Journal of Memory and Language 145 (2025) 104670

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Memory for prediction: A Transformer-based theory of sentence processing

Soo Hyun Ryu [a,b,*], Richard L. Lewis [a,b,c]

[a] Department of Psychology, University of Michigan, United States of America
[b] Weinberg Institute for Cognitive Science, University of Michigan, United States of America
[c] Department of Linguistics, University of Michigan, United States of America

ARTICLE INFO

Dataset link: github.com/soohyunryu/memory-for-prediction/

Keywords:
Sentence processing theory
Language and memory
Transformers

ABSTRACT

We demonstrate that Transformer-based neural network language models provide a new foundation for mechanistic theories of sentence processing that seamlessly integrate expectation-based and memory-based accounts. First, we show that the attention mechanism in GPT2-small operates as a kind of cue-based retrieval architecture that is subject to similarity-based interference. Second, we show that it provides accounts of classic memory effects in parsing, including contrasts involving relative clauses and center-embedding. Third, we show that a simple word-by-word entropy metric computed over the internal attention patterns provides an index of memory interference that explains variance in eye-tracking and self-paced reading time measures (independent of surprisal and other predictors) in two natural story reading time corpora. Because the cues and representations are learned, there is no need for the theorist to postulate representational features and cues. Transformers provide practical modeling tools for exploring the effects of memory and experience, given the increasing availability of both pre-trained models and software for training new models, and the ease with which surprisal and attention entropy metrics may be computed.

Introduction

The aim of this paper is to demonstrate that recent state-of-the-art neural network language models using the Transformer architecture (Vaswani et al., 2017) provide a new foundation for mechanistic theories of human sentence processing that seamlessly integrate probabilistic expectation-based (Hale, 2001, Hale, 2016, Levy, 2008, Levy, 2013) and working memory interference-based accounts (Lewis et al., 2006; Lewis & Vasishth, 2005; Vasishth & Engelmann, 2021). Because the language models compute a probability distribution over the lexicon conditioned on left context, they yield word-by-word *surprisals* which may be used to predict reading times, and we leverage this property in the analyses we present. Our primary contribution is to show that a simple word-by-word *entropy metric* computed from the Transformer's internal attention patterns has independent empirical power in accounting for processing difficulty and reading times.

The insight motivating this entropy metric is that the Transformer is a kind of multi-level *cue-based retrieval* architecture, and the entropy measure is an index of *similarity-based interference*—the more similar (with respect to retrieval cues) the candidates are that compete to be retrieved, the more diffuse is the attention (Ryu & Lewis, 2021). The theoretical connection between cue-based retrieval parsing and attention architectures is deep, because they are based on the same functional motivation: provide a memory mechanism that is capable of handling long-distance dependencies in natural language, given sufficiently discriminating cues and distinct memory representations (Lewis et al., 2006; Lewis & Vasishth, 2005). But modern Transformers differ sharply from earlier cue-based parsers in that the cues and representations are *learned*. Furthermore, the revolutionary AI advances made possible by Transformer-based language models provide important evidence for the *functional sufficiency* of cue-based architectures as the basis of human language processing.

The paper is structured as follows. We first provide brief background on the two main theoretical approaches that we integrate here: probabilistic surprisal-based approaches and working memory-based approaches. We then show, through processing visualizations and quantitative analyses, that the learned attention mechanism in GPT2 (an existing pretrained transformer, Radford et al., 2019) implements a kind of cue-based retrieval model that is consistent with earlier proposals for cue-based retrieval parsers, but moves beyond them in important ways. These analyses provide the motivation for the entropy metric that we use in the remainder of the paper.

We then show that entropy metrics from GPT2 provide accounts of classic phenomena that have motivated memory-based accounts, including contrasts between subject and object relative clauses in English and contrasts between center-embedding and right-branching, and

* Corresponding author.
E-mail address: soohyunr@umich.edu (S.H. Ryu).

https://doi.org/10.1016/j.jml.2025.104670
Received 31 October 2023; Received in revised form 15 June 2025; Accepted 18 June 2025
Available online 25 July 2025

furthermore they do so in ways that recapitulate the accounts of earlier cue-based retrieval models. Finally, we show that the surprisal and entropy metrics computed from GPT2 together provide an account of reading times in two separate reading-time corpora: the Ghent Eye-tracking Corpus (GECO) (Cop et al., 2017) (14 participants reading 5,031 sentences from a complete novel), and the Natural Stories Corpus, a self-paced reading corpus (Futrell et al., 2021) developed specifically to introduce complex syntactic structures that put pressure on working memory.

We chose to use GPT2-small model for two reasons. First, small models make analysis more widely accessible and future *interpretability* analyses more tractable. Second, Oh et al. (2022) found that the GPT2-small model is the most predictive of human reading times among a set of larger models. There have been more recent and complex models after Oh et al. (2022), but we do not think it is likely that the recent models will provide the basis for better predictions human reading times despite their higher performance in downstream tasks. In short, larger size and longer training do not make Transformer-based models more predictive of human reading times (Oh & Schuler, 2023a, Oh & Schuler, 2023b).

Our analyses suggest that Transformers are promising candidates for developing richer, adaptive and mechanistic models of human language processing and acquisition, due to their relative simplicity, the ease with which the surprisal and entropy metrics may be computed, and the increasing availability of both pre-trained models and software for training new models. This contribution thus provides an alternative approach to noisy memory surprisal models (Futrell et al., 2020; Hahn et al., 2022) as a route to integrating expectation and memory. Specifically, our approach complements lossy memory surprisal models by addressing a different aspect of memory: while lossy surprisal emphasizes the degradation of encoded memory over time, our approach focuses on memory retrieval interference during the interpretation of new input.

Background: Surprisal and memory-based accounts of sentence processing

We build here on a rich history of psycholinguistic work that aims to provide an account of word-by-word incremental sentence processing, especially work that focuses on intra-sentential dependency formation and the effects of structural complexity on processing and perception (e.g. Boland et al., 2001; L., 1987; Frazier & Fodor, 1978; Frazier & Rayner, 1982; Gibson, 1998; Hale, 2001; Hawkins, 1994; Levy, 2008; Lewis et al., 2006; Lewis & Vasishth, 2005; Miller & Chomsky, 1963; Vasishth & Engelmann, 2021). In particular, we build on two enduring ideas in sentence processing theory.

The first idea is that comprehenders form moment-to-moment predictions of upcoming words, and perception and comprehension processes depend on these predictions in important ways. The idea received a clear formal and general treatment in Hale (2001) introduction of *surprisal* into psycholinguistic theory. The surprisal or *Shannon information* (Shannon, 1948) of an event (or word) is the negative log probability of that event (see Hale (2016) for a tutorial on information theoretic metrics in psycholinguistics). Hale (2001) and Levy (2008) have shown that surprisal holds promise for accounting for a wide range of garden path effects (though precise quantitative tests reveal that surprisal may underestimate the magnitude of some garden path effects (Van Schijndel & Linzen, 2021). Subsequent work has demonstrated that word reading times vary linearly with surprisal (equivalently, they are a logarithmic function of probability), and that this relationship holds over a range of several orders of magnitude—a kind of quantitative relationship that is rare in psychology. Why should reading times covary linearly with surprisal? Levy (2008) offers one interpretation in terms of a rational update of a belief distribution over possible sentence interpretations. Optimal Bayesian perception (Lewis et al., 2013; Norris, 2009) provides another account. Our analyses here are agnostic about the underlying mechanism (though building more comprehensive dynamical models on a Transformer base could commit to such mechanisms).

The second idea is that human sentence processing is subserved by a bounded linguistic working memory (Caplan & Waters, 1990; Carpenter & Just, 1989; Chomsky & Miller, 1963; Gibson, 1991; Gordon et al., 2001; Lewis, 1993; Marcus, 1980; Miller & Isard, 1964). For example, Dependency Locality Theory (Gibson, 2000) assumes that the distance among dependents (measured in number of words or number of discourse referents, Gibson (2000)) affects the difficulty of dependency formation. In cue-based retrieval theory (Lewis et al., 2006; Lewis & Vasishth, 2005) dependency formation is bounded by similarity-based interference (Chomsky, 1965; Lewis, 1996; Lewis, 1998)—retrieving previous partial representations becomes more difficult when there is interference from similar candidate representations. In this view, linguistic working memory operates under principles governing all other kinds of human memory: the limiting factor is the capacity to discriminate target memoranda from similar distractors (Lewis, 1996; Shiffrin, 2003).

Surprisal has extraordinary empirical scope in accounting for sentence processing phenomena, and can be given a rational basis in a belief-update view of the function of linguistic perception (Levy, 2008). But two considerations motivate including bounded memory in an integrated model, one empirical and one theoretical.

The empirical motivation is that surprisal does not provide a complete account of comprehension difficulty associated with complex syntactic structures. A specific instance is that surprisal does not account for the difficulty at the embedded verb in an English object relative clauses, compared to an English subject relative (Levy & Gibson, 2013). Object relative clauses are less frequent, but the higher surprisal comes at the onset of the relative noun phrase, not the embedded verb (though there is some evidence for slowing at the onset NP as well; Staub (2010)). This single construction is interesting in part because it is just one of a class of difficult embedded structures that have motivated memory-based accounts (Futrell et al., 2020; Gibson, 1998; Lewis & Vasishth, 2005). Recent analyses of data obtained from participants reading natural stories or other content provide further evidence for a contribution of structural complexity independent of surprisal. Word-by-word metrics motivated by Dependency Locality Theory and a left-corner retrieval parser have been shown to account for reading times independent of surprisal (Shain et al., 2016), and structural complexity metrics have been shown to account for variance in fMRI (Brennan et al., 2020) and EEG (Hale et al., 2018) signals that is independent of surprisal (Hale et al., 2022). We add to this body of work here by testing a theoretically motivated attention-entropy metric from Transformers.

The theoretical motivation for integrating memory and surprisal is that surprisal does not itself provide an algorithmic processing account that specifies the computational architecture of sentence processing (Hale et al., 2022; Lewis, 2000)—the memories, representations, processes, and control structure of comprehension. In other words, regardless of whether surprisal provides a complete account of reading times, there is still the theoretical task of uncovering the computational mechanisms of sentence processing. The fact that memory and complexity structural metrics appear to leave empirical signatures that are independent of surprisal makes this task easier.

The contributions we report here follow in a line of previous work pursuing the goal of integrating surprisal and memory. Demberg and Keller (2009) introduced *Psycholinguistically Motivated Tree-Adjoining Grammar*, and extension of Tree-Adjoining Grammar (Joshi et al., 1975), to explain sentence processing difficulty that reflects both locality and surprisal effects. The lossy context surprisal theory of Futrell et al. (2020) adds noise to contextual memory when computing surprisal, and the surprisal values computed with this assumption of noisy context better predict human sentence reading time. Kuribayashi et al. (2022) showed that neural network language models better estimate

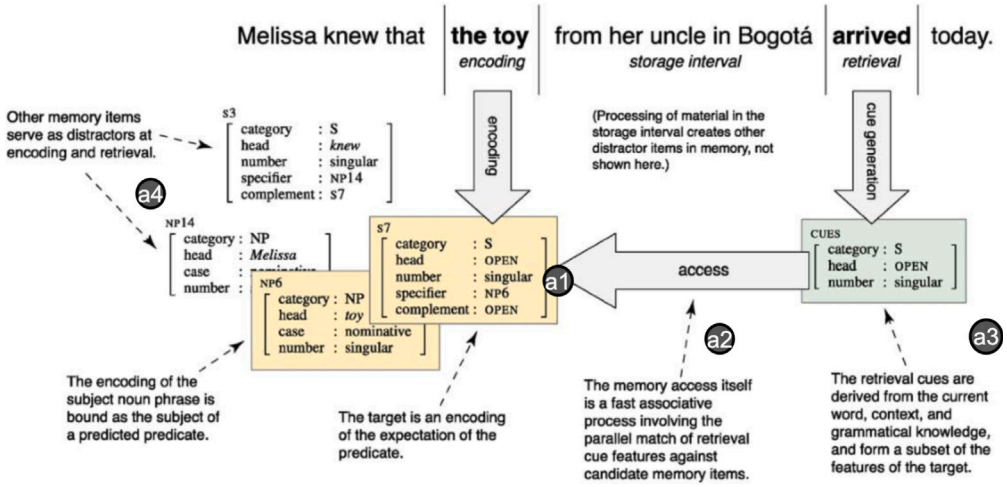


Fig. 1. Cue-based retrieval parsing, adapted from Lewis et al. (2006). The labels (a1)–(a4) refer to assumptions enumerated in the text.

human sentence reading patterns with limited context access. The resource rational surprisal model of Hahn et al. (2022) combines lossy context surprisal with the assumption that comprehenders make rational inferences given noisy memory representations and distributional facts about language.

Our primary contribution in this paper is a new integrative model in which memory retrieval interference, as quantified by Transformer attention entropy, makes a direct and independent contribution to reading time. In the next section we show that the internal attention patterns of GPT2 show the signatures of cue-based retrieval parsing, and we formally introduce the attention entropy metric.

Transformers as cue-based retrieval sentence processors

The purpose of this section is to establish the conceptual link between Transformer attention and cue-based retrieval sentence processing and define formally the attention entropy metrics used in subsequent analyses. We first briefly review the key aspects of cue-based processing and current challenges in developing it as a psycholinguistic theory, challenges that Transformers help to address.

Cue-based retrieval processing and its methodological challenges

Fig. 1, adapted from Lewis et al. (2006), illustrates the core assumptions of cue-based retrieval sentence processing (Lewis et al., 2006; Lewis & Vasishth, 2005; Lissón et al., 2021, Lissón et al., 2023; Van Dyke & Lewis, 2003) (we number these here (a1)–(a4) for later reference): (a1) word-by-word incremental formation of linguistic relations is mediated by the retrieval from memory of relevant prior linguistic representations; (a2) the retrieval process happens via a parallel match of cues against all candidate memory items; (a3) the cues are derived from the current word being processed via the application of proceduralized knowledge, which includes grammatical knowledge; and (a4) similarity-based interference arises to the degree that cues partially match distractor items in memory. This interference may lead to longer retrieval times of the correct target. It can also lead to increased probability of retrieving an incorrect distractor, which in the case of grammatical illusions can lead to a mixed distribution of correct and incorrect retrievals, and retrieval times with a mean that is faster than non-interfering baselines.

This broad sketch leaves unspecified many key details that are required to make empirical predictions. These include commitments to basic parsing strategies, whether single or multiple interpretations are maintained during processing, the nature of the intermediate memory representations, and the possible features that may be used as cues. For example, the ACT-R model of Lewis and Vasishth (2005) used a left-corner parser and syntactic representations and cues modeled after X-bar phrase structure trees (Baker, 1989; Chomsky, 1970).

The requirement that cue-based sentence processing models must commit to linguistic representations and specific cues is both a methodological virtue and challenge. One benefit is that it allows for empirical tests of different assumptions about cues (e.g., Yadav et al., 2022). The challenge is that the application to arbitrary linguistic stimuli requires a broad-coverage cue-based parser, and there have been few attempts to create one (Boston et al., 2011; Dotlačil, 2021). A related challenge is that it has not been clear how to model a limited beam multi-path parser that would presumably require multiple memory retrievals at each word. Pre-trained Transformers provide one way to address all of these challenges: they are broad coverage language models that may be applied to any linguistic stimuli, the representations and cues (*queries* in the context of transformers) are learned, and the model provides a concrete mechanism that supports multiple parallel retrievals and, at least implicitly, multi-path parsing.

Memory retrieval in the transformer: Scaled dot-product attention

The function of attention in the Transformer is to create dynamic pathways from representations of earlier parts of the sentence to the computations that create representations for the current word or token. Explicit memory retrievals open up these pathways through simple computations localized in each *attention head*. There are multiple attention heads operating at multiple layers of abstract representation. We describe here the computations local to a single attention head and layer[1] and refer the reader to the original Transformer paper (Vaswani et al., 2017) and the many online tutorials (e.g. 3Blue1Brown , Director) and implementations (e.g., Karpathy, 2024) for further details about how attention building blocks are composed into multiple layers.

In each attention head, the input embeddings of words in a sentence are converted into *keys*, *queries*, and *values*,[2] which are vectors of real numbers. *Queries* serve as the cue in cue-based retrieval processing, and *values* are the representations associated with prior words. Consider reading the sixth word (*were*) of a sentence, as in Fig. 2: there are five value vectors for the previous words (denoted $v_1, v_2, \ldots, v_5$ in Fig. 2), excluding the value vector for the sixth word itself.

[1] Recent work by Timkey and Linzen (2023) shows that interference effects arise in a recurrent neural net with only one attention head.

[2] This is unfortunate terminology when combining Transformers with value-based decision making. Value vectors in Transformers do not represent *value* in the decision theoretic or reinforcement learning sense.

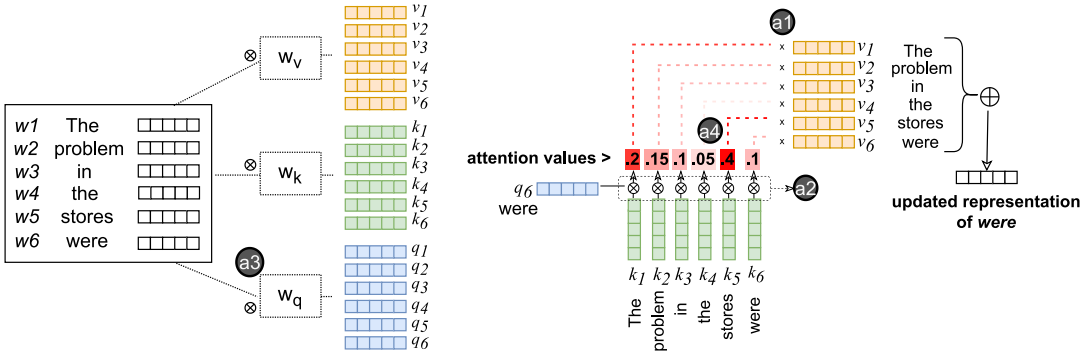


Fig. 2. How attention weights between a word (*were*) and representations associated with preceding words are computed in a Transformer's self-attention head. *Left:* Input embeddings of each word are multiplied by learned weight matrices and converted into value, key, and query vectors. *Right:* The query vector of *were* is matched against key vectors of preceding words. The value vectors of preceding words are summed using the attention weights to contribute to a contextual representation of *were* (we omit here the further processing of the summed vectors through a multi-layer perceptron). The labels (a1)–(a4) refer to assumptions enumerated in the text.

At this point, if the Transformer were a direct implementation of the simplest kind of cue-based processor, it would be natural to assume that q_6 is matched against $v_1 \ldots v_5$, and the closest matching value vector is retrieved. If this were the case, attention would be a *content-based retrieval* because the cues would be matched directly against the content of candidate retrieval items. But there are two twists here that make the Transformer different.

The first twist is that the query vectors are not matched directly against the value vectors. Instead, they are matched against *key* vectors of the same size that are associated with prior words (denoted $k_1 \ldots k_5$ in Fig. 2). The match produces a positive real scalar quantity that represents the degree of match, and it is computed by simply taking the dot-product of the query and key vectors and dividing by a scaling constant that is a function of the vector size; Vaswani et al. (2017) call this "scaled dot-product attention". The use of query, key, and value vectors and the calculation of attention weights through query–key similarity was an original design choice in the work of Vaswani et al. (2017) and all other Transformer models have followed it. Given the widespread adoption of Transformers across machine learning we think that if the separation of keys and values did not confer some functional benefit, simpler models would have been proposed by now. But we do not know of explicit systematic ablation studies that establish this benefit.

The second twist is the following. Rather than retrieving the value vector associated with the maximum score, a softmax computation transforms the set of match scores into a vector of quantities between 0 and 1 that sum to one, and these quantities are used to compute the *weighted sum of the value vectors*. This weighted sum of value vectors is the output of the attention process. These retrieved weighted-sums of values are then passed through further computations including a multi-layer perceptron to compute new vector representations as output. The weighted sum retrieval and the use of separate key vectors are a generalization of cue-based retrieval; it subsumes the simple case where the key vectors are just equal to the value vectors and where a maximum is taken to retrieve a single value.

Formally, if all the query vectors are combined as columns of a matrix Q, and keys and values combined into matrices K and V, the output of attention is a matrix of output vectors defined by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right) \quad (1)$$

where d_k is the dimensionality of the key and query vectors and QK^T accomplishes the dot products of all the query and key vectors. We are concerned here only with "causal" attention—the attention associated with a given word position is only allocated to prior words and not future words; this is now the standard paradigm in generative AI as well.

Transformer language models are built with multiple layers. In each layer, multiple attention heads perform multiple scaled dot-product attention retrievals and the outputs of each head are concatenated. These concatenated outputs are then projected to vectors that form the input to a subsequent layer of multi-headed attention, until finally a probability distribution over next-words is computed which is the final output of the Transformer at each word. In GPT2-small, there are 12 layers each with 12 independent attention heads. There are thus two distinct kinds of parallelism: each attention head can make a retrieval that combines multiple items in parallel, and there are multiple attention heads operating in parallel at each layer.

What is remarkable about the Transformer is that it is possible to train the architecture end-to-end so that the word embeddings, keys, values, query representations, and softmax parameters are all learned from the task of predict-the-next-token. (And with sufficient data, yields multi-modal models that form the basis of systems like ChatGPT.)

Attention as cue-based retrieval subject to interference

We can now specify the mapping of Transformer attention to cue-based retrieval sentence processing: all word-by-word computations in the Transformer are mediated by retrieval from prior representations (assumption (a1) above); attention happens via a parallel match of queries against all candidates for retrieval (a2); and queries are computed from representations of each word (a3). What remains to be established is that attention patterns show evidence of interference from distractor items that are not in grammatical relations with the current word (a4). The remainder of the paper provides extensive evidence for this, but we provide one simple illustrative example here.

Consider the example sentences in (1) involving subject–verb agreement in English. We are interested in the attention patterns at the verb, either *was* in (1 a) or *were* in (1 b). In both cases the preceding words are identical, and the head noun of the subject phrase is *problem*, although there is an agreement violation in (1 b). Prior work has shown that some attention heads in GPT2 distribute attention in ways that suggest they are processing specific grammatical relations such as subject–verb (Vig & Belinkov, 2019; Voita et al., 2019). We might expect therefore that if an attention head is processing subject–verb relations, this head would show relatively focused attention on the head noun *problem* in (1 a). But under the assumption that the plural number feature is part of the retrieval query, there should be additional attention weight given to the distractor *stores* in (1 b).

(1) Agreement interfering examples.

a. NON-INTERFERING The **problem** in the stores **was** solved by
b. INTERFERING The **problem** in the stores **were** solved by

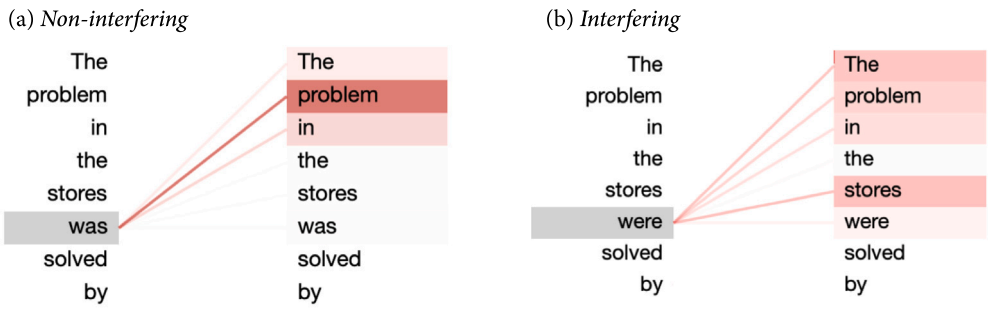


Fig. 3. Attention distributions at the critical verb (*was* and *were*) in one attention head (the third head in the fourth layer of GPT2-small) that may be implicated in computing subject–verb relations. The distribution is more diffuse in the interfering condition because of attraction to the distractor (*stores*) that matches the number feature of *were*.

This is exactly the pattern that we observe (Ryu & Lewis, 2021), as visualized in Fig. 3, which shows the attention weights computed by the attention head in GPT2-small that most reliably allocates maximal attention to subjects in subject–verb dependencies, using a visualization tool that Vig (2019) introduced. (The specific head is the third head in the fourth layer, henceforth referred to as $head_{4,3}$. This head was identified using the method introduced in Voita et al. (2019); see Appendix A). The attention in (1 a) is relatively focused and the attention in (1 b) is more diffuse, because of the additional weight given to the distractor NP. For a more in-depth illustration of similarity-based interference in agreement phenomena, we refer readers to Ryu and Lewis (2021), where both inhibitory and facilitatory interference effects are examined using attention entropies and surprisals. In that work, we argue that facilitatory effects are the result of *decreased* surprisal that can arise when interfering representations are used for prediction.

Attention entropy

The key idea we explore in this paper is that diffuse attention patterns in pre-trained Transformers are a signature of similarity-based interference. To quantify this at each word, we introduce *attention entropy*, following Ryu and Lewis (2021) and Ryu and Lewis (2022).

Recall that at each word w_i, each attention head computes a softmax vector of attention weights allocated to each previous word position. We denote the attention weight from a source word at position i to a target word at position $j < i$ by attention head h in layer l as

$$\text{Attn}_{l,h}(w_i, w_j) \tag{2}$$

which must be between 0 and 1.

To quantify the diffuseness of the attention from a source word i at a given attention head, we use Shannon (1948)'s information entropy, which has the properties that we need: it is well-defined over weights that sum to 1, it is at maximum when the attention weights are equal, and minimum when all the attention weight is on one element. Formally, the attention entropy at word i for attention head h in layer l is:

$$\text{AttnEntr}_{l,h}(w_i) = -\sum_{j=1}^{i-1} \text{Attn}_{l,h}(w_i, w_j) \times \log_2 \text{Attn}_{l,h}(w_i, w_j) \tag{3}$$

where i refers to the location of the current source word, and the j's are locations of prior words.

Shannon entropy is a measure of uncertainty defined over probability distributions, but we are using entropy for its convenient mathematical properties—we know of no evidence that Transformers treat attention weights as probability distributions.

Aggregate attention entropy as a word-by-word processing metric

The attention entropy can be measured either using a single attention head (for example one that is conjectured to specialize for a specific grammatical dependency), or an *aggregate attention entropy* over multiple attention heads. While we will visualize attention entropies for selected heads, for our reading time analyses we compute an aggregate metric over a set of multiple attention heads. We use the simple mean in Eq. (4) as a measure of aggregate attention. If H is a set of attention heads of size $|H|$ in which each head is identified by a head index h in layer l of a Transformer, then:

$$\text{AggrAttnEntr}(w_i) = \frac{\sum_{(l,h)\in H} \text{AttnEntr}_{l,h}(w_i)}{|H|} \tag{4}$$

where again, l and h indicate indices of layers and heads in the Transformer.

For GPT2-small the simplest aggregate measure takes the mean over all 144 attention heads. A concern in doing this is that many attention heads have linguistically uninteresting attention patterns (such as always attending to the first token or the previous token) and some may not even be functionally implicated in the Transformer's final output (Voita et al., 2019), thereby adding noise to the predictors. In the analyses we report here, we select a subset of attention heads (about 14% of the total set) for which there is some evidence that the heads are computing intra-sentential grammatical dependencies.

Our selection method is based on the method in Voita et al. (2019) mentioned earlier, and the details are in Appendix A. In a nutshell, what our method does is find attention heads that reliably allocate most of their attention to the dependent or governor of one of the 43 grammatical dependency types in the CoreNLP Stanford dependency parser (Manning et al., 2014), which we use to provide the "gold standard" against which to compare the attention patterns. As Voita et al. (2019) points out, simple fixed position-based heuristics can account for many dependencies—for example, the subject head is very often (in fact 41% of the time) in the position immediately before the verb. So we impose a threshold that an attention head must exceed at least 10% of this positional baseline; e.g. for subject–verb relations, it must allocate maximum attention to subjects in at least 45.1% of cases. For some dependency types, multiple attention heads meet this threshold, and we select the one with the highest score. Using this criterion, twenty out of 144 total attention heads are selected, covering 29 of the 43 dependency types.

We conducted a preliminary analysis to assess whether this selected subset of 20 attention heads enhances the predictive power of attention entropy for human reading times from Natural Stories Corpus (Futrell et al., 2021). Specifically, we computed: (a) attention entropy using the 20 selected syntactic attention heads, (b) attention entropy using the 122 non-selected attention heads, and (c) attention entropy using all 144 attention heads as a baseline. We then fit three simple frequentist models incorporating these different entropy measures, surprisals, and other predictors (e.g., word frequencies, word lengths, word position in a sentence), using a model to be specified in detail later in the section *Modeling Sentence Reading Times*, and compared the log-likelihood differences between (a) vs. (c) and (b) vs. (c). The log-likelihood difference between (a) and (c) was 117.29, whereas the difference between (b) and (c) was −11.73. These findings suggest that this head selection

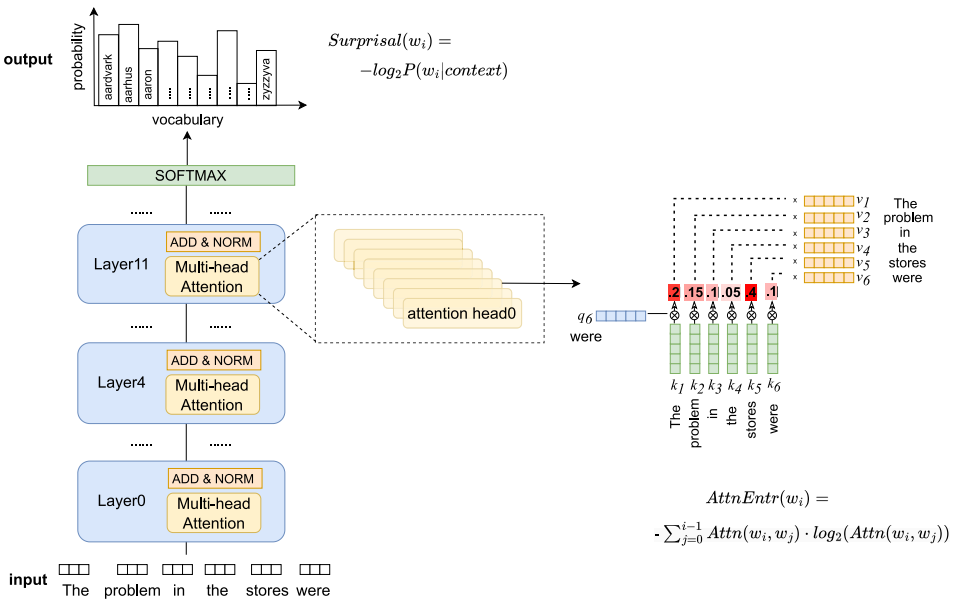


Fig. 4. The internal working mechanisms of GPT2, showing how the Transformer architecture integrates expectation-based and memory-based aspects of sentence processing. Surprisals are computed with the last hidden state of the model, and attention entropies are computed based on attention distributions in attention heads. Attention entropy is proposed as an index of similarity-based retrieval interference (Ryu & Lewis, 2021).

procedure improves the ability of attention entropy to predict human reading times, arguably by reducing noise introduced by nonfunctional heads or heads with degenerate attention patterns.

In this paper, we utilize aggregate attention entropy and attention entropies from specific heads depending on the research questions we aim to address or phenomena we wish to illustrate.

Other Transformers' attention-based reading time metrics have been proposed. Ryu and Lewis (2022) (the work forming the basis of the analyses we present here) used a global metric computed over all attention heads. Oh and Schuler (2022) used an entropy-based metric computed from attention heads in the final *topmost* layer. The metric we used here is based on a different selection procedure and we offer it as a complementary approach. Indeed, none of the 20 heads selected by our method is in the final layer, consistent with findings that intermediate layers are computing representations of sentence structure (Hoover et al., 2019; Zini & Awad, 2022).

Fig. 4 provides an overview of how we use GPT2-small to compute both word-by-word surprisals and word-by-word aggregate attention entropies. In the next main section we use these two predictors in analyses of classic embedded structures, followed by analyses of reading time measures from both self-paced reading and eye-tracking corpora.

Relationship to the entropy reduction hypothesis

Before we explore the empirical implications of the attention entropy metrics, it is useful to make a few clarifying remarks concerning the relationship of attention entropy to the *Entropy Reduction Hypothesis* (ERH) of Hale (2006) (see also Hale (2016)). *Entropy* in ERH is a quantity measuring uncertainty about sentence interpretation. ERH posits that information processing work is done when the sentence processor reduces this uncertainty and that this uncertainty reduction work takes time that should be measurable in word-by-word empirical measures. The entropy of *attention entropy* is a quantitative measure of diffuseness of the attentional pattern; there is no commitment in the theory that it characterizes uncertainty about what to attend to or that it characterizes uncertainty about sentence interpretation. Nonetheless, there are possible interesting connections among these quantities and we take this up in the final Discussion.

Relative clauses and center embeddings

English object relative clauses (vs. subject relative clauses) and center-embedding sentences (vs. right-branching sentences) are two classes of constructions that have played an important role in motivating memory bounds in sentence processing since Chomsky and Miller (1963); (e.g. Christiansen & MacDonald, 2009; Gibson, 1991; Lewis, 1993). If attention entropy is an index of similarity-based interference, and if such interference is the source of memory-based difficulty in sentence processing, then it should be possible to directly see entropy effects in these constructions. In this section we examine such effects, and show in particular that attention entropy dissociates from surprisal as it predicts increased difficulty at the verb in English object relative clauses.

English subject- and object-relative clauses

As Levy and Gibson (2013) have pointed out, the locus of difficulty in object-extracted clauses is useful in distinguishing expectation-based and memory-based theories. Given that subject-extracted relative clauses (SRCs) are more frequent than object-extracted clauses (ORCs), expectation-based theories predict that language comprehenders may anticipate a verb after processing *that*. Thus, the locus of the difficulty of ORCs predicted by expectation-based theories is the onset of the noun phrase (*the* in (2 b)) since it is the place where the expectation towards a verb is unrealized, and surprisals computed from early PCFG-based models as well as recent large language models (including the analyses we show below) align with this intuitive prediction. But importantly, surprisal does *not* predict an increase in difficulty at the embedded verb of the ORC. In contrast, memory-based theories predict the locus of ORCs' processing difficulty is the verb in the relative clauses for a variety of reasons; for retrieval interference accounts the embedded clause verb (*attack* in (2 b)) has two candidates as its dependent subject in ORCs (*reporter* and *senator* in (2 b)) while only one candidate in SRC (*reporter* in (2 a)), and two syntactic relations must be computed.

(2) a. *Subject Relative Clause (SRC)*
The reporter$_i$ that t_i **attacked$_{rcv}$ the$_{rcn}$** senator admitted the error.

Table 1
An example set of manipulated materials used for the experiment on relative clause processing.

	Materials used to compute difficulty metrics at the noun onset
SRC	Yesterday, the girl who watched *the* parents changed a critical part of the story.
ORC	Yesterday afternoon, the girl who *the* parents watched changed a critical part of the story.
	Materials used to compute difficulty metrics at the embedded verb
SRC	Yesterday at noon, the girl who *watched* the parents changed a critical part of the story.
ORC	Yesterday, the girl who the parents *watched* changed a critical part of the story.

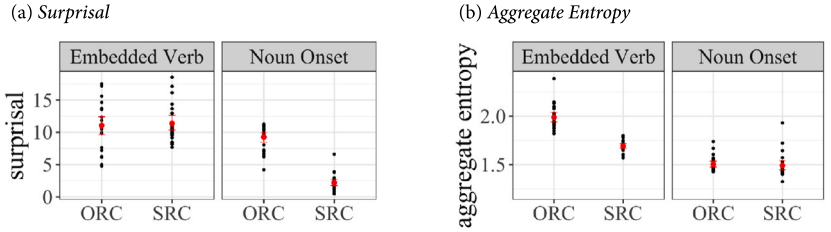


Fig. 5. Surprisals and aggregate attention entropies are measured at the noun onsets and at the embedded verbs (e.g., at *the* and *followed* in Table 1) in SRC and ORC sentences. The red dots and lines indicate means and 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

b. *Object Relative Clause (ORC)*
The reporter$_i$ that **the$_{rcn}$** senator **attacked$_{rcv}$** t_i admitted the error.

Most reading time studies have shown that the significantly greater processing difficulty differential between ORCs and SRCs is indeed found at the relative clause verb (Grodner & Gibson, 2005; Levy et al., 2013), which is aligned with the prediction of memory-based theories. But Staub (2010) also found increased difficulty at the RC noun onset of the ORC, where there were increased regressions. The maze task results of Vani et al. (2021) also showed that slowdowns were observed in the RC noun onset.

In the following analysis, we examine both surprisal and aggregate attention entropy at critical words (embedded verbs and noun onsets) in the SRC and ORC materials of Staub (2010). We anticipate replicating the finding that higher surprisals occur at the ORC noun onset but not the embedded verb. But we also anticipate greater attention entropy at the ORC embedded verb.

Methods

In each of the critical sentences in the materials from Staub (2010), we used GPT2-small to compute surprisal and aggregate attention entropy (Eq. (4)) at two critical regions: the onset of the relative clause noun and the relative clause verb (see (2) above).

Materials

24 sets of sentences from Staub (2010) were included. In order to prevent confound effects from the position of the critical word, the example sentences were manipulated to have an adverbial phrase at the beginning of the sentences (e.g., *'yesterday'*, *'yesterday afternoon'* or *'yesterday at noon'*). Different manipulations were required to control for the position of the relative clause noun onset and for the position of the relative clause verb. The manipulated structures are as in Table 1.

Results

As shown in Fig. 5, greater surprisals of ORCs are observed at the onset of the relative clause noun, but not the embedded verbs, consistent with earlier surprisal analyses (Levy & Gibson, 2013). But greater aggregate attention entropies of ORCs are observed at the relative clause verb, consistent with the assumption that greater retrieval interference occurs at the embedded ORC verb.

Fig. 6 illustrates how the attention at head$_{4,3}$ from the embedded verb is more diffuse in an object relative clause than in a subject relative clause. We show this head because it is most likely to be one of the heads functionally important at the verb site because it is may be computing subject–verb relations—but the graphs in Fig. 5 show the same aggregate entropy (over 20 heads) that we use in later analyses of reading times in the self-paced reading and eye-tracking corpora.

Right vs. center embeddings

Center-embedded constructions have constituent structures of the form

(3) Center embedding: $[_\alpha \ \ldots \ [_\beta \ \ldots \ ] \ \ldots \ ]$

where the constituent β is embedded within constituent α with material on either side. If α and β are the same type of constituent (under some theory of types) then the construction is an instance of *self-embedding*. In contrast, right (or left) embedding (or branching) involves structures of the form

(4) Right embedding: $[_\alpha \ \ldots \ [_\beta \ \ldots \ ] \ ]$

where the material in β closes both the constituent β as well as α. Intuitively, there is a memory load difference: in (3), after β is processed a memory of the first part of α is required to compute further relations within α. Formally, arbitrary center embeddings are precisely the structures that render grammars outside the scope of finite state (limited memory) automata (Chomsky & Miller, 1963).

In (5 a) for example, a constituent *that the judge sentences* is center-embedded in the constituent *that the inmates . . . played* which is center embedded inside the constituent *the game. . . involved bats and balls.* Such *double* center embeddings are famously difficult to process and most people reject them as ungrammatical on first listening or reading (Miller & Isard, 1964; e.g., Hakes et al., 1976).

(5) a. *Center embedding*
The game$_{N\text{-level}1}$ [that the inmates$_{N\text{-level}2}$ [that the judge$_{N\text{-level}}$ sentenced$_{V\text{-level}3}$] played$_{V\text{-level}2}$] involved$_{V\text{-level}1}$ bats and balls.
b. *Right branching*
The judge$_{N\text{-level}3}$ sentenced$_{V\text{-level}3}$ the inmates$_{N\text{-level}2}$ who played$_{N\text{-level}2}$ the game$_{N\text{-level}1}$ that involved$_{V\text{-level}1}$ bats and balls.

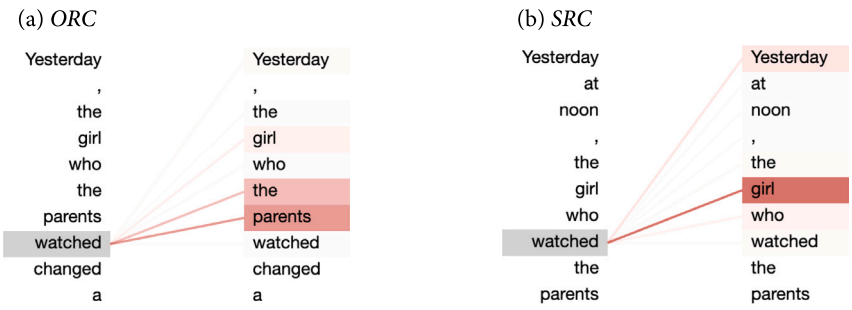


Fig. 6. Visualizing attention patterns at the embedded verb (*watched*) of object and subject relative clauses. Shown are the attention patterns of $head_{4,3}$, selected for visualization because it reliably allocates its maximum attention to the subject when processing a verb and is thus plausibly implicated in processing subject–verb relations (see Appendix A for details). The distribution is more diffuse in the object relative clause at the embedded verb because attention is distributed to two noun phrases (*girl, the, parents*).

The right-branching counterpart (5 b) is processed easily even though the number of items to process is exactly the same in both types of sentences (Lewis, 1996; Yngve, 1960). Note that the nouns and verbs have been labeled with level numbers 1, 2 and 3, where 3 is the most embedded level. We use this descriptive level reference in our analyses below.

The study of Lewis and Vasishth (2005) showed that a memory-based model that hypothesizes sentence processing is shaped by similarity-based interference can successfully account for the differential processing difficulty of center-embedded and right-branching sentences. Specifically, it showed that the processing time does not dramatically increase for deep right branching sentences at embedded verbs contrary to center embedded sentences. In addition, their model also simulated misanalyses of dependencies in center-embedded sentences that result in the elimination of the correct candidate for subsequent retrievals. In other words, failure to attaching the retrieval verb to the correct target can lead to the final verb being left without its correct unattached subject, generating further difficulty. For instance, when the level2 verb (*played* in (5 a) is incorrectly attached to the level 1 noun (*game* in (5 a)) rather than the correct level 2 noun (*inmates*), the correct dependent is not available for the final verb.

In the following analyses we examine attention entropy and surprisal at the verbs in the center and right-branching sentences from Stolz (1967) study. We anticipate observing higher attention entropies at the level 2–3 verbs in center-embedded sentences compared to right branching due to the larger number of potentially interfering noun phrase dependents. Also, we anticipate the higher surprisal at the final verb (level 1 verb) in the center embedded constructions than in the right branching constructions. This is expected because the language model has computed a state in which the most likely subject dependent of the middle verb is the first noun phrase, making the continuation with a third verb surprising. This anticipation is consistent with an account of the grammatical illusion identified by Janet Fodor many years ago: dropping the middle verb in double center embeddings leads to increased acceptability (Christiansen & MacDonald, 2009; Gibson & Thomas, 1996; Häussler & Bader, 2015; Huang & Phillips, 2021).

In addition to aggregate attention entropy, we visualize the attention patterns of a single attention head in GPT2-small whose maximum attention falls on the subject in the majority of subject–verb relations ($head_{4,3}$ that was introduced earlier). Although our analyses do not provide direct evidence of mis-retrievals or relations established in error, we provide indirect evidence of the possibility of such errors in the attention patterns associated with this head.

Methods

Three metrics from GPT2-small were computed at all three verbs in the sentences: surprisal, aggregate attention entropy (Eq. (4)), and the amount of attention given to prior nouns from the verbs in the head selected for its correlation with the subject–verb dependency (Eq. (2)).

In order to prevent confound effects from the word position, we added adverbial phrases at the beginning of right branching sentences at levels 2 and 3 (e.g., *'Yesterday evening'* or *'Before the sun went down'*). By doing so, we ensured that verbs of interest appear at the same position for both center-embedded and right-branching sentences. Additionally, in order to see whether early misattachments of an embedded verb and its subject might be possible in the center-embedded sentences, we also measured how much attention is paid to each noun in embedded sentences at all verb levels.

Materials

Fifteen sets of sentences from Stolz (1967) were included, with the same structure as in Table 2. Each sentence includes three pairs of subjects and verbs, which are listed by level in the table.

Results

Consider first the visualization in Fig. 7 of attention patterns at $head_{4,3}$, the head that most reliably allocates maximum attention to subjects in subject–verb dependencies. There is a striking contrast between center and right branching sentences in how sharply focused attention is allocated to the correct subject dependent. This is so even though the lexical items and thus semantic constraints are identical.

Fig. 8 visualizes the amount of attention allocated to each noun by $head_{4,3}$ at the level 2 (middle verb) and level 3 verbs. In right branching sentences, at the level 2 verb by far the most attention is allocated to the correct level 2 subject noun. But in the center-embedded sentences, the attention is allocated more uniformly across the three nouns, with the greatest amount of attention allocated to the noun that is neither the subject of the level 1 (main) verb (Fig. 8 a) nor the correct object gap-filler.

Finally, Fig. 9 shows four different quantities computed at each of the three verbs: surprisal, aggregate attention entropy, the attention paid to the correct subject target by attention $head_{4,3}$, and the attention entropy at attention $head_{4,3}$. Both attention entropy metrics are higher at the second and third verbs in center-embedded sentences, and attention to the correct subject noun is lower for the second and third verb in center-embedded sentences[3]. Surprisal is higher at the final verb (level

[3] The processing of the CE and RB level 2 and 3 verbs differ in another important way: in the CE constructions, there are two dependencies being processed: a subject dependency and the object filler-gap dependency, but in the RB constructions, only the subject filler-gap dependency is being processed. Could the slow down be due to the fact that there are two dependencies rather than one being computed? The issue comes down to what extent attention heads such as $head_{4,3}$ are really specialized for subject and not object relations. The analysis of the correlation of attention patterns with grammatical dependencies in Appendix A does suggest that $head_{4,3}$ is patterning primarily with subject and not filler-gap object relations, consistent with the interpretation that the increased interference observed in Fig. 9 for $head_{4,3}$ is due to subject interference and not the addition of object gap filling. Furthermore, the noun that $head_{4,3}$ pays most attention to at level 2 is *neither* the correct subject nor the correct object gap-filler. But systematically disentangled the effects of multiple dependencies vs. interference in single dependencies will require further computational experiments.

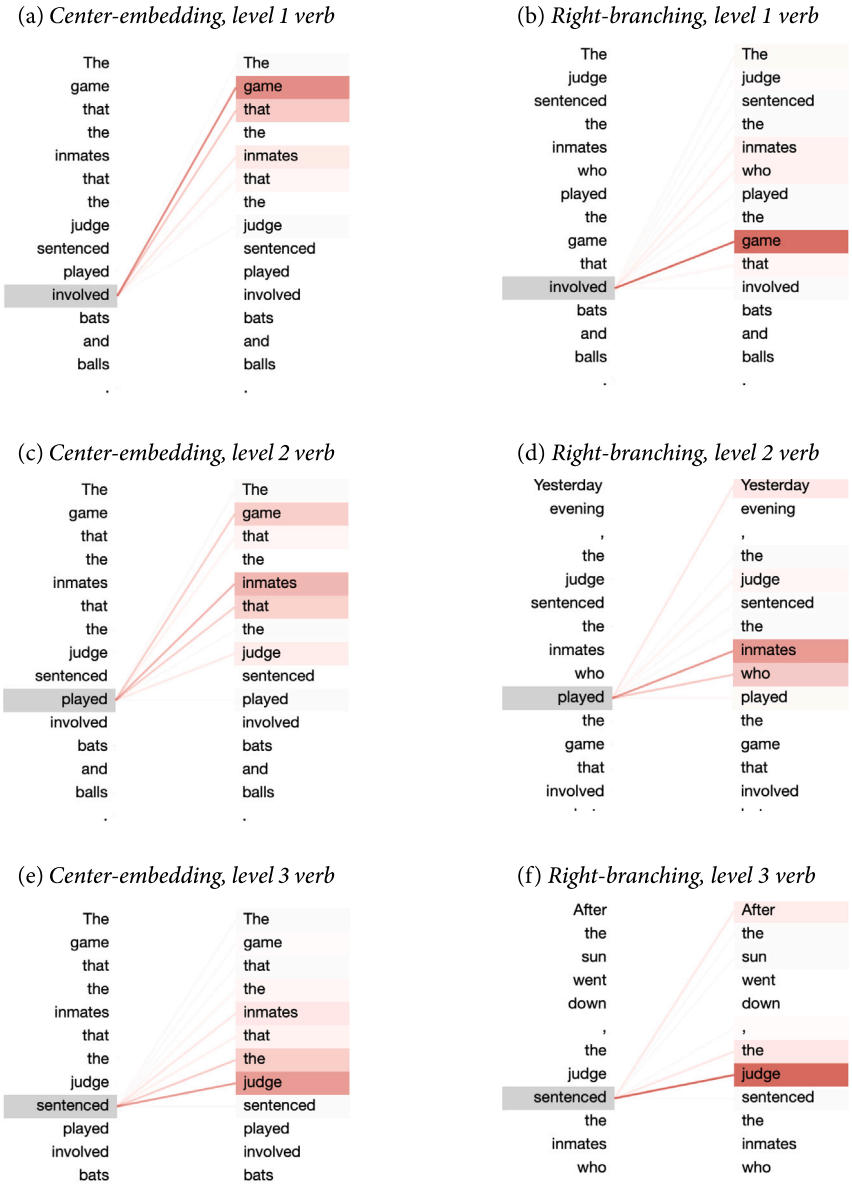


Fig. 7. Attention patterns at the embedded verbs at different levels of center-embedded and right-branching structures. Shown are patterns at $head_{4,3}$, the head that most reliably allocates maximum attention to subjects in subject–verb dependencies. There is a striking contrast between right branching and center embedded structures in the degree to which attention is sharply focused on the correct subject at the two innermost verbs.

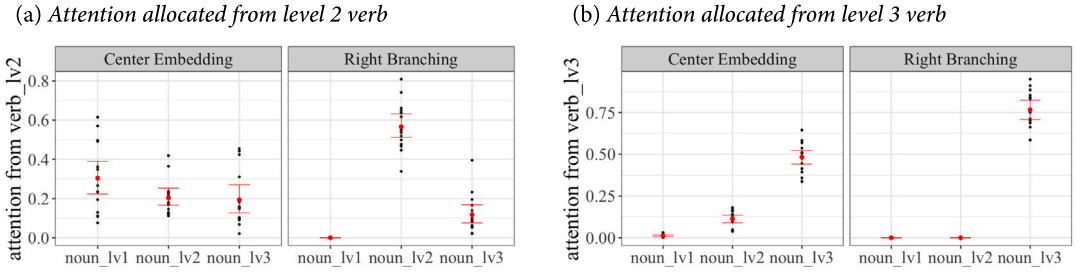


Fig. 8. Attention allocated to nouns from the level 2 and level 3 verbs (*played* and *sentenced* respectively, in Table 2) by the attention head ($head_{4,3}$) that most reliably allocates maximum attention to subjects in subject–verb dependencies. Attention is sharply allocated to the correct subject noun in the right branching sentences, but in the center-embedded sentences, at the middle level 2 verb (*played* in Table 2) attention is diffusely allocated, with most attention on a noun (*game* in Table 2) which is neither the correct subject nor the correct object gap-filler. The red dots and lines indicate means and 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
An example set of materials used for the experiment on embedded sentence processing.

Center-embedded	LEVEL1	The **game**$_{target}$ that the inmates that the judge sentenced played **involved**$_{cue}$ bats and balls.
	LEVEL2	The game that the **inmates**$_{target}$ that the judge sentenced **played**$_{cue}$ involved bats and balls.
	LEVEL3	The game that the inmates that the **judge**$_{target}$ **sentenced**$_{cue}$ played involved bats and balls.
Right-branching	LEVEL1	The judge sentenced the inmates who played the **game**$_{target}$ that **involved**$_{cue}$ bats and balls.
	LEVEL2	Yesterday evening, the judge sentenced the **inmates**$_{target}$ who **played**$_{cue}$ the game that …
	LEVEL3	Before the sun went down, the **judge**$_{target}$ **sentenced**$_{cue}$ the inmates who …

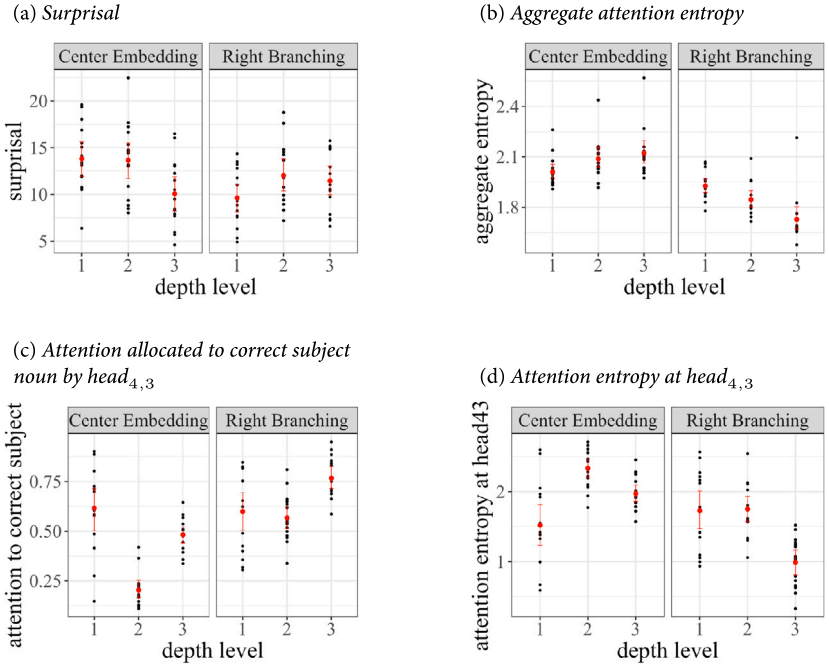


Fig. 9. Four metrics computed at the embedded verb at three levels of center and right embedding. Depth levels 1–3 indicate *involved*, *played*, and *sentenced* in Table 2 respectively. The top two panels (a) *surprisal* and (b) *aggregate attention entropy* are those metrics used in subsequent reading time analyses. Panels (c) and (d) are derived from the attention patterns of head$_{4,3}$, a head selected because it reliably allocates attention to subjects of verbs. Note that at level 2 this head shows both high entropy and very low allocation of attention to the correct subject. The red dots and lines indicate means and 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1 verb) in the center embedded sentence than in the right branching sentence.

These patterns are well aligned with the memory retrieval interference-based explanation about processing difficulty of center-embedded sentences (Lewis & Vasishth, 2005), which has a role for both increased similarity-based interference as well as incorrect attachments of subjects and verbs.

Modeling sentence reading times with surprisal and attention entropy

In this experiment, we examine whether word-by-word aggregate attention entropy, computed using GPT2-small, provides additional predictive power beyond surprisal in accounting for word-level reading times collected in naturalistic reading settings, building on our prior work (Ryu & Lewis, 2022). We follow here a similar analysis by Shain et al. (2016) which used alternative memory-based predictors. We ran two analyses using different types of reading time measurements: self-paced reading times and eye-tracking data.

Materials

For the self-paced reading time data, we used the Natural Stories corpus (Futrell et al., 2021), which includes self-paced reading times for 485 sentences read by 180 participants. Sentences in this corpus were designed to include syntactic constructions that generate psycholinguistically interesting phenomena which show the role of memory in sentence processing more clearly than most commonly-used sentence structures. For eye-tracking data, we used the Ghent Eye tracking corpus (GECO, Cop et al., 2017) in which 5031 sentences from novels are read by 14 monolingual English speakers.

Methods

For every word in the corpus, we computed both surprisal and aggregate attention entropy (Eq. (4)) using GPT2-small. To take spillover effects into account, we also included predictors from the immediately preceding word. Surprisal for each word was calculated using the largest possible context window size, which is 1024. This decision is made based on previous finding that GPT2's surprisal becomes the most predictive when the context window is about 1000 (Gao

& Yu, In prep). Aggregate attention entropy was calculated with the context window size of 30 as we found this makes the aggregate attention entropy the most predictive of human reading times. A window size of 30 means that the entropy metric is focused on attentional patterns mostly within the current sentence and previous sentence—aligning with our selection of attention heads based on intra-sentential grammatical dependencies. See Appendix B for the details.

We excluded data points where a word is read in 0 ms or in more than 3,000 ms from analyses in both reading time data. When GPT2's byte-pair encoding (BPE) tokenizer recognizes a word as a combination of multiple tokens, we took the maximum values of surprisal and aggregate entropy of the subtokens. We chose maximum rather than sum or mean to avoid disproportionately increasing (with sum) or decreasing (with mean) the entropy metric for individual low frequency words which happened to be split into subtokens.

We fit a Bayesian linear mixed model with surprisal and entropy predictors for the current and previous words and many other control variables including position, word length and word frequency computed from Google N-gram corpus (Michel et al., 2011) (see Table 3). The predictors computed at the current word will be subscripted with w and the ones computed at the previous word will be subscripted with $w-1$. We included participant IDs as a random slope and word types as a random intercept. All predictors were transformed into z-scores except for the word position. The model was fit using the R package *brms* (Bürkner, 2017). Uninformative priors were specified as follows: The intercept prior assumes a normal distribution centered at 1000 with a standard deviation of 1000; for the regression coefficients (b), we assume a normal distribution with a mean of 0 and a standard deviation of 500; The standard deviation (sd) prior also follows a normal distribution with a mean of 0 and a standard deviation of 500; the correlation matrix prior employs an LKJ distribution with a shape parameter of 1. The model specification follows the approach outlined in previous work on reading time modeling by Boyce and Levy (2023).

For the self-paced reading time model, the dependent variable is the time a reader takes before moving on to the next word; and for the eye-tracking model, we tested two dependent variables: the first fixation duration, which is the time spent on the first fixation of the current word and the go-past time, which is the total of all fixations before moving to the right of the current word, including any regressions to earlier words. We included these three reading time measures for two reasons. First, we aimed to assess whether attention entropy effects, if present, would emerge consistently across multiple datasets. Second, we hypothesized that different reading time measures may capture different stages of sentence processing. Specifically, first fixation duration is thought to reflect early processing stages, such as word recognition and initial lexical access. In contrast, go-past time is assumed to reflect a combination of early and later processing stages, as it includes time spent on regressions, which are often associated with processes like memory retrieval or syntactic dependency formation. We did not make strong a priori assumptions about the specific stage of processing indexed by self-paced reading times, due to the inherent challenges in localizing effects within this paradigm (Boyce & Levy, 2023; Witzel et al., 2012) that make it difficult to disentangle whether observed effects stem from initial processing difficulty or delayed integration and reanalysis processes.

Results

Results from the models are provided in Tables 3, 4, 5.

The posterior distribution of estimates for predictors of interest—*surprisal* and *attention entropy*—are shown in Fig. 10. Not surprisingly we replicate the quite strong surprisal and spillover surprisal effects. In addition to that, the results show the attention entropy has an influence on reading times in both self-paced reading and eye-tracking, independent of surprisal and the other effects.

Table 3
Posterior estimates for the fixed effects of predictors on self-paced word reading times in the Natural Stories Corpus (Futrell et al., 2021).

Parameters	Estimate	95% CrI
intercept	321.64	[310.54, 333.68]
word position$_w$	−0.01	[−0.10, 0.07]
word length$_w$	4.44	[2.47, 6.43]
surprisal$_w$	7.74	[6.70, 8.84]
aggregate entropy$_w$	2.27	[1.47, 3.05]
word frequency$_w$	−5.32	[−7.82, −2.57]
surprisal$_w$ * aggregate entropy$_w$	0.90	[0.46, 1.35]
word length$_{w-1}$	−0.09	[−1.13, 0.87]
surprisal$_{w-1}$	5.03	[4.30, 5.80]
aggregate entropy$_{w-1}$	0.36	[−0.27, 0.96]
word frequency$_{w-1}$	−2.33	[−3.10, −1.56]
surprisal$_{w-1}$ * aggregate entropy$_{w-1}$	0.83	[0.44, 1.20]

Table 4
Posterior estimates for the fixed effects of predictors on first-fixation durations in the GECO corpus (Cop et al., 2017).

Parameters	Estimate	95% CrI
intercept	213.70	[194.19, 234.24]
word position$_w$	−0.11	[−0.25, −0.03]
word length$_w$	−0.23	[−1.66, 1.24]
surprisal$_w$	3.29	[2.46, 4.12]
aggregate entropy$_w$	−0.98	[−2.40, 0.39]
word frequency$_w$	−4.06	[−5.46,−2.68]
surprisal$_w$ * aggregate entropy$_w$	0.39	[0.01, 0.77]
word length$_{w-1}$	−2.77	[−3.97, −1.61]
surprisal$_{w-1}$	1.75	[0.93, 2.57]
aggregate entropy$_{w-1}$	2.35	[0.86, 3.91]
word frequency$_{w-1}$	−0.06	[−0.83, 0.71]
surprisal$_{w-1}$ * aggregate entropy$_{w-1}$	0.28	[−0.06, 0.63]

Table 5
Posterior estimates for the fixed effects of predictors on go-past times in the GECO corpus (Cop et al., 2017).

Parameters	Estimate	95% CrI
intercept	331.23	[293.99, 367.93]
word position$_w$	0.38	[−0.14, 0.90]
word length$_w$	16.38	[8.62, 24.58]
surprisal$_w$	9.84	[7.45, 12.23]
aggregate entropy$_w$	3.27	[0.06, 6.43]
word frequency$_w$	−2.53	[−6.44, 1.14]
surprisal$_w$ * aggregate entropy$_w$	1.34	[−0.31, 3.02]
word length$_{w-1}$	−10.16	[−13.69, −6.72]
surprisal$_{w-1}$	10.95	[7.87, 14.01]
aggregate entropy$_{w-1}$	−3.12	[−6.40, 0.16]
word frequency$_{w-1}$	−5.11	[−8.32, −1.76]
surprisal$_{w-1}$ * aggregate entropy$_{w-1}$	−1.08	[−2.65, 0.51]

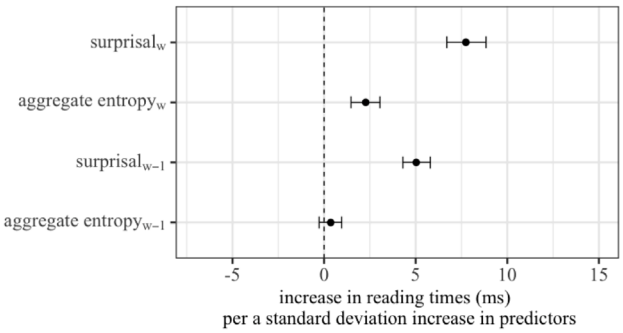
More specifically, positive entropy effects (slowdowns caused by increase in attention entropy) are found on the current target words for self-paced reading times and go-past times in eye-tracking data. In first-fixation durations in eye-tracking data, we find positive entropy effects of the previous word.

There are also positive interaction effects between surprisal and attention entropy, indicating an over-additive effect. This suggests that the attention entropy effect is increased for more unexpected words.

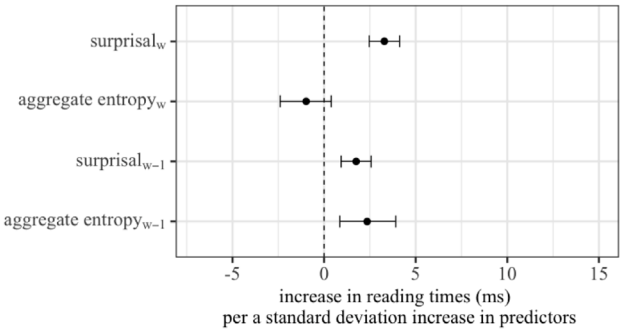
Provisional theoretical interpretation

The four effects of interest here are (i) the effect of surprisal of the current word, (ii) the effect of surprisal of the previous word, (iii) the positive effect of aggregate attention entropy of the current word in self-paced reading times and go-past time in eye-tracking data, or of the previous word with first fixation duration in eye-tracking data, and (iv) the positive interaction between surprisal and attention entropy (even though it was not found in go-past times). We now briefly sketch a theoretical interpretation of these effects.

(a) *Coefficient estimates from Natural Stories corpus.*



(b) *Coefficient estimates from GECO corpus (first fixation duration).*



(c) *Coefficient estimates from GECO corpus (go-past time).*

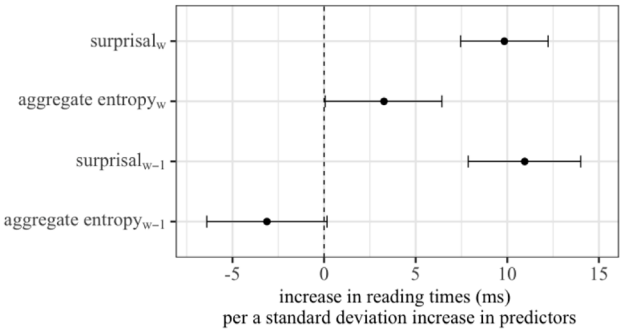


Fig. 10. Results from Bayesian regression models on predicting reading times. The distribution of posterior estimates of the coefficients of standardized predictors from self-paced reading time and eye tracking data. The lines indicate 95% credible intervals, and the dots indicate means of estimates.

Probabilistic expectations of upcoming words speed up lexical identification and access to the degree that the word is expected (Ehrlich & Rayner, 1981; Rayner & Well, 1996; Schustack et al., 1987). In particular, assuming noisy sequential sampling of visual evidence at some fixed rate, an optimal Bayesian perceiver will require a number of samples (and thus time) that is a linear function of surprisal (Norris, 2009). This effect on word identification would be expected to show up in earlier reading time measures in most accounts of eye movement control, such as EZ-Reader (Reichle et al., 2003), Fig. 11. Under the interpretation that surprisal is also a measure of the cost of probabilistic belief update (Levy, 2008) (which itself might be post memory retrieval), surprisal might manifest as a spillover effect *in addition* to a separate early lexical identification effect.

An eye movement control system that is optimized for time efficiency would also aggressively schedule saccades to the next words in a manner that would make later stage, higher level linguistic dependency formation effects spillover to upcoming words, as in EZ-Reader or SWIFT (Engbert et al., 2005). As we take aggregate attention entropy as an indicator of interference during memory retrieval, we hypothesize that the entropy effects would extend into spillover regions. This would be especially true if the analyzed reading times index early stages of sentence processing or if participants are motivated to read quickly (as they surely are to some degree).

The pattern of attention entropy effects we find in the eye-tracking data aligns with our interpretation that attention entropy reflects the sentence processing difficulty associated with later-stage memory retrieval. Specifically, when considering go-past time, which includes the regression to the previous words presumably required for dependency formation, the entropy effects are observed at the target region, suggesting entropy's role in later memory retrieval. In contrast, when the model is fit to first fixation, the entropy effects appear as spillover indicating that later memory retrieval difficulty carries over to subsequent regions. Entropy effects are also observed at the target region in self-paced reading data. Even though the self-paced reading time data are less straightforward to interpret as that effects from self-paced reading data tend to be distributed across multiple regions (Boyce & Levy, 2023; Witzel et al., 2012), we see the results suggest self-paced reading times indexes the later memory integration stage to some extent. To test the role of attention entropy in explaining sentence processing as a memory-integration-index further, future work is needed to explore how varying levels of reading speed and task motivation influence the locus of aggregate attention entropy effects.

The positive interaction between surprisal and attention entropy aligns with prior research showing that memory retrieval effects are attenuated when expectations are strong (i.e., when surprisal is low) (Campanelli et al., 2018; Husain et al., 2014; Tung & Brennan, 2023). For instance, Campanelli et al. (2018) demonstrated that when a verb at the retrieval site was highly predictable, memory interference effects on reading time were significantly reduced.

Summary and discussion

We have demonstrated that Transformers may serve as useful new foundations for integrative models of human sentence processing that combine elements of expectation-based and memory-based theories, specifically cue-based retrieval models. We presented three kinds of evidence for this view. First, we showed that the computational architecture of the Transformer, with its key-query-value dot-product attention mechanism, embodies the key assumptions of cue-based retrieval models. This is perhaps not surprising, because they share an underlying *functional* motivation: provide a memory retrieval mechanism that supports the computation of long-distance linguistic dependencies. For artificial intelligence, this addresses the problem of decaying memories in recurrent neural nets (Vaswani et al., 2017), and for psycholinguistics, it provides a way to explain both the astonishing functional capacities of human processing (via parallel cue-based access) and its apparently sharp limits (via similarity-based interference) (Lewis, 1998; Lewis & Vasishth, 2005).

Second, we showed that the internal attention patterns of a pre-trained Transformer, GPT2, show the signatures of similarity-based interference expected under a cue-based retrieval account. An attention-entropy metric provides one simple way to quantify the degree of retrieval interference, and we showed that attention entropy accounts for difficulty contrasts between right-branching and center-embedded structures, and predicts increased difficulty at the embedded verb of object relative clauses, an empirical effect that surprisal alone does not capture.

Third, we used word-by-word attention entropy averaged over twenty attention heads in GPT2 (selected for the correlation of their attention patterns with intra-sentential grammatical dependencies as described in Appendix A) to predict reading times in the Natural Stories

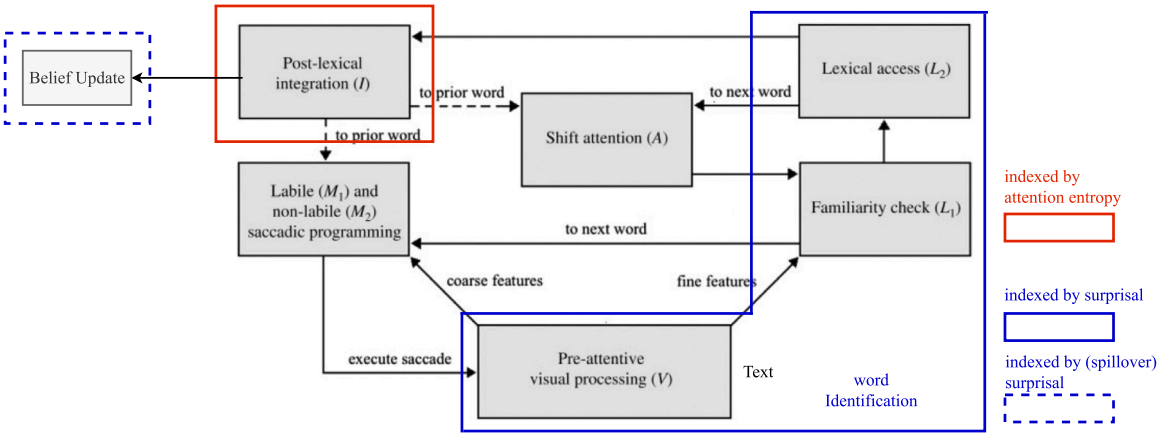


Fig. 11. Representation of influences from surprisal and attention entropy being integrated on the E-Z Reader model of eye-movement control. The original diagram is from Mancheva et al. (2015).

self-paced reading corpus (Futrell et al., 2021) and in the GECO eye-tracking corpus (Cop et al., 2017). An effect of attention entropy emerged in different loci depending on the type of reading times in a way that it is consistent with the interpretation of attention entropy as an index of a post-lexical memory integration stage.

Pre-trained Transformers address many of the methodological and theoretical challenges facing the use of cue-based retrieval theories in sentence processing research. The representations and cues are learned, so the theorist does not need to posit features (Smith & Vasishth, 2020). They provide a concrete mechanism for realizing multiple parallel memory retrievals. The models are powerful as language models, and so they address, at least implicitly, the need for probabilistic belief update over possibly multiple interpretations, and they address the concern that cue-based parsing cannot be scaled up to achieve human-level functionality.

Furthermore, several factors make Transformers promising components of widely applicable and practical modeling and analysis tools: the increasing availability of powerful pre-trained models and software for training new models, the relative transparency of internal word-by-word attention patterns, and the easy of computing attention entropy and surprisal.

What diffuse attention reflects, and revisiting the entropy reduction hypothesis

As we noted earlier, the use of entropy over attention values to create metrics intended to capture retrieval interference does not commit to an interpretation of attention entropy as a measure of uncertainty about what to attend to, or equivalently, attention values as representing a probability distribution over attention targets. It seems unlikely that all Transformer attention patterns that target more than one previous token or position indicate *confusion* about what the correct target is. Attentional spread over multiple targets could be functional—gathering information from multiple parts of prior context to inform the incremental interpretation of current input. Labeling such patterns as *interference* does not seem apt because it implies that more sharply focused attention (reducing interference) would lead to better performance. This might indeed be the case for some attention functions—perhaps those that subserve grammatical dependency formation would indeed perform better if somehow the attention did not spread to non-dependents. It seems likely to us that both kinds of attention spread arise in Transformers. It is also possible that all kinds of attentional spread, whether functional or not, are associated with processing slow downs in humans.

In cases of ambiguity, attentional spread might be a functional mechanism by which the Transformer is hedging its bets about what the correct interpretation is, i.e. it might be a mechanism that allows the model to perform something like a beam search, maintaining multiple interpretations in parallel (Gibson, 1991; Gorrell, 1987; Hale et al., 2018; Jurafsky, 1996; Levy, 2008). In that case, attention entropy would indeed be related to the entropy of the Entropy Reduction Hypothesis (Hale, 2006). But the relationship would be complex, and whatever that relationship, it is not clear that the finding of an attention-entropy slow down on the *current word* would be consistent with it. Again, even in this case, though the attentional spread might be functional, it may still be associated with processing slowdowns in humans.

In short, there are many possibilities why attention might be more or less diffuse, and these possibilities interact with different assumptions about what might lead to slow-downs. Teasing apart these possibilities will require the interventionist tools of mechanistic interpretability (Elhage et al., 2021; Nanda & Bloom, 2022).

Limitations and looking ahead: Empirical challenges, mechanistic interpretability and dynamic mechanism models

There are clear limitations to the work we presented here. One key limitation is that we have not yet shown that a single model with fixed parameters can account for quantitative reading time contrasts in both natural corpora (such as Natural Stories and GECO) *and* experimental materials. This remains an outstanding challenge for the field more broadly (Huang et al., 2024) and is not a problem that is unique to the entropy approach we advance here.

Another clear limitation is that our analyses are not grounded in a clear understanding of the *causal functional* roles of various attention heads. Our method of attention head selection is intended to mitigate this concern to some degree, to the extent that it makes it more likely that the selected heads are in fact functionally important and that they are computing something linguistically interesting. But developing causal accounts of Transformer representations and attention head functions will require applying the new causal interventionist tools of mechanistic interpretability and representation visualization—manipulating attention patterns, patching in different activations in controlled experiments, and using methods to uncover human-interpretable features in representations. Such tools are developing rapidly and this is yet another area where psycholinguistics stands to benefit from the enormous investment in such tools both inside and outside academia.

Another possible limitation concerns the use of unnormalized attention weights in the analysis. Prior research has raised concerns about relying on these weights (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019): an attention head may assign a high weight to a token, but if the

corresponding value vector has a low norm, that token may have minimal influence on the final output. Thus, high attention weights do not necessarily indicate the importance of the attended token in sentence processing. To address this, Kobayashi et al. (2020) proposed using norm-adjusted attention weights (see also Oh and Schuler (2022)). This adjustment accounts for the interaction between attention scores and value vector magnitudes, potentially providing a more accurate reflection of the information actually utilized by the model. We plan to adopt this normalization in future analyses, but note that the head-selection method used here mitigates the concern to some degree, because it makes it more likely that the attention patterns that enter analysis are implicated in dependency formation.

Finally, another clear limitation is that while the current analyses use attention entropy to predict reading time, we do not have an explicit theory or computational model of the *dynamics* of memory access. Transformers do the same computation at each word in a single feed-forward pass—they do not offer a mechanistic account of memory dynamics. Using attention entropy in this way is like using surprisal without a mechanistic account of why surprisal varies linearly with reading time.

One possibility for adding dynamics to a Transformer is to incorporate noise along with an internal sequential evidence accumulator, so that, for example, multiple memory accesses are integrated over time until some threshold of quality is reached (Shadlen & Shohamy, 2016). That threshold may be entropy related. This would be akin to incorporating sequential sampling of noisy perceptual evidence of word identity until some belief threshold is met (Norris, 2006; Wald & Wolfowitz, 1948), which would provide a mechanistic account of the linear surprisal effect. In this way a Transformer could serve as the basis of a bounded rational comprehender, bounded by perceptual and memory noise, but rationally adapted to that noise and speed–accuracy tradeoffs imposed by current task demands (Lewis et al., 2014).

More comprehensive mechanistic and dynamic models could be developed by including encoding and storage noise or interference, as was adopted in the lossy-context surprisal (Futrell et al., 2020) model, and the recent transient binding model (Keshev et al., 2024) of sentence processing. Finally, there is a much richer space now of neural net sequence modelers that include not only Transformer variants but new recurrent architectures (e.g., Sun et al., 2024).

Why is there interference in sentence processing?

Despite the clear limitations of the current work, we believe Transformers are promising foundations for richer models of human sentence processing and that addressing the challenges outlined above is both worthy of effort and increasingly tractable. We would like to close on a positive note and point out that, even with our current limited understanding, Transformer models already provide the outlines of a deep answer to a fundamental question: Why is human sentence processing subject to sharp processing limitations? Give the bounds that similarity-based interference seems to impose, it appears at first to be a kind of system defect.

The answer is that such interference naturally arises in systems that are under great pressure to develop representations that *generalize*, because the only way for physical computing systems to generalize is to learn and use representations in a similarity space (Shepard, 1987). For systems that have any amount of representational noise, such useful representations come at a cost of limited processing capacity (Frankland et al., 2021). The sharp limits we observe in humans with center-embeddings and the more subtle effects of interference evident in reading times reflect this fundamental tradeoff. We can now see reflections of such interference in the learned processes and representations of artificial language models built with purely functional aims in mind. In short, these limits are not signatures of design flaws; on the contrary, they are signatures of optimization.

CRediT authorship contribution statement

Soo Hyun Ryu: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Richard L. Lewis:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Selection of attention heads for the aggregate entropy metric

Voita et al. (2019) presents a method to identify Transformer's attention heads that may be computing specific grammatical dependencies. This method works by comparing attention patterns to the dependencies uncovered by a dependency parser. More specifically, the method compares the proportion of times an attention head pays the highest attention to a grammatical dependent of a given type with the proportion of the most frequent relative positions observed between dependents of a given type. This comparison ensures that the identified heads assign greater attention to a grammatical dependent as the position varies in a way that is not simply due to positional information. We used this approach (1) to identify attention heads that may reflect nominal subject and verb dependency as used in *Right vs. center embeddings* and (2) to select attention heads for our aggregate attention entropy metric. Here, we provide a detailed information of attention head identification.

1. **Step 1. Counting relative positions of grammatical dependents**
   To determine the proportion of each grammatical dependency that can be detected based on dependents' relative positions, we first counted the relative positions of all instances in each grammatical dependency. We used the CoreNLP dependency parser (Manning et al., 2014) to annotate 148,376 sentences from the Brown Corpus and the Gutenberg corpus, provided via the Natural Language Toolkit (NLTK; Bird et al. (2009)). For each grammatical dependency, we then recorded the locations of grammatical dependents relative to their governors' positions in all instances. In the case of the nominal subject and verb dependency, the most common relative position is -1, indicating that subjects appear right before verbs, and this accounts for 41% of all instances.
2. **Step 2. Examine attention heads' accuracies in relating grammatical dependents**
   To determine which attention heads capture certain dependency better than simple prediction based on relative position, we compute the percentage of instances for each attention head where the largest attention is paid to the corresponding grammatical dependents from their governors, or to the governors from the dependents if the governor precedes the dependents. In the case of the nominal subject and verb (*nsubj*) dependency, the highest percentage was found in $\text{head}_{4,3}$ (the third head in the fourth layer of GPT2), where the largest attention is paid to the subject from the verb in 59% of nsubj instances.
3. **Step 3. Determination of attention heads**

Just as Voita et al. (2019) did, we considered attention heads to be capable of detecting a certain grammatical dependency if their accuracy in allocating the largest attention to grammatically corresponding dependents is at least 10% higher than the proportion of instances that can be explained by the most frequent relative positions. For example, for a head to be selected for the *nsubj* dependency, it must pay the highest attention to the subject in at least 41+4.1 = 45.1% of the *nsubj* dependencies.

The results of the analysis are provided in Table 6.

Table 6
Results from the head selection analysis.

Dependency type	Frequency-based accuracy	Attention-based accuracy
acl	2, 40.64%	**(2,9), 70.18%**
		(3,7), 53.37%
acl:relcl	2, 29.65%	**(6,7), 42.74%**
		(6,1), 36.64%
		(7,8), 35.64%
		(4,6), 35.14%
		(2,9), 35.10%
		(4,3), 34.87%
		(3,7), 34.06%
advcl	2, 10.76%	**(3,5), 20.42%**
		(0,0), 17.35%
		(3,4), 17.31%
		(4,9), 16.78%
		(6,3), 16.64%
		(4,6), 14.81%
		(5,9), 14.51%
		(3,9), 14.43%
		(0,6), 14.43%
		(4,8), 14.32%
		(1,5), 12.32%
advmod	−1, 42.55%	**(4,11), 53.81%**
		(3,11), 52.77%
		(6,8), 50.15%
		(5,6), 49.52%
amod	−1, 78.72%	None
appos	2, 31.05%	None
aux	−1, 52.39%	**(3,8), 74.04%**
		(3,9), 69.71%
		(1,0), 62.43%
		(2,2), 61.04%
		(2,4), 59.95%
		(2,8), 58.07%
aux:pass	−1, 88.00%	**(2,8), 97.09%**
case	−2, 41.50%	**(2,0), 85.98%**
		(3,8), 82.54%
		(4,0), 75.53%
		(2,8), 73.23%
		(5,4), 68.10%
		(2,5), 53.73%
		(2,2), 52.20%
		(1,2), 51.92%
		(8,7), 50.79%
		(3,3), 46.97%
		(6,8), 46.44%
		(6,0), 46.35%
		(7,4), 46.23%
		(6,11), 46.02%
cc	−1, 42.74%	**(3,8), 65.92%**
		(4,1), 65.40%
		(2,5), 61.12%
		(2,4), 56.28%
		(2,0), 55.94%
		(1,1), 55.28%
		(2,6), 49.83%
cc:preconj	−1, 37.40%	**(10,5), 74.05%**
		(4,1), 73.28%
		(8,0), 63.36%
		(9,10), 62.60%
		(3,11), 53.44%
		(6,5), 53.44%
		(3,1), 48.09%

(*continued on next page*)

Table 6 (*continued*).

Dependency type	Frequency-based accuracy	Attention-based accuracy
		(6,11), 44.27%
		(8,2), 41.22%
ccomp	3, 17.53%	**(4,2), 29.65%**
		(5,3), 27.24%
		(4,1), 22.63%
		(4,6), 21.69%
		(6,2), 21.38%
		(0,0), 20.93%
		(1,5), 19.63%
compound	−1, 86.22%	None
compound:prt	1, 84.65%	**(1,0), 95.27%**
		(2,4), 94.98%
		(2,8), 94.59%
		(3,9), 93.34%
conj	2, 30.44%	None
cop	−1, 36.78%	**(2,8), 82.98%**
		(3,9), 73.14%
		(3,8), 63.91%
		(4,0), 61.25%
		(2,5), 59.12%
		(1,1), 50.36%
		(6,0), 48.79%
		(5,4), 46.66%
		(8,7), 44.62%
		(7,0), 42.75%
csubj		**(3,2), 40.72%**
det	−1, 63.22%	**(2,3), 81.92%**
		(3,2), 80.18%
det:predet	−2, 70.66%	None
discourse	−3, 19.53%	None
expl	−1, 65.51%	None
fixed	1, 89.55%	None
flat	1, 92.09%	None
goeswith	−1, 100%	None
iobj	1, 88.08%	**(4,0), 97.93%**
		(5,4), 97.93%
list	2, 30.00%	**(0,6), 41.67%**
mark	−1, 47.33%	**(2,0), 63.09%**
		(3,8), 62.42%
nmod	3, 38.20%	None
nmod:npmod	−1, 73.15%	None
nmod:poss	−1, 62.55%	**(3,6), 75.08%**
		(2,3), 73.95%
		(3,2), 70.36%
nmod:tmod	2, 30.00%	**(7,8), 50.00%**
		(9,3), 50.00%
		(6,0), 47.50%
		(10,9), 47.50%
		(2,9), 47.50%
		(9,0), 47.50%
		(11,11), 45.00%
		(8,7), 45.00%
		(11,8), 45.00%

(*continued on next page*)

Table 6 (*continued*).

Dependency type	Frequency-based accuracy	Attention-based accuracy
		(6,7), 45.00%
		(6,1), 42.50%
		(11,3), 42.50%
		(7,0), 42.50%
		(4,3), 42.50%
		(9,10), 42.50%
		(5,7), 40.00%
		(8,5), 40.00%
		(3,7), 40.00%
		(5,4), 37.50%
		(3,5), 37.50%
		(11,10), 37.50%
		(0,6), 37.50%
		(1,7), 35.00%
		(8,4), 35.00%
		(4,6), 35.00%
nsubj	−1, 40.63%	**(4,3), 56.85%**
		(6,0), 47.13%
		(3,6), 46.30%
		(2,9), 46.27%
nsubj:pass	−2, 39.59%	**(4,3), 67.63%**
		(3,7), 53.90%
		(2,9), 52.12%
nummod	−1, 54.27%	**(3,6), 71.54%**
		(1,0), 66.04%
		(5,6), 65.88%
		(4,11), 63.30%
		(6,8), 61.46%
		(0,7), 60.47%
obj	2, 37.98%	**(2,8), 84.15%**
		(4,0), 81.57%
		(3,9), 76.11%
		(5,4), 73.74%
		(3,8), 69.99%
		(6,11), 59.23%
		(5,7), 55.01%
		(8,7), 54.01%
		(9,10), 53.31%
		(8,4), 51.21%
		(11,11), 50.38%
		(10,5), 47.59%
		(3,11), 47.19%
		(3,3), 45.26%
		(7,4), 44.93%
		(11,10), 44.80%
		(7,8), 44.55%
		(2,9), 44.44%
		(4,6), 43.44%
		(10,11), 45.26%
obl	3, 24.24%	**(4,9), 35.17%**
		(3,9), 28.98%
obl:npmod	−1, 62.43	**(2,9), 82.01%**
		(5,6), 70.90%
		(2,4), 29.58%
obl:tmod	2, 16.07%	**(4,9), 54.29%**
		(6,11), 39.64%
		(7,8), 36.79%
		(3,5), 35.71%
		(2,8), 35.36%
		(5,7), 35.18%
		(9,10), 33.57%
		(11,11), 33.39%
		(5,3), 30.89%
		(10,9), 30.36%
		(3,9), 30.18%
		(11,10), 28.93%

(*continued on next page*)

Table 6 (*continued*).

Dependency type	Frequency-based accuracy	Attention-based accuracy
		(7,4), 28.57%
		(8,4), 28.39%
		(4,6), 27.68%
		(6,3), 26.43%
		(10,5), 26.43%
		(4,3), 25.89%
		(3,8), 25.18%
		(7,5), 25.00%
		(3,7), 24.46%
		(7,3), 24.46%
		(0,2), 24.11%
		(8,5), 23.75%
		(2,9), 22.68%
		(6,7), 22.50%
		(8,7), 22.50%
		(3,11), 22.14%
		(10,11), 21.61%
		(0,0), 21.25%
		(11,8), 20.89%
		(7,0), 20.71%
		(7,9), 20.71%
		(0,6), 20.54%
		(11,3), 20.00%
		(11,7), 19.29%
		(1,7), 18.57%
		(6,8), 18.21%
		(9,3), 18.21%
		(4,0), 18.04%
parataxis	3, 13.64%	**(3,5), 20.00%**
		(0,0), 17.70%
		(3,4), 16.83%
		(5,9), 15.86%
		(0,6), 15.86%
punctuation	−1, 16.57%	**(1,1), 37.02%**
		(2,6), 36.05%
		(2,5), 35.95%
		(9,11), 35.58%
		(10,2), 33.02%
		(10,2), 31.84%
		(10,4), 31.03%
		(4,11), 30.29%
		(11,2), 29.91%
		(1,3), 29.33%
		(3,7), 28.99%
		(7,11), 28.42%
		(1,7), 28.12%
		(9,3), 27.19%
		(3,3), 26.42%
		(7,0), 25.42%
		(11,4), 25.34%
		(5,6), 24.68%
		(3,1), 24.58%
		(11,11), 24.58%
		(4,1), 24.33%
		(5,4), 24.31%
		(10,10), 24.24%
		(5,3), 24.02%
		(8,5), 23.68%
		(6,1), 23.55%
		(8,6), 23.30%
		(8,2), 23.20%
		(11,0), 23.13%
		(6,8), 22.31%
		(3,8), 22.25%
		(11,3), 22.02%
		(4,6), 21.68%
		(4,2), 21.65%
		(8,3), 21.48%
		(8,0), 21.19%
		(5,2), 20.94%
		(1,2), 20.92%

(*continued on next page*)

Table 6 (*continued*).

Dependency type	Frequency-based accuracy	Attention-based accuracy
		(7,9), 20.77% (2,9), 20.73% (11,10), 20.68% (11,9), 20.48% (10,7), 20.47% (8,11), 20.32% (9,6), 20.22% (6,7), 19.95% (5,8), 19.24% (8,7), 19.10% (7,8), 19.08% (9,9), 18.97% (3,2), 18.80% (6,5), 18.79% (1,8), 18.54% (7,7), 18.27%
root	0, 100.00%	None
vocative	2, 26.32%	None
xcomp	2, 54.90%	**(3,9), 75.04%**

A.1. Attention heads identified for nominal subject dependency

Heads selected for the *nsubj* dependencies were $head_{4,3}$, $head_{6,0}$, $head_{3,6}$, $head_{2,9}$, ordered by accuracy. In *Right vs. center embeddings*, our analysis is based on $head_{4,3}$, whose accuracy is the highest.

A.2. Selection of attention heads for the aggregate attention entropy computation

Among the 43 dependency types supported by the CoreNLP package, 14 types failed to be associated with the patterns of any attention heads. There are twenty attention heads that obtain the highest scores in one or more of the remaining 29 dependency types, and we included them for the aggregate attention computation as described in *Transformers as Cue-based Retrieval Sentence Processors*.

Appendix B. The predictive power of attention entropy as a function of the window size

Gao and Yu (In prep) have shown that surprisal becomes more predictive of human reading times when a larger context window (around 1,000 tokens) is used. This should not be surprising, as the larger context likely allows readers to draw on broader contextual information to form expectations about upcoming words.

If the influence of attention entropy on reading times reflects memory integration, it is not clear what the optimal window will be for computing entropy. If attention entropy effects are largely an index of intra-sentential dependency formation, much smaller windows might be best for computing attention entropy. We explored this question empirically.

B.1. Methods

To assess the predictive power of aggregate attention entropy calculated with different window sizes, we compared the coefficients of each predictor from frequentist mixed-effects regression models. These models were fit with attention entropies of varying context window sizes, along with other predictors like word length, word position, and word frequency. The token representations were derived using the maximum context window size, and attention values for the most recent N tokens were included for attention entropy with window size N.

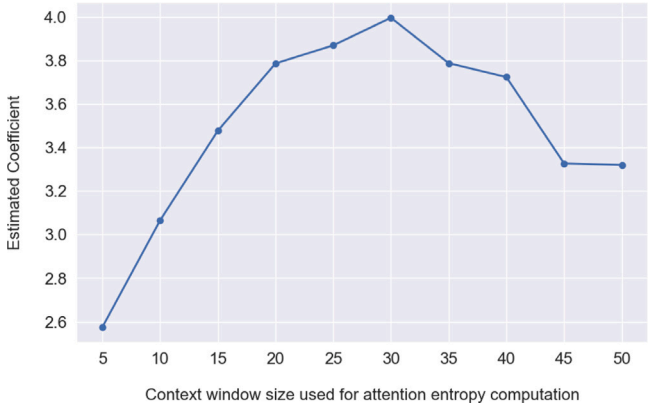


Fig. 12. Change in coefficients of attention entropies by the different context window size used for attention entropy computation.

B.2. Results

As shown in Fig. 12, attention entropy has the highest coefficient at a window size of 30. This suggests that attention entropy benefits from a limited window size compared to the optimal context window for surprisal. Given this, we use the aggregate attention entropy with the context window size of 30 for our analysis in Section *Modeling Sentence Reading Times with Surprisal and Attention Entropy*.

Data availability

Data files and analysis codes for the present study and the following studies in the paper are available on GitHub: github.com/soohyunryu/memory-for-prediction/.

References

3Blue1Brown (Director) (2024). But what is a GPT? Visual intro to transformers | Chapter 5. *Deep Learning*.
Baker, C. L. (1989). *English syntax*. MIT Press.
Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
Boland, J., Lewis, R., & Blodgett, A. (2001). Distinguishing generation and selection of modifier attachments: Implications for lexicalized parsing and competition models. In *14th Annual CUNY Conference on Human Sentence Processing, Philadelphia, PA*. Citeseer.
Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.
Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*, *2*(1).
Brennan, J. R., Dyer, C., Kuncoro, A., & Hale, J. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, *146*, Article 107479. http://dx.doi.org/10.1016/j.neuropsychologia.2020.107479.
Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.
Campanelli, L., Van Dyke, J. A., & Marton, K. (2018). The modulatory effect of expectations on memory retrieval during sentence comprehension. In *Proceedings of the annual meeting of the cognitive science society, vol. 40*.
Caplan, D., & Waters, G. S. (1990). Short-term memory and language comprehension: A critical review of the neuropsychological literature. In G. Vallar, & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 33–89). Cambridge University Press.
Carpenter, P. A., & Just, M. A. (1989). The role of working memory in language comprehension. In D. Klahr, & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert a. Simon*. Erlbaum.
Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
Chomsky, Noam (1970). In R. Jacobs, & P. Rosenbaum (Eds.), *Remarks on nominalization*. Ginn, Readings in English Transformational Grammar.
Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In D. R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol. II* (pp. 269–321). John Wiley.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 126–161.
Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615.
Demberg, V., & Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the annual meeting of the cognitive science society, vol. 31*.
Dotlačil, J. (2021). Parsing as a cue-based retrieval model. *Cognitive Science*, *45*(8), Article e13020.
Ehrlich, S. F., & 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., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, *1*(1), 12.
Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813. http://dx.doi.org/10.1037/0033-295X.112.4.777.
Frankland, S. M., Webb, T., & Cohen, J. D. (2021). No coincidence, george: Capacity-limits as the curse of compositionality.
Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.
Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.
Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), Article e12814.
Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories Corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*(1), 63–77.
Gao, R., & Yu, C.-L. (In prep). Humans consider extensive prior contexts during natural reading: An eye-tracking examination with the GECO dataset. (in preparation).
Gibson, E. A. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown* (Ph.D. thesis), Carnegie Mellon University.
Gibson, E. A. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.
Gibson, E. A. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Mirantz, & W. O'Neil (Eds.), *Image, language, brain*. MIT Press.
Gibson, E. A., & Thomas, J. (1996). The processing complexity of english center-embedded and self-embedded structures. In C. Schutze (Ed.), *Proceedings of the NELS 26 workshop on language processing*.
Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(6), 141–423.
Gorrell, P. (1987). *Studies of human syntactic processing: Ranked-parallel versus serial models*. The University of Connecticut.
Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.
Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), Article e2122602119. http://dx.doi.org/10.1073/pnas.2122602119.
Hakes, D. T., Evans, J. S., & Brannon, L. L. (1976). Understanding sentences with relative clauses. *Memory and Cognition*, *4*(3), 283–290.
Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Second meeting of the North American chapter of the association for computational linguistics on language technologies 2001* (pp. 1–8). http://dx.doi.org/10.3115/1073336.1073357.
Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. http://dx.doi.org/10.1207/s15516709cog0000_64.
Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.
Hale, J., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, *8*, 427–446. http://dx.doi.org/10.1146/annurev-linguistics-051421-020803.
Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. R. (2018). Finding syntax in human encephalography with beam search. http://dx.doi.org/10.48550/arXiv.1806.04127.
Häussler, J., & Bader, M. (2015). An interference account of the missing-VP effect. *Frontiers in Psychology*, *6*, 766.
Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.
Hoover, B., Strobelt, H., & Gehrmann, S. (2019). Exbert: A visual analysis tool to explore learned representations in transformers models. arXiv preprint arXiv:1910.05276.
Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, Article 104510. http://dx.doi.org/10.1016/j.jml.2024.104510.
Huang, N., & Phillips, C. (2021). When missing NPs make double center-embedding sentences acceptable. *Glossa: A Journal of General Linguistics*, *6*(1), http://dx.doi.org/10.5334/gjgl.1292.
Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from hindi. *PLoS One*, *9*(7), Article e100986.
Jain, S., & Wallace, B. C. (2019). Attention is not explanation. arXiv preprint arXiv:1902.10186.
Joshi, A. K., Levy, L. S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*, *10*(1), 136–163.
Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.
Karpathy, A. (2024). nanoGPT, [Computer software]. (Original work published 2022).
Keshev, M., Cartner, M., Meltzer-Asscher, A., & Dillon, B. (2024). A working memory model of sentence processing as binding morphemes to syntactic positions. *Topics in Cognitive Science*.
Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7057–7075). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.574.
Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. https://arxiv.org/abs/2205.11463.
L., Frazier (1987). Sentence processing: A tutorial review.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. http://dx.doi.org/10.1016/j.cognition.2007.05.006.
Levy, R. (2013). Memory and surprisal in human sentence comprehension.
Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, *69*(4), 461–495.
Levy, R., & Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*, 229.
Lewis, R. L. (1993). An architecturally-based theory of sentence comprehension. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 108–113).
Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–115.
Lewis, R. L. (1998). Working memory in sentence processing: Retroactive and proactive interference in parsing. In *The 11th annual CUNY conference on human sentence processing*.
Lewis, R. L. (2000). Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. In M. W. Crocker, M. Pickering, & C. Clifton, Jr. (Eds.), *Architectures and mechanisms for language processing*. Cambridge University Press.
Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.
Lewis, R. L., Shvartsman, M., & Singh, S. (2013). The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs. *Topics in Cognitive Science*, *5*(3), 583–610.
Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.
Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.
Lissón, P., Paape, D., Pregla, D., Burchert, F., Stadie, N., & Vasishth, S. (2023). Similarity-based interference in sentence comprehension in Aphasia: A computational evaluation of two models of cue-based retrieval. *Computational Brain & Behavior*, *6*(3), 473–502. http://dx.doi.org/10.1007/s42113-023-00168-3.
Lissón, P., Pregla, D., Nicenboim, B., Paape, D., van het Nederend, M. L., Burchert, F., Stadie, N., Caplan, D., & Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in Aphasia. *Cognitive Science*, *45*(4), Article e12956. http://dx.doi.org/10.1111/cogs.12956.
Mancheva, L., Reichle, E. D., Lemaire, B., Valdois, S., Ecalle, J., & Guérin-Dugué, A. (2015). An analysis of reading skill development using EZ Reader. *Journal of Cognitive Psychology*, *27*(5), 657–676.
Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).
Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. MIT Press.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., & Brockman, W. (2011). In J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, & E. L. Aiden (Eds.), *The google books team* (pp. 176–182).
Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In D. R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol. II*. John Wiley.
Miller, G. A., & Isard, S. (1964). Free recall of self-embedded english sentences. *Information and Control*, *7*, 292–303.
Nanda, N., & Bloom, J. (2022). Transformerlens. URL: https://Github.Com/Neelnanda-Io/TransformerLens.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357. http://dx.doi.org/10.1037/0033-295X.113.2.327.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*(1), 207–219. http://dx.doi.org/10.1037/a0014259.

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

Oh, B.-D., & Schuler, W. (2022). Entropy-and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9324—9334).

Oh, B.-D., & Schuler, W. (2023a). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. arXiv preprint arXiv:2304.11389.

Oh, B.-D., & 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.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–+.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 61–71).

Ryu, S. H., & Lewis, R. L. (2022). Using transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th annual conference on human sentence processing*.

Schustack, M. W., Ehrlich, S. F., & Rayner, K. (1987). Local and global sources of contextual facilitation in reading. *Journal of Memory and Language*, *26*(3), 322–340.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. http://dx.doi.org/10.1016/j.neuron.2016.04.036.

Shain, C., Van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the workshop on computational linguistics for linguistic complexity* (pp. 49–58).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shiffrin, R. (2003). Modeling memory and perception. *Cognitive Science*, *27*(3), 341–378. http://dx.doi.org/10.1016/S0364-0213(03)00027-2.

Smith, G., & Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cognitive Science*, *44*(12), Article e12918.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.

Stolz, W. S. (1967). A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 867–873.

Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., et al. (2024). Learning to (learn at test time): Rnns with expressive hidden states. arXiv preprint arXiv:2407.04620.

Timkey, W., & Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 8705–8720). Association for Computational Linguistics., http://dx.doi.org/10.18653/v1/2023.findings-emnlp.582.

Tung, T.-Y., & Brennan, J. R. (2023). Expectations modulate retrieval interference during ellipsis resolution. *Neuropsychologia*, *190*, Article 108680.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A retrieval interference theory of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), Article e12988.

Vani, P., Wilcox, E. G., & Levy, R. (2021). Using the interpolated maze task to assess incremental processing in english relative clauses. In *Proceedings of the annual meeting of the cognitive science society, vol. 43*.

Vasishth, S., & Engelmann, F. (2021). *Sentence comprehension as a cognitive process: A computational approach*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems, vol. 30*.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th annual meeting of the association for computational linguistics: System demonstrations* (pp. 37–42).

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 63–76).

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5797–5808).

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*(3), 326–339.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. arXiv preprint arXiv:1908.04626.

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*, 105–128.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate Bayesian computation. *Open Mind*, *6*, 1–24. http://dx.doi.org/10.1162/opmi_a_00052.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466.

Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys*, *55*(5), 1–31.