are subject to working memory limitations and naturally forget parts of the input over time (Baddeley and Hitch, 1974; Baddeley, 2003; Lewis and Vasishth, 2005). The decay in the availability of earlier linguistic inputs influences readers' expectations about upcoming words, and should be taken into account by any model of language comprehension (Futrell et al., 2020).

In contrast to humans, mainstream LLMs are able to build much more robust representations of a large number of previous elements in the linguistic input and use these representations to make predictions. Part of this is attributable to the Transformer neural network architecture (Vaswani et al., 2017) that underlies most mainstream LLMs at the time of writing.

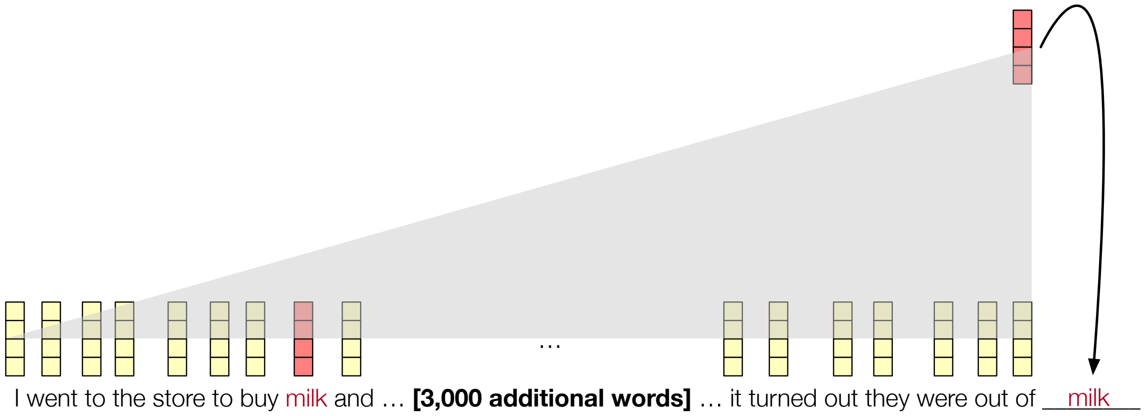

I went to the store to buy milk and … [3,000 additional words] … it turned out they were out of milk
…

Figure 3: **Transformers have perfect access to the input sequence.** Even with thousands of words in between, Transformers can retrieve the earlier occurrence of the word *milk* to make the correct prediction of the second *milk*.

At a high level, these networks build a vector representation of the input sequence by taking a weighted average of the representations of each of the previous words in the sequence (this is referred to as "attention"). During this process, the model has perfect access to a very large context window which can span tens of thousands or even millions of words — typically enough to fit the entire length of the text being processed — and the resulting representation of the sequence is not subject to any form of temporal decay that humans must face. Mainstream LLMs' ability to look at every word in the sequence and mix the representations of these words to perform next-word prediction is especially helpful for predicting discourse entities that are likely to be repeated throughout the text: the probability of a second mention of the same entity is much higher than of the first one (Church, 2000, see also Figure 3).

There is evidence that Transformer-based language models in fact use this property of their architecture to make predictions: with sufficient training, they become capable of assigning probabilities of near one to sequences of words that were observed earlier in the input (Armeni et al., 2022). This suggests that these models can learn to implement a form of copying mechanism that leverages their decay-free access to the input (Elhage et al., 2021; Olsson et al., 2022; Bietti et al., 2023; Jelassi et al., 2024). Indeed, direct comparisons between humans and LLMs during the processing of texts that contained repeated passages have shown that while mainstream LLMs assigned high probabilities of near one to the repeated portion of the text, humans were not as accurate at explicitly predicting the repeated words (Vaidya et al., 2023) and did not speed up as drastically when reading the repeated portion as might be expected if they had verbatim memory of the first occurrence of the text (Gruteke Klein et al., 2024).

Mainstream LLMs' perfect access to earlier words in the text input may underlie other discrepancies with human processing. This is the case, for example, when readers need to identify the subject for a particular verb in sentences that have multiple candidate subjects. In the sentence *The rat the cat the dog chased loved ate the malt*, for example, readers often struggle to determine that the subject of *ate* is *the cat* rather than any of the other noun phrases. LLM probabilities fail to accurately predict the increase in reading times displayed by participants at the main verb of such deeply embedded sentences (Hahn et al., 2022). This suggests that Transformers do not face interference from multiple candidate subjects, possibly because they can perform many parallel retrieval operations at once (Timkey and Linzen, 2023). Interference from multiple candidate subjects in human sentence processing also arises in the phenomenon of agreement attraction (Bock and Miller, 1991; Pearlmutter et al., 1999; Wagers et al., 2009), where participants, again because they interpret an incorrect noun as the subject of the sentence, consider ungrammatical sentences such as *The keys on the table is ...* to be acceptable. Transformer LLMs show a much lower rate of agreement attraction errors than people do, suggesting again that they display considerably less memory interference than humans (Arehalli and Linzen, 2024).

A third mismatch between LLM and human prediction that could be due to LLMs' superhuman short-term memory concerns syntactic disambiguation. When a sentence turns out to have a structure that differs from the one it appeared likely to have based on the first few words of the sentence — so called garden-path sentences, such as *The old man the boat* — humans often experience severe processing difficulty. Mainstream LLMs, by contrast, predict only very mild unpredictability in those contexts, quite possibly because they have the memory capacity to consider multiple possible structures of the sentence concurrently (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024).

Further support for the hypothesis that mainstream LLMs' short-term memory is too accurate to model human language comprehension comes from studies that manipulate the LLMs' access to the input. In deeply embedded sentences, downweighting the contribution of high-frequency words that are unlikely to be retained in memory reduces the mismatch between LLMs and humans in the prediction of the main verb (Hahn et al., 2022). More generally, the alignment between LLM probabilities and human reading times improves when the LLM's ability to access earlier words in the sentence is reduced by increasing the attention allocated to recent words (de Varda and Marelli, 2024; Clark et al., 2025) or providing only a few recent words as input (Kuribayashi et al., 2022). While questions about the exact form of human-like decay remain, these results converge to the conclusion that the decay-free representation in mainstream LLMs is problematic for modeling human linguistic prediction.

Towards LLMs that are better aligned with humans

In this section, we highlight how changes to language model training procedures and neural network architectures could address the issues outlined in the previous sections. The most obvious discrepancy to address is the mismatch in the quantity and quality of training data; we hypothesize that limiting training data would limit LLMs' opportunities to memorize knowledge that humans may not have, and could hinder the development of superhuman short-term memory mechanisms, which emerge in Transformers only after considerable training (Armeni et al., 2022).

A recent effort to spur research on language models with human-scale training data is the BabyLM Challenge (Warstadt et al., 2023; Wilcox et al., 2025). This shared task invited teams to train language models on curated datasets that consist of 100 million words or less, about 40% of which is transcribed child-directed speech or text material intended for children. Evaluation on grammatical minimal pair discrimination (Warstadt et al., 2020) demonstrated the potential of language models to learn robust linguistic generalizations from small amounts of data. While this kind of human-scale training seems promising, it remains to be seen whether it will address the superhuman long-term and short-term memory issues described above. In addition, a fundamental question remains about the correct mixture of training data that will give rise to the next-word predictions that are representative of those of the typical human reader. Identifying the right kind of training data is complicated by the fact that humans learn through other sensory experience in addition to linguistic input; for example, humans may learn the fact that bananas are typically yellow from having seen bananas, rather than from reading text that describes the color of bananas.

Even with the ideal training data, the misalignment will persist if models draw different generalizations from it than humans. The typical training objective of LLMs only rewards predicting the correct word (the one that in fact occurred in the text), and therefore mainstream LLMs are incentivized to make lexically sharp predictions that assign a high probability to the correct continuation whenever possible. For example, given *Elvis Presley was born in the city of* ______, they are likely to assign a high probability to *Tupelo* and low probabilities to names of other cities. Given mainstream LLMs' ability to store training examples, these lexically sharp predictions are unlikely to change once they are learned. In contrast, humans make broader predictions based on meaning, which facilitate the processing of words that share semantic properties with the target word, even if the target word itself is unpredictable (Federmeier and Kutas, 1999; Roland et al., 2012; Luke and Christianson, 2016). Therefore, a reader who does not know the birthplace of Elvis is likely to find the name of a plausible yet incorrect city to be somewhat predictable. For LLMs to make such diffuse next-word predictions like humans, alternative training objectives that upweight semantically similar words together could be helpful. Such a training objective could minimize a continuous loss based on the distance between vector representations (Kumar and Tsvetkov, 2019), upweighting the prediction of not only *Tupelo* but also other cities, to the extent that their representations are similar.

The misalignment in prediction due to mainstream LLMs' long-term memory can also be addressed by intervening on their representations — first locating some information of interest, and then adjusting it to steer the model's predictions. Model editing techniques (Meng et al., 2022; Wang et al., 2023), in particular, typically train the model on a curated set of examples so as to lead it to make predictions that are more factually correct than it would otherwise. For instance, if a model mispredicts the birthplace of Elvis Presley as *Orlando*, one could further train on sentences like *Elvis Presley was born in Tupelo* to get it to predict the correct birthplace of Elvis. For purposes of cognitive modeling, these techniques can be applied in the opposite direction to unlearn knowledge that most human readers wouldn't have, and bring it closer to human-like knowledge that influences real-time processing.

What about LLMs' superhuman short-term memory? Existing approaches for reducing the accessibility of the representations of earlier words in Transformer LLMs make simplifying assumptions about human memory, in particular, that the input decays solely as a function of time, such that earlier words in the text always decay more than later words (Kuribayashi et al., 2022; de Varda and Marelli, 2024; Clark et al., 2025) or that it decays as a function of frequency, such that high-frequency words always decay more than low-frequency words (Hahn et al., 2022). As a starting point, experiments should evaluate the impact of more flexible methods for limiting the accuracy of memory retrieval of linguistic input in Transformers. Those could include implementing decay at the level of syntactic constituents rather than individual tokens (Lewis and Vasishth, 2005), or implementing memory retrieval bottlenecks that could give rise to the interference humans experience between similar units in short-term memory (Timkey and Linzen, 2023).

Alongside these approaches, which aim to restrict the scope of Transformers' attention mechanisms, we recommend pursuing approaches that abandon Transformers altogether, and place a renewed focus on neural network architectures that rely on recurrence (e.g. Elman, 1991; Hochreiter and Schmidhuber, 1997; Dao and Gu, 2024). In comparison to Transformers, which have direct access to the representations of prior words, recurrent models are forced to compress the input into a representation of a fixed size, and as such are more likely to implement a human-like form of decay when making predictions.

Targeted human experiments for benchmarking modeling progress

We have outlined the hypothesis that the discrepancy between the linguistic predictions of mainstream LLMs and humans is rooted in the models' superior long-term and short-term memory. Existing evidence for this hypothesis is observational and preliminary, as many of the reading datasets that have been studied focus on naturalistic text (e.g. newspaper articles), and as such do not manipulate short- and long-term memory in a controlled way. To better ground our hypothesis, and to benchmark efforts to narrow the gap between LLMs and humans, we argue that targeted human experiments are necessary that can measure the role of long-term and short-term memory in human reading.

For the first issue surrounding long-term memory, a more direct link between the readers' knowledge of text and their reading behavior needs to be established. For the example of factual knowledge, this could be achieved by collecting reading times of sentences from datasets that contain information about real-world entities (e.g. Levy et al., 2017), coupled with a pre-experiment questionnaire that gauges the relevant piece of knowledge beforehand. The same experimental paradigm of explicitly measuring humans' familiarity with some text and linking it to their reading behavior can likewise apply to multiword expressions and famous documents like the Declaration of Independence.

For the second issue of short-term memory, the influence of temporal decay in reading — which needs to be evaluated at the discourse level — has been far less studied than the influence of complex subject-verb relations within the same sentence. As such, experiments using text that requires readers to recall crucial information (such as a short list of items) with varying amounts of text that are interleaved should be conducted. Experiments like these will complement existing naturalistic reading time datasets and provide more direct evidence for the misalignment between the memory capabilities of humans and LLMs that influence real-time language comprehension. The data collected as part of these experiments can additionally be used to benchmark research progress in developing language models that are more human-like through approaches like those outlined in the previous section.

Conclusion

This position paper reviews recent studies that use neural language models as models of linguistic prediction during language comprehension. While predictability estimates from early neural language models showed promise in predicting human behavioral measurements of linguistic prediction, in the last few years mainstream LLMs have become much more accurate than humans at next-word prediction, leading to an increasing misalignment in prediction between LLMs and humans. We argue that this growing superhumanness of LLMs is due to their superior long-term memory of training examples and short-term memory of previous words in the text. To address this issue, we have advocated for creating language models with human-like long-term and short-term memory through the use of alternative training procedures and neural network architectures, and argued that new human experiments should be conducted to benchmark progress on this front.

Acknowledgments

This project is supported by the National Science Foundation (NSF) under grants BCS-2020945 and IIS-2504953 and the National Institute of Biomedical Imaging and Bioengineering under grant R01EB038873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation or the National Institutes of Health.

References

Arehalli, S., Dillon, B., and Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 301–313.

Arehalli, S. and Linzen, T. (2024). Neural networks as cognitive models of the processing of syntactic constraints. *Open Mind*, 8:558–614.

Armeni, K., Honey, C., and Linzen, T. (2022). Characterizing verbatim short-term memory in neural language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 405–424.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36(3):189–208.

Baddeley, A. D. and Hitch, G. (1974). *Working Memory*. University of Stirling, Stirling, Scotland.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. (2023). Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pages 1560–1588.

Bock, K. and Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Bransford, J. D. and Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, 2:331–350.

Brothers, T., Wlotko, E. W., Warnke, L., and Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, 1(1):135–160.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. (2023). Quantifying memorization across neural language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Church, K. W. (2000). Empirical estimates of adaptation: The chance of two noriegas is closer to $p/2$ than p^2. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Clark, C., Oh, B.-D., and Schuler, W. (2025). Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747.

Dao, T. and Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 10041–10071.

de Varda, A. G. and Marelli, M. (2024). Locally biased transformers better align with human reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for Transformer circuits. *Transformer Circuits Thread*.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Federmeier, K. D. and Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Fitzsimmons, G. and Drieghe, D. (2013). How fast can predictability influence word skipping during reading? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4):1054–1063.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–69.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3).

Goodkind, A. and Bicknell, K. (2018). 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*, pages 10–18.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim,

S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The Llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783v2.

Gruteke Klein, K., Meiri, Y., Shubi, O., and Berzak, Y. (2024). The effect of surprisal on reading times in information seeking and repeated reading. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1195–1205.

Hahn, M., Futrell, R., Gibson, E., and Levy, R. P. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Brookes Publishing, Baltimore, MD.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., and Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10:409–416.

Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. (2024). Repeat after me: Transformers are better than state space models at copying. *arXiv preprint*, arXiv:2402.01032v2.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint*, arXiv:1602.02410v2.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge.

Kumar, S. and Tsvetkov, Y. (2019). Von mises-fisher loss for training sequence to sequence models with continuous outputs. In *7th International Conference on Learning Representations (ICLR)*.

Kuribayashi, T., Oseki, Y., Brassard, A., and Inui, K. (2022). Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Lesci, P., Meister, C., Hofmann, T., Vlachos, A., and Pimentel, T. (2024). Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635.

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. (2017). Zero-shot relation extraction via reading comprehension. In Levy, R. and Specia, L., editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Luke, S. G. and Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

Luke, S. G. and Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., and Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

McDonald, S. A. and Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35*.

Merkx, D. and Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Merrill, W., Smith, N. A., and Elazar, Y. (2024). Evaluating *n*-gram novelty of language models using Rusty-DAWG. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14459–14473.

Oh, B.-D., Clark, C., and Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Oh, B.-D. and Schuler, W. (2023a). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Oh, B.-D. and Schuler, W. (2023b). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Oh, B.-D., Yue, S., and Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2644–2663.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*.

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement Processes in Sentence Comprehension. *Journal of Memory and Language*, 41(3):427–456.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Rambelli, G., Chersoni, E., Senaldi, M. S. G., Blache, P., and Lenci, A. (2023). Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 87–98.

Roland, D., Yun, H., Koenig, J.-P., and Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122:267–279.

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4086–4094.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Smith, R., Snow, P., Serry, T., and Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42(3):214–240.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Timkey, W. and Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720.

Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38274–38290.

Vaidya, A., Turek, J., and Huth, A. (2023). Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 58–69.

van Dijk, T. A. and Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press, New York, NY.

van Schijndel, M. and Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45:e12988.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., and Li, J. (2023). Knowledge editing for large language models: A survey. *arXiv preprint*, arXiv:2310.16218.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors (2023). *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., and Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.