


Opinion

To model human linguistic prediction, make LLMs less superhuman

Byung-Doh Oh ^{1,*}, and Tal Linzen^{2,3}

When we read, we make predictions about upcoming words; these predictions influence our reading behavior. The success of large language models (LLMs), which, like humans, make predictions about upcoming words, has motivated their use as models of human linguistic prediction. Surprisingly, in the last few years, as LLMs' ability to predict the next word has improved, their ability to explain reading behavior has declined. We argue this is because current LLMs can predict upcoming words much better than human readers can. This 'super-humanness' is driven by LLMs' extensive training data, stronger long-term memory of training examples, and stronger short-term memory. We advocate for LLMs with human-like memory and for new experiments to measure the alignment between humans and LLMs, and we outline directions toward achieving these goals.

LLMs as potential models of prediction in human language comprehension

Prediction is a core principle of human cognition, where it plays a central role in domains ranging from visual object recognition to action planning [1,2]. Language comprehension, in particular, is arguably as rapid and efficient due to its predictive nature: when we encounter predictable words, we are able to read them faster, and our brains display reduced neural activity in response to those words [3–5].

Studying how linguistic predictions are generated and how they shape language comprehension can help shed light on human cognitive processes in language and beyond. Word predictability estimates from large language models (LLMs) [6–8]—neural networks that, like humans, make predictions about upcoming words—could be immensely useful for this project [9,10]. However, for LLMs to serve this purpose, their word predictions need to align with those made by humans. In the last few years, it has become clear that the predictions made by mainstream LLMs are, in fact, increasingly diverging from those of humans: LLMs are much better able to predict the next word than humans are; in other words, their predictions are superhuman [9,11,12]. In many AI applications, superhumanness is benign and even desirable; that is not, of course, the case for cognitive modeling.

This opinion article highlights two sets of issues that contribute to the growing discrepancy between humans and LLMs. The first is LLMs' long-term memory: they can consume and retain massive amounts of data, and consequently, they know dramatically more than human readers. The second is their near-perfect short-term memory of previous words in the text: unlike humans, LLMs can faithfully recall specific words that occurred many pages before the current word. After discussing these issues, we outline modeling efforts toward developing more human-like LLMs and argue that, because the empirical data that measures this discrepancy is currently very partial, new human experiments will be necessary to benchmark research progress on this front.

Highlights

Prediction is a core principle of human cognition and language comprehension, but the details of the process and the range of empirical findings it can account for are not well understood.

Large language models could provide infrastructure for studying prediction because they yield estimates of the probability of each upcoming word.

But mainstream large language models have developed in a direction that is increasingly unhelpful for the study of human linguistic prediction; the accuracy of their predictions now far exceeds the accuracy of human predictions.

Large language models' superhumanness can be attributed to larger training data, stronger long-term memory for training examples, and stronger short-term memory.

We advocate for addressing these issues through alternative architectures and training procedures, together with new human experiments to benchmark large language model–human alignment.

¹Division of Linguistics and Multilingual Studies, Nanyang Technological University, Singapore, Singapore

²Department of Linguistics, New York University, New York, NY, USA

³Center for Data Science, New York University, New York, NY, USA

*Correspondence: byungdoh.oh@ntu.edu.sg (B.-D. Oh).

How LLMs are used to study human language comprehension

While it is fairly clear that prediction plays a role in language comprehension [13,14], there are many open questions about this process. What kinds of information about the first few words of a sentence do we draw on to predict upcoming words, and what partial representations of the sentence do we construct? Which aspects of the behavioral processes involved in language comprehension, such as eye movements during reading, can be explained by prediction, and which aspects need to be explained by other principles [15]?

Any quantitative study of this process relies on estimates of how predictable a word in a particular context is to readers (Table 1). For example, given the partial sentence 'I purchased a _____', how much more predictable is 'banana' compared to 'cassowary'? Traditionally, predictability was estimated using responses to the cloze task [16], where subjects are asked to provide a completion (usually the next word) given an incomplete sentence; words that are more frequently produced by subjects in a particular context are deemed to be more predictable in that context. However, this method is not very practical: to reliably 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.

An alternative to the cloze task relies on conditional probabilities derived from *n*-gram language models [17,18], which estimate the probability of a word in the context of a handful of preceding words (typically one to four), based on the number of times it occurred in this context in a training corpus. While predictability estimates derived in this way have substantial predictive value for human word-by-word reading [5,19–22], *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 when making predictions [23,24], and the ability of these models to generalize to new contexts that were not seen in the training corpus is very limited.

LLMs [6–8], which predict upcoming words using neural networks, address both of these issues to a large extent, as the probabilities they produce are determined by longer sequences of words, and their representations can generalize better to new contexts. These models were quickly adopted by cognitive scientists, who showed that predictability estimates derived from these models are superior to those from *n*-gram models [25–27]: they explained more variance in word-by-word reading times, measured as the time taken between keystrokes in an experiment where each keystroke reveals the next word (self-paced reading) [28], or as the duration of eye fixations on each word collected through eye tracking [29]. A consistent finding from these earlier studies was that language models that predict the next word more accurately—that is, place higher probability on the word that in fact occurred next in the sentence—also yielded conditional probabilities that align more closely with human reading times.

Table 1. Methods for estimating word predictability and their limitations

Method	Definition	Limitations
Cloze task [16]	Human subjects are asked to provide a completion to an incomplete sentence such as 'I purchased a _____'	Many responses are required to differentiate between low-predictability words, raising concerns about feasibility
<i>n</i> -gram language models [17,18]	Conditional probabilities are calculated based on sequence counts from a corpus, for example, $\frac{c(\text{I purchased a banana})}{c(\text{I purchased a})}$	Some sequences may not be attested in the corpus, and the context window is too short to model human linguistic prediction
LLMs [6–8]	Conditional probabilities are calculated from a large neural network that is typically trained to predict the next word	The probabilities may be based on long-term and short-term memory capacities that far exceed those of humans

Intriguingly, this relationship completely reversed in studies published more recently (from around 2022 onward). These studies—which used more powerful language models, typically based on the Transformer neural network architecture [30] and trained on a much larger amount of data—showed that these models find the words they encounter to be more predictable than humans do, resulting in a growing divergence between the models’ predictions and human reading times (Figure 1) [9,11,12]. This suggests that these newer models—which we will call ‘mainstream LLMs’ throughout this article—make predictions that average human readers readily cannot, making them superhuman as models of linguistic prediction. We argue that this superhumanness is best characterized by two underlying issues: LLMs’ long-term memory of massive amounts of training data and their near-perfect short-term memory of previous words in the text.

Superhuman long-term memory

A core reason that LLMs exceed the ability of human readers to predict the next word is their ability to remember examples from the training data much more faithfully than humans can. This is particularly clear in cases where successful prediction relies on factual knowledge. The fact that readers bring their background knowledge to bear on the comprehension process

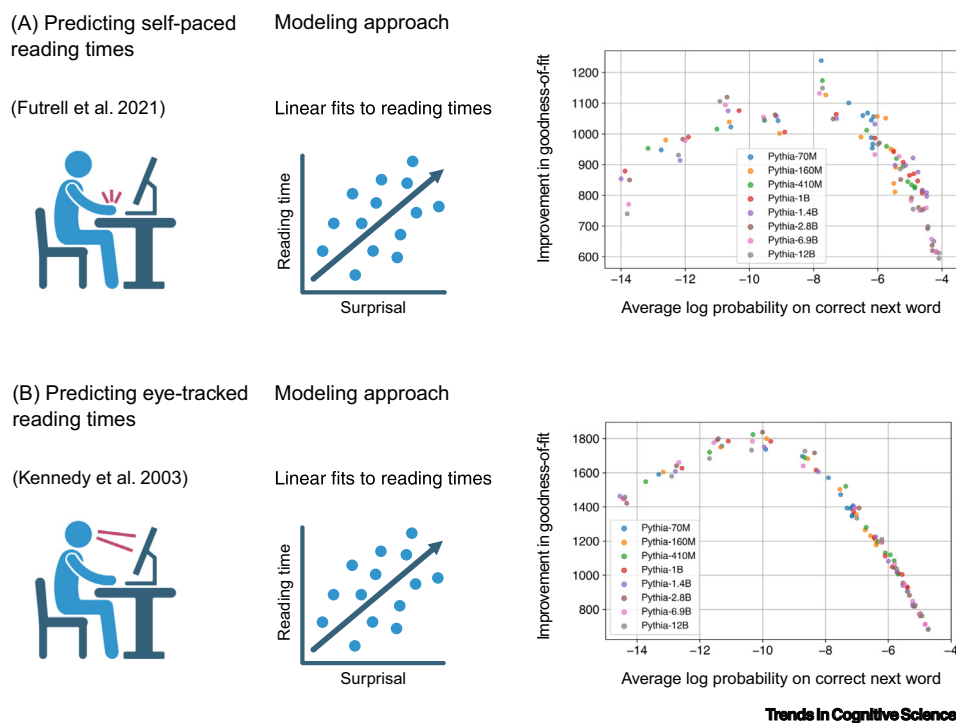
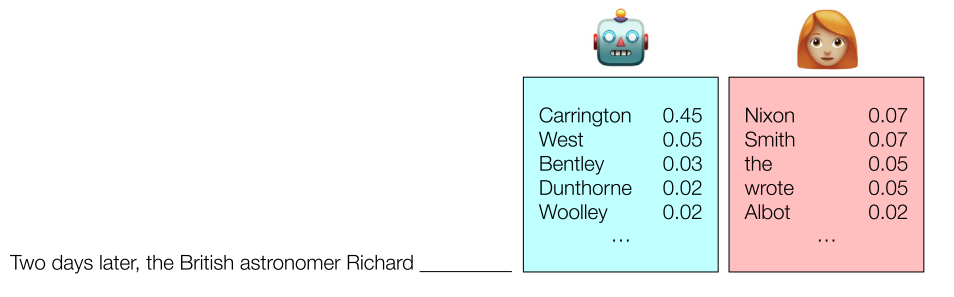


Figure 1. Relationship between an LLM’s word prediction accuracy and fit to human data. (A) Using the Natural Stories Corpus [31], which contains self-paced reading times of expository articles and short stories, we reproduce the analysis of [32] by calculating word-by-word predictability estimates from a wide range of LLMs [33] and evaluating their linear fit to reading times as the improvement in regression model log-likelihood over a common set of baseline predictors. Each dot represents one LLM; dots with the same color are LLMs with the same number of parameters, but different amounts of training data. The improvement in word prediction accuracy initially improves an LLM’s alignment with human data, which is consistent with earlier studies [25–27]. In contrast, after a certain point, further improvement in word prediction accuracy drives a misalignment with human data, which is consistent with later studies that used the same dataset [9,11,12]. (B) The same pattern of results is observed with data from the Dundee Corpus [34], which contains eye-tracked reading times of newspaper editorials. LLM: large language model.

has been emphasized in theories of reading comprehension [35,36] and studied as a source of individual differences in language comprehension [37]. Consider, for example, 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. If they do, they will be more likely to correctly predict ‘Tupelo’ and experience less difficulty upon encountering this word. But many readers, plausibly, will have never been exposed to information about the birthplace of Elvis Presley or, if they have been exposed to it at some point, may have completely forgotten it (see Figure 2 for a related example).

Given the central role of background knowledge in next-word prediction, an accurate cognitive model of linguistic prediction would need to make predictions that are consistent with the reader’s knowledge. However, mainstream LLMs acquire far more knowledge than humans could ever acquire. This discrepancy can be attributed to three reasons. First, the data typically used to train mainstream LLMs contain a substantial amount of material that conveys factual information, such as Wikipedia articles. Second, the amount of data used to train mainstream LLMs is often orders of magnitude larger than what humans are exposed to [40,41]: a typical English-speaking child will have experienced at most 100 million words by the age of 12, while mainstream LLMs such as Llama 3 [38] are trained on as many as 15 trillion words. Finally, and

(A) Superhuman long-term memory



(B) Superhuman short-term memory

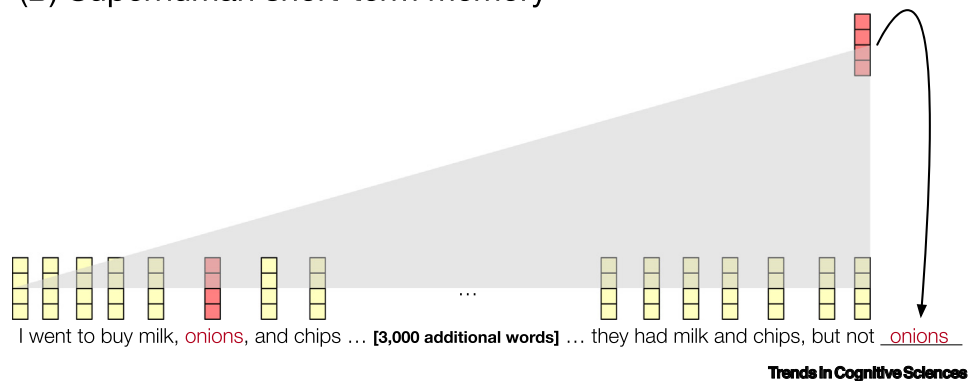


Figure 2. The memory capacity of LLMs far exceed those of humans. (A) Mainstream LLMs have much more factual knowledge than average humans due to their superior long-term memory. Top five words with highest probabilities from the Llama 3 LLM (blue) [38] and proportions of top five frequent completions collected from 41 human subjects (red) [39]. The LLM places high probabilities on names of astronomers, including the correct continuation ‘Carrington’. (B) Transformers have perfect store of the input sequence. Even with thousands of words in between, Transformers can retrieve the earlier occurrence of the word ‘onions’ to make the correct prediction of the second ‘onions’. LLM: large language model.

perhaps most fundamentally, mainstream LLMs do not easily forget the text 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 [42–44]. 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’, in the aforementioned example.

Supporting this conjecture, recent studies provide preliminary correlational evidence that the mismatch between LLMs and human linguistic prediction is, in part, due to the LLMs’ superior long-term memory. A regression analysis of reading times showed that LLMs were much less surprised than human readers by proper names like ‘Tupelo’, which can only be predicted with accurate factual knowledge [12]. The discrepancy with humans 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—if anything, the opposite is true.

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 with human reading times [32]. Although limiting the amount of training data could affect language model probabilities in many different ways [45,46], one of these ways is by preventing the exposure to factual information, such as the birthplace of Elvis, that could eventually be memorized. Further support for this hypothesis comes from the finding that, while training language models on some amount of text from specialized domains such as biology and physics textbooks improves their ability to model the reading behavior of subjects with relevant domain expertise, training on too much data from those domains has a deleterious effect, suggesting again that the models acquire superhuman factual expertise [47].

We hypothesize that the mismatch in familiarity with the text between humans and LLMs goes beyond factual, knowledge-based prediction. Consider, for example, multiword expressions like the idiom ‘spill the beans’, where the later words are easy to predict given the first ones [48]. Mainstream LLMs are likely able to store a much larger range of such expressions from different dialects and registers of the language than humans can, thanks to their vast training data and strong memorization capabilities. Consequently, they are likely to predict upcoming words in these expressions with greater accuracy than humans. An extreme case of this issue could be illustrated by famous documents like the Declaration of Independence, which are likely to be memorized and predicted accurately by mainstream LLMs [42,44] 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 [49,50].

In summary, a crucial factor that limits the ability of LLMs to serve as models of linguistic prediction is their greater familiarity with facts and turns of phrase compared to the average human reader. This is due to the difference in the quantity and quality of language exposure, as well as 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 [51,52]. During this process, humans are subject to limitations in working memory and naturally forget parts of the input over time [53–55]. The decay in the availability of earlier linguistic input influences readers' expectations about upcoming words [56,57].

In contrast to humans, mainstream LLMs are able to build and maintain robust representations of a much larger number of previous words and use these representations to make next-word predictions. Part of this is attributable to the Transformer neural network architecture [30] that underlies most mainstream LLMs at the time of writing. These networks build vector representations of each word by taking a weighted average of the representations of each previous word in the sequence (this is referred to as 'attention'). During this process, the model has access to its representations of the words in 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. The resulting representation of the sequence is not subject to any form of temporal decay, unlike human memory.

Mainstream LLMs' substantially stronger short-term memory could allow them to make more accurate predictions. Names and concepts are much more likely to be mentioned again once they have already been mentioned [58]. An LLM could easily predict such repeated entities, even in cases where they are difficult for humans to remember—in a 19th-century Russian novel with an extensive cast of characters, for example (Figure 2).

There is evidence that Transformer LLMs, in fact, utilize this property of their architecture to make predictions: with sufficient training, they can assign probabilities near one to sequences of words that they observed earlier [59]. This suggests that these models can learn to implement a form of copying mechanism that leverages their decay-free store of the input [60–62]. 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 near one to the repeated portion of the text, humans were not able to predict the repeated words as accurately [63], and when reading the repeated portion, they did not speed up as drastically as might be expected if they had verbatim memory of the first occurrence of the text [64].

Mainstream LLMs' perfect storage of earlier words in the text input may lead to other discrepancies with human language processing. One such case could arise when readers form syntactic dependencies between words, for example, when they need to identify the subject for a particular verb from 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 rat' rather than any of the other noun phrases [65–67]. LLM probabilities fail to accurately predict the increase in reading times displayed by participants at the main verb of such deeply embedded sentences [57]. This suggests that Transformers face little interference from multiple candidate subjects, possibly because they can perform many parallel memory retrieval operations at once [68]. Interference from multiple candidate subjects in language processing also arises in the phenomenon of agreement attraction [69–71], 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 rusty' 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 [72].

A third mismatch between LLM and human prediction, which could be due to 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 [73,74], such as 'The old man the boat'—humans often

experience severe processing difficulty [74–76]. Mainstream LLMs, by contrast, predict only very mild difficulty in those contexts, possibly because they have the memory capacity to consider multiple possible structures of the sentence concurrently [10,15,77].

Further support for the hypothesis that mainstream LLMs' short-term memory is too strong to model human linguistic prediction comes from studies that manipulate or limit LLMs' access to the representations of previous words. 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 [57]. 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 [78,79, but cf. 80] or by providing only a few recent words as input [81]. Finally, limiting access to earlier words also yields attention patterns in Transformer LLMs that are better suited to capture processing difficulty during reading [82]. Overall, while questions about the exact form of human-like decay remain, these results converge on the conclusion that mainstream LLMs' access to decay-free representations negatively impacts their ability to model human linguistic prediction.

Toward LLMs that are better aligned with humans

How could changes to language model architectures and training procedures address the issues we have identified? The most obvious discrepancy that needs to be addressed is the mismatch in the quantity and quality of training data. Limiting training data would not only limit LLMs' opportunities to memorize knowledge that humans may not have but also hinder the development of superhuman short-term memory mechanisms, which emerge in Transformers only after considerable training [59]. Such efforts could follow in the footsteps of the BabyLM Challenge [41,83], a yearly shared task that invites teams to train language models on curated datasets consisting of 100 million words or less, about 40% of which is transcribed child-directed speech or text material intended for children. This shared task has shown that language models can learn robust linguistic generalizations from human-scale amounts of data, pointing to the feasibility of this paradigm. However, it is as yet unknown if training on this corpus addresses the superhuman long-term and short-term memory issues described earlier.

A fundamental question remains about the correct mixture of training data that will give rise to the next-word predictions 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 experiences in addition to linguistic input [84]; for example, most humans might learn the fact that bananas are typically yellow from having seen bananas rather than from reading text that describes the color of bananas. This raises the possibility that fully human-like word predictions would require training on multimodal data or additional textual data that would compensate for the lack of multimodal training.

Even with the ideal training data, the misalignment will persist if models draw different generalizations from the data than humans. The typical training objective of LLMs only rewards predicting the correct word—the one that, in fact, occurred in the text—and, consequently, mainstream LLMs are incentivized to make lexically sharper 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 lower 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. These predictions facilitate the processing of words

that share semantic properties with the target word, even if the target word itself is unpredictable [85–87]. 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. While we expect that, in such cases, LLMs also assign some probability mass to words other than the most likely word, we hypothesize that they do so to a lesser extent than people.

To incentivize LLMs to make next-word predictions that are as diffuse as those of 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 [88], which would upweight the prediction of not only ‘Tupelo’ but also other cities, to the extent that their representations are similar. This can be coupled with a reduction in the size of vector representations for each word, which would further prevent the model from encoding specific information about each city. Another way to make LLMs’ predictions more diffuse could involve temperature-scaling prior to probability calculation to increase the uncertainty of LLMs’ predictions [89]; this intervention has been shown to improve the fit of LLM probabilities to human data [90].

The misalignment in predictions, which is due to mainstream LLMs’ long-term memory, can also be addressed *post hoc* by intervening in their representations—first locating a piece of information of interest and then adjusting it to steer the model’s predictions. Model editing techniques [91–93], 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 it 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 to cause it to unlearn knowledge that most human readers would not have.

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 [78,79,81], or as a function of frequency, such that high-frequency words always decay more than low-frequency words

Box 1. 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-term and long-term memory in a controlled way. To better ground this hypothesis, and to benchmark efforts to close the gap between LLMs and humans, new, targeted human experiments are necessary.

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., 97], coupled with a questionnaire that gauges familiarity with the relevant piece of knowledge. A similar experimental paradigm, which measures humans’ familiarity with some text and links 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, we recommend collecting measurements during the reading of texts that implicitly require readers to recall information, such as a short list of items or the name of a particular character in the story. These experiments should vary factors that are known to affect human memory, such as the size of the list, the amount of intervening text between mentions, and the position of the entity in the first presentation of the list [98].

[57]. As a starting point, experiments should evaluate the impact of more flexible methods for limiting Transformers' memory retrieval accuracy. These could include decay at the level of syntactic constituents rather than individual tokens [55], or memory retrieval bottlenecks that could give rise to the interference humans experience between similar units in short-term memory [68].

Alongside these approaches, which aim to restrict the scope of Transformers' attention mechanisms, we recommend exploring approaches that abandon Transformers altogether and place a renewed focus on recurrent neural network architectures [e.g., 94–96]. 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, therefore, are more likely to implement a human-like form of decay when making predictions.

These efforts to narrow the gap between LLMs and humans should be accompanied by targeted human experiments that measure the role of long-term and short-term memory in human reading (Box 1). Such experiments will help benchmark research progress on this front and complement existing datasets by providing a more detailed quantitative characterization of the misalignment between the memory capabilities of humans and LLMs as it affects linguistic prediction.

Concluding remarks

Prediction is a core principle of human cognition and language processing: we use our linguistic experience to make predictions about how a sentence will unfold. While language models could help elucidate this process, in the last few years, mainstream LLMs have become much more accurate than humans at next-word prediction, limiting their applicability for cognitive modeling. This discrepancy between humans and LLMs partly parallels findings in the vision domain, where deep neural networks use prediction strategies that exceed humans' capabilities [99,100]. We highlighted two areas where LLMs are superhuman in this sense: their superior long-term memory of training examples and their superior short-term memory of previous words in the text. We call for research efforts to address these issues (see also Outstanding Questions): we should develop language models with human-like memory through alternative training procedures and neural network architectures, and conduct new human experiments to benchmark research progress.

Acknowledgments

This project is supported by the National Science Foundation under grants BCS-2020945 and IIS-2504953 and by 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.

Declaration of interests

Tal Linzen is a part-time employee at Google.

References

1. Bar, M. (2009) The proactive brain: memory for predictions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1235–1243
2. Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204
3. Ehrlich, S.F. and Rayner, K. (1981) Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* 20, 641–655
4. Kutas, M. and Hillyard, S.A. (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163
5. Smith, N.J. and Levy, R. (2013) The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319
6. Jozefowicz, R. et al. (2016) Exploring the limits of language modeling. *arXiv* <https://doi.org/10.48550/arXiv.1602.02410>
7. Gulordava, K. et al. (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* (Walker, M. et al., eds), pp. 1195–1205, Association for Computational Linguistics
8. Radford, A. et al. (2019) *Language Models are Unsupervised Multitask Learners*, OpenAI Technical Report

Outstanding questions

What kinds of training data and training objectives are most likely to give rise to the next-word predictions that average human readers make?

How does extralinguistic experience—to which many large language models typically do not have exposure—shape humans' linguistic predictions during reading?

Compared to large language models, how much linguistic and nonlinguistic knowledge do average human readers lack? Conversely, is there knowledge that humans have but large language models do not, which also introduces a mismatch in predictions?

What are the most effective machine learning techniques for erasing the knowledge encoded by large language models without drastically distorting their representations and predictions?

What kinds of linguistic information are retained in human short-term memory, and how can neural network architectures be modified to match this mechanism?

9. Shain, C. *et al.* (2024) Large-scale evidence for logarithmic effects of word predictability on reading time. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2307876121
10. Huang, K.-J. *et al.* (2024) Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *J. Mem. Lang.* 137, 104510
11. Oh, B.-D. *et al.* (2022) Comparison of structural parsers and neural language models as surprisal estimators. *Front. Artif. Intell.* 5, 777963
12. Oh, B.-D. and Schuler, W. (2023) Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Trans. Assoc. Comput. Linguist.* 11, 336–350
13. Kuperberg, G.R. and Jaeger, T.F. (2016) What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59
14. Staub, A. (2025) Predictability in language comprehension: prospects and problems for surprisal. *Annu. Rev. Linguist.* 11, 17–34
15. van Schijndel, M. and Linzen, T. (2021) Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci.* 45, e12988
16. Taylor, W.L. (1953) "Cloze procedure": a new tool for measuring readability. *J. Mass Commun. Q.* 30, 415–433
17. Jelinek, F. *et al.* (1975) Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Inf. Theory* 21, 250–256
18. Baker, J. (1975) The DRAGON system—an overview. *IEEE Trans. Acoust. Speech Signal Process.* 23, 24–29
19. McDonald, S.A. and Shillcock, R.C. (2003) Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vis. Res.* 43, 1735–1751
20. Boston, M.F. *et al.* (2008) Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus. *J. Eye Mov. Res.* 2, 1–12
21. 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* (Reitter, D. and Levy, R., eds), pp. 61–69, Association for Computational Linguistics
22. 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* (Burstein, J. *et al.*, eds), pp. 4086–4094
23. Fitzsimmons, G. and Drieghe, D. (2013) How fast can predictability influence word skipping during reading? *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1054–1063
24. Brothers, T. *et al.* (2020) Going the extra mile: effects of discourse context on two late positivities during language comprehension. *Neurobiol. Lang. (Camb.)* 1, 135–160
25. 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* (Sayeed, A. *et al.*, eds), pp. 10–18, Association for Computational Linguistics
26. Wilcox, E.G. *et al.* (2020) On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Denison, S. *et al.*, eds), pp. 1707–1713, Cognitive Science Society
27. Merx, D. and Frank, S.L. (2021) Human sentence processing: recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Chersoni, E. *et al.*, eds), pp. 12–22, Association for Computational Linguistics
28. Just, M.A. *et al.* (1982) Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238
29. Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422
30. Vaswani, A. *et al.* (2017) Attention is all you need. In *Advances in Neural Information Processing Systems* (Guyon, I. *et al.*, eds), pp. 6000–6010, Curran Associates, Inc
31. Futrell, R. *et al.* (2021) The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Lang. Resour. Eval.* 55, 63–77
32. Oh, B.-D. and Schuler, W. (2023) 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* (Bouamor, H. *et al.*, eds), pp. 1915–1921, Association for Computational Linguistics (ACL)
33. Biderman, S. *et al.* (2023) Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning* (Krause, A. *et al.*, eds), pp. 2397–2430, PMLR
34. Kennedy, A. *et al.* (2003) The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement* (van Gompel, R. *et al.*, eds)
35. Kintsch, W. (1998) *Comprehension: A Paradigm for Cognition*, Cambridge University Press
36. van Dijk, T.A. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*, Academic Press
37. Smith, R. *et al.* (2021) The role of background knowledge in reading comprehension: a critical review. *Read. Psychol.* 42, 214–240
38. Grattafiori, A. *et al.* (2024) The Llama 3 herd of models. *arXiv* <https://doi.org/10.48550/arXiv.2407.21783>
39. Luke, S.G. and Christianson, K. (2018) The Provo Corpus: a large eye-tracking corpus with predictability norms. *Behav. Res. Methods* 50, 826–833
40. Hart, B. and Risley, T.R. (1995) *Meaningful Differences in the Everyday Experience of Young American Children*, Brookes Publishing
41. Wilcox, E.G. *et al.* (2025) Bigger is not always better: the importance of human-scale language modeling for psycholinguistics. *J. Mem. Lang.* 144, 104650
42. McCoy, R.T. *et al.* (2023) How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Trans. Assoc. Comput. Linguist.* 11, 652–670
43. Carlini, N. *et al.* (2023) Quantifying memorization across neural language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, OpenReview
44. Merrill, W. *et al.* (2024) Evaluating *n*-gram novelty of language models using Rusty-DAWG. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Al-Onaizan, Y. *et al.*, eds), pp. 14459–14473, Association for Computational Linguistics
45. Kaplan, J. *et al.* (2020) Scaling laws for neural language models. *arXiv* <https://doi.org/10.48550/arXiv.2001.08361>
46. Chang, T.A. *et al.* (2024) Characterizing learning curves during language model pre-training: learning, forgetting, and stability. *Trans. Assoc. Comput. Linguist.* 12, 1346–1362
47. Škrjanec, I. and Demberg, V. (2026) Language models that match reader experience are better predictors of reading times. *J. Mem. Lang.* 146, 104677
48. Rambelli, G. *et al.* (2023) Are frequent phrases directly retrieved like idioms? An investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multitask Expressions* (Bhatia, A. *et al.*, eds), pp. 87–98, Association for Computational Linguistics
49. Tirumala, K. *et al.* (2022) Memorization without overfitting: analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems* (Koyejo, S. *et al.*, eds), pp. 38274–38290, Neural Information Processing Systems Foundation, Inc
50. Lesci, P. *et al.* (2024) Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Ku, L.-W., ed.), pp. 15616–15635, Association for Computational Linguistics
51. Branford, J.D. and Franks, J.J. (1971) The abstraction of linguistic ideas. *Cogn. Psychol.* 2, 331–350
52. Jarvella, R.J. (1971) Syntactic processing of connected speech. *J. Verbal Learn. Verbal Behav.* 10, 409–416
53. Baddeley, A.D. and Hitch, G. (1974) Working memory. In *Psychology of Learning and Motivation* (Vol. 8) (Bower, G.H., ed.), pp. 47–89, Academic Press

54. Baddeley, A. (2003) Working memory and language: an overview. *J. Commun. Disord.* 36, 189–208
55. Lewis, R.L. and Vasishth, S. (2005) An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419
56. Futrell, R. et al. (2020) Lossy-context surprisal: an information-theoretic model of memory effects in sentence processing. *Cogn. Sci.* 44, e12814
57. Hahn, M. et al. (2022) A resource-rational model of human processing of recursive linguistic structure. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2122602119
58. 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*, International Committee on Computational Linguistics (ICCL)
59. Armeni, K. et al. (2022) Characterizing verbatim short-term memory in neural language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning* (Fokkens, A. and Srikumar, V., eds), pp. 405–424, Association for Computational Linguistics
60. Olsson, C. et al. (2022) In-context learning and induction heads. *arXiv* <https://doi.org/10.48550/arXiv.2209.11895>
61. Bietti, A. et al. (2023) Birth of a transformer: a memory viewpoint. *Adv. Neural Inf. Proces. Syst.* 37, 1560–1588
62. Jelassi, S. et al. (2024) Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning* (Salakhutdinov, R. et al., eds), pp. 21502–21521, PMLR
63. Vaidya, A. et al. (2023) Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning* (Jiang, J. et al., eds), pp. 58–69, Association for Computational Linguistics
64. Gruteke Klein, K. et al. (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* (Barak, L. and Alikhani, M., eds), pp. 219–230, Association for Computational Linguistics
65. Gibson, E. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76
66. McElree, B. et al. (2003) Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91
67. Yngve, V.H. (1960) A model and an hypothesis for language structure. *Proc. Am. Philos. Soc.* 104, 444–466
68. 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*, pp. 8705–8720, Association for Computational Linguistics
69. Bock, K. and Miller, C.A. (1991) Broken agreement. *Cogn. Psychol.* 23, 45–93
70. Pearlmutter, N.J. et al. (1999) Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456
71. Wagers, M.W. et al. (2009) Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237
72. Arehalli, S. and Linzen, T. (2024) Neural networks as cognitive models of the processing of syntactic constraints. *Open Mind* 8, 558–614
73. Bever, T.G. (1970) The cognitive basis for linguistic structures. In *Cognition and the Development of Language* (Hayes, J.R., ed.), pp. 279–362, Wiley & Sons
74. Frazier, L. and Rayner, K. (1982) Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210
75. Garnsey, S.M. et al. (1997) The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* 37, 58–93
76. Sturt, P. et al. (1999) Structural change and reanalysis difficulty in language comprehension. *J. Mem. Lang.* 40, 136–150
77. Arehalli, S. et al. (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* (Fokkens, A. and Srikumar, V., eds), pp. 301–313, Association for Computational Linguistics
78. 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*, pp. 30–36, Association for Computational Linguistics
79. Clark, C. et al. (2025) Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7735–7747, International Committee on Computational Linguistics
80. Thamma, A. and Heilbron, M. (2025) Human-like fleeting memory improves language learning but impairs reading time prediction in transformer language models. *arXiv* <https://doi.org/10.48550/arXiv.2508.05803>
81. Kuribayashi, T. et al. (2022) Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y. et al., eds), pp. 10421–10436, Association for Computational Linguistics
82. Ryu, S.H. and Lewis, R.L. (2025) Memory for prediction: a Transformer-based theory of sentence processing. *J. Mem. Lang.* 145, 104670
83. Warstadt, A. et al., eds (2023) *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Association for Computational Linguistics
84. Bender, E.M. and Koller, A. (2020) Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. et al., eds), pp. 5185–5198, Association for Computational Linguistics
85. Federmeier, K.D. and Kutas, M. (1999) A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495
86. Roland, D. et al. (2012) Semantic similarity, predictability, and models of sentence processing. *Cognition* 122, 267–279
87. Luke, S.G. and Christianson, K. (2016) Limits on lexical prediction during reading. *Cogn. Psychol.* 88, 22–60
88. 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*, International Conference on Learning Representations (ICLR)
89. Guo, C. et al. (2017) On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Precup, D. and Teh, Y.H., eds), pp. 1321–1330, PMLR
90. Liu, T. et al. (2024) Temperature-scaling surprisal estimates improve fit to human reading times—but does it do so for the “right reasons”? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Ku, L.-W. et al., eds), pp. 9598–9619, Association for Computational Linguistics
91. Meng, K. et al. (2022) Locating and editing factual associations in GPT. *Adv. Neural Inf. Proces. Syst.* 36, 17359–17372
92. Wang, S. et al. (2023) Knowledge editing for large language models: a survey. *arXiv* <https://doi.org/10.48550/arXiv.2310.16218>
93. Cohen, R. et al. (2024) Evaluating the ripple effects of knowledge editing in language models. *Trans. Assoc. Comput. Linguist.* 12, 283–298
94. Elman, J.L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 195–225
95. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.* 9, 1735–1780
96. 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* (Salakhutdinov, R. et al., eds), pp. 10041–10071, PMLR

97. Levy, O. *et al.* (2017) Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (Levy, R. and Specia, L., eds), pp. 333–342, Association for Computational Linguistics
98. Oberauer, K. *et al.* (2018) Benchmarks for models of short-term and working memory. *Psychol. Bull.* 144, 885–958
99. Linsley, D. *et al.* (2023) Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Adv. Neural Inf. Proces. Syst.* 37, 28873–28891
100. Linsley, D. *et al.* (2025) Better artificial intelligence does not mean better models of biology. *Trends Cogn. Sci.* <https://doi.org/10.1016/j.tics.2025.11.016>