

STRUCTURALCOST: A controlled reading time dataset for modeling human sentence processing difficulty

Nina Nusbaumer¹, Iria de-Dios-Flores^{2, 4}, Corentin Bel^{3, 4},
Christophe Pallier^{3, 5}, Guillaume Wisniewski¹, Benoît Crabbé¹

¹LLF, CNRS, Université Paris Cité, ²COLT, Universitat Pompeu Fabra,

³Unicog, Neurospin, CEA, ⁴LNC2, ENS-PSL, ⁵INSERM, CNRS

{nina.nusbaumer, guillaume.wisniewski, benoit.crabbe}@u-pariscite.fr, iria.dedios@upf.edu,
corentin.bel@ens.psl.eu, christophe@pallier.org

Abstract

We introduce STRUCTURALCOST, a self-paced reading dataset of 475 participants and 40,800 observations isolating the processing cost of long-distance subject-verb dependency resolution. We replicate a low-powered psycholinguistic finding at NLP scale, namely that human reading times at the main verb increase with dependency length, driven by syntactic embedding beyond linear distance. Different language models – spanning n -gram models, SSMs, and transformers – partially mirror this graded difficulty profile, yet underestimate the integration cost humans incur, with a gap that persists across architectures and model sizes. This suggests these models capture the predictive component of human processing but not the full integration cost that working memory imposes. STRUCTURALCOST provides data needed to drive progress toward evaluating the cognitive plausibility of language models.

1 Introduction

An actively debated question in computational psycholinguistics is whether language models (LMs) can serve as computational models of human sentence processing (Wilcox et al., 2020; Oh and Schuler, 2023); essentially, whether the mechanisms driving LMs’ text processing mirror those humans deploy during language comprehension. The question has practical stakes: a cognitively plausible LM could be used to test hypotheses at scale that are otherwise difficult or impossible to probe experimentally and to leverage model failures to sharpen sentence processing theories.

Language comprehension is incremental and constrained by two key components: prediction and working memory. As readers encounter each word, they continuously predict what comes next, and processing effort rises when those predictions fail (Hale, 2001; Levy, 2008). At the same time, comprehension relies on working memory, which

temporarily hold syntactic information until it can be integrated (Gibson, 2001; Lewis et al., 2006), e.g., the subject of a sentence remains active in memory until its main verb is later encountered.

LMs assign probabilities to words in context, which are a natural proxy for simulating the predictive component of human language processing. Numerous studies show that the surprisal of models, a simple log-transformation of these probabilities, correlates reliably with reading times – the primary behavioral proxy for online cognitive effort in psycholinguistics (e.g., Smith and Levy, 2013; Goodkind and Bicknell, 2018; Futrell et al., 2021). However, surprisal fit alone does not reveal whether LMs capture the memory-based component of language processing, one of its key signatures being the cost of resolving long dependencies: the further apart a syntactic head and its dependent are, the greater the cognitive effort required to resolve that dependency (Gibson, 2001; Lewis et al., 2006). This is a theoretically important distinction: LM architectures differ greatly in how they handle distance, from n -gram models with a hard local window to transformers with unconstrained self-attention. Thus, evaluating whether LMs replicate the graded effect of dependency length is a direct test of cognitive plausibility.

Measuring this graded difficulty profile is inherently challenging, as it requires isolating the impact of syntactic structure from confounding lexical factors, such as word frequency and length, which also influence processing times (Oh et al., 2024). Large naturalistic corpora provide the necessary scale for model evaluation but are often confounded by these lexical variables, making it difficult to separate the effects of dependency length from other influences. On the other hand, controlled psycholinguistic experiments are effective at isolating structural effects but typically lack the scale needed to evaluate models comprehensively. No existing dataset successfully combines both experimental control

and large-scale data. To fill this gap, we introduce STRUCTURALCOST, a manually curated dataset of 40,800 self-paced reading observations from 475 participants, including individual working memory scores.¹ STRUCTURALCOST is built following psycholinguistic standards applied at NLP scale: it holds the main subject-verb pair constant while systematically varying the distance between them, isolating the processing cost of dependency length from lexical confounds.

Using STRUCTURALCOST, we show that reading times at the main verb increase with dependency length – an effect driven by syntactic depth beyond linear distance alone – and that existing LMs only partially mirror this graded difficulty profile despite achieving strong global surprisal-RT fit. Models tend to underestimate the integration cost humans incur, consistently across architectures and model sizes, pointing to a fundamental limitation for treating them as plausible mechanistic models of sentence comprehension: current LMs seem to capture the predictive component of human processing but not the memory-based integration cost.

Our contributions are threefold: (1) we release STRUCTURALCOST, a large-scale controlled dataset that isolates dependency length effects on human reading times, replicating at large-scale previously low-powered psycholinguistic findings; STRUCTURALCOST is a controlled diagnostic tool, targeting a specific phenomenon, and not a general benchmark; this limited scope is a design choice, not a limitation. The use of established measures (surprisal-RT alignment, mixed models) is deliberate, so that the results remain directly comparable to existing psycholinguistic and NLP resources, the novelty being in the resource and its controlled manipulation. (2) In this controlled context, we observe that syntactic structure beyond mere linear distance modulates processing difficulty. (3) We observe that current LMs, across architectures and sizes, only partially replicate this graded difficulty profile despite strong global surprisal-RT fit, revealing a systematic blind spot in how cognitive plausibility is evaluated. We argue this motivates a shift toward controlled, memory-sensitive datasets as a criterion for cognitive plausibility.

¹STRUCTURALCOST is available at <https://doi.org/10.34847/nk1.7c467lyv>. Experiment: <https://github.com/NinaNusb/Self-paced-reading-task>. Analysis: <https://github.com/NinaNusb/lm-human-processing.git>.

2 Related Work

Toward cognitively plausible LMs. Because LMs assign probabilities to words in context, their surprisal estimates, defined as $-\log_2 p(\text{word}|\text{context})$, are a natural proxy for the predictive component of human language processing, and a growing body of work confirms they correlate with human reading times (Goodkind and Bicknell, 2018; Futrell et al., 2021). However, recent work finds an inverse scaling trend: larger models do not always yield better psychometric fit (Oh et al., 2022; Oh and Schuler, 2023; Shain et al., 2024), likely because state-of-the-art models have become *superhuman* at next-word prediction (Oh and Linzen, 2025), causing their surprisal to underestimate the difficulty humans experience.

A parallel line of work asks whether architectural constraints can bring models closer to human behavior. De Varda and Marelli (2024) and Timkey and Linzen (2023) show that inductive biases, beyond raw prediction accuracy, shape psychometric fit. Kuribayashi et al. (2022) provide a mechanistic explanation: transformer self-attention gives equal weight to all prior context, removing the distance-dependent decay that produces locality effects in human processing. They show that imposing context limitations on transformers partially restores these effects.

However, these constrained models have been evaluated primarily on reaction times from naturalistic datasets: Dundee (Kennedy and Pynte, 2005), Provo (Luke and Christianson, 2018), MECO (Siegelman et al., 2022), UCL (Frank et al., 2013), the SPR Brown corpus (Smith and Levy, 2013), and Natural Stories (Futrell et al., 2021)². While these corpora provide valuable insights into the global ability of LMs to fit general behavioral predictability, the uncontrolled variability in naturalistic data – where frequency, length, and syntactic structure are confounded – makes these datasets less well suited for modeling processing cost arising from structural complexity. Therefore, to be able to evaluate the cognitive plausibility of LMs, we must move beyond global fit on natural text and adopt controlled datasets that systematically manipulate structural and memory demands.

Limitations of existing datasets. Evaluating the cognitive plausibility of LMs involves testing their sensitivity to structural effects alongside their

²Except Timkey and Linzen, who studied disambiguation.

Condition	Dep.	Example Sentence
<i>Critical sentences</i>		
Baseline	0	The violinist leaves the stage before the audience loudly applauds the great performance.
PP	4	The violinist in the large orchestra leaves the stage before the audience loudly applauds.
SRC	4	The violinist that followed the conductor leaves the stage before the audience loudly applauds.
ORC	4	The violinist that the conductor followed leaves the stage before the audience loudly applauds.
2×SRC	9	The violinist that followed the conductor that worked in the orchestra leaves the stage.
2×ORC	9	The violinist that the conductor that the whole orchestra admired followed leaves the stage.
<i>Control sentences</i>		
Baseline	0	The waiter stacks the plates with the dirty apron next to the kitchen sink.
PP-short	4	The waiter with the dirty apron stacks the plates next to the kitchen sink.
PP-long	9	The waiter with the dirty apron next to the kitchen sink stacks the plates.

Table 1: Example sets of critical and control sentences. Dep. (dependency length) = number of intervening words between the **main subject** and **main verb**. *Critical sentences* cross syntactic structure (SRC/ORC) with embedding depth (single/double), plus a 0-word baseline and a 4-word linear-distance control (PP). *Control sentences* manipulate linear distance via prepositional phrases, with no syntactic embedding, providing a reference class for dissociating distance from structural complexity.

general ability to predict reading times. We identify four criteria that, we argue, datasets must satisfy to support such evaluation: **(C1) Isolation** – sentences must appear in isolation to remove the influence of prior context and focus solely on working memory limits; **(C2) Controlled variables** – the dataset must systematically vary dependency length while keeping vocabulary and lexical properties constant, ensuring that any increase in difficulty is due to memory load, not harder words; **(C3) Sufficient items** – a broad and varied set of unique sentences to provide robust model evaluation; **(C4) Sufficient participants** – enough participants per sentence to ensure reliable statistical estimates. Existing resources are insufficient for the evaluation of LMs’ potential sensitivity to structural complexity as neither meets all three criteria: NLP-oriented datasets offer large-scale naturalistic data but lack experimental control to isolate structural effects, while psycholinguistic datasets provide strong control but lack the items and participants needed for model evaluation. STRUCTURAL-COST is designed to satisfy all criteria simultaneously (see Appendix A for a detailed evaluation of existing datasets against these criteria).

3 Dataset Creation

Design rationale. Our dataset targets a well-documented prediction from memory-based theories of sentence processing: longer dependencies between a subject and a verb yield higher integration costs at the verb (Gibson, 2001; Lewis et al., 2006). To isolate this effect, we contrast sentences

that have maximally overlapping vocabulary but varying distance between the main subject and main verb, which we manipulate using intervening material.³

Dependency length can increase in two ways: linearly, by inserting prepositional phrases or also hierarchically, by embedding relative clauses. Both increase subject-verb distance, but differ in processing demands: linear modifications incrementally extend dependencies, while hierarchical embeddings introduce qualitatively greater complexity by nesting clauses and requiring the reader to manage multiple layers of syntactic information, increasing storage cost on working memory (Gibson, 2001; Liu et al., 2026), though see Section 6 for an unexpected reversal of this prediction at double embedding. Structural complexity may also contribute independently of dependency length: object relative clauses are consistently harder than subject ones even at matched distances (Gordon et al., 2001). Individual differences in working memory capacity (WMC) may further modulate these effects, with lower-WMC readers showing larger slowdowns under long-distance dependencies (King and Just, 1991; Just and Carpenter, 1992). The controlled structure of our design allows us to tease apart these potentially confounded factors, through two sentence types: *critical sentences* and *control sentences*.

Critical sentences are of primary interest: they

³We focus on English, whose relative clause and center-embedding structures are well-documented in the psycholinguistic literature, providing robust predictions against which to evaluate model behavior.

manipulate syntactic structure while keeping dependency length constant across conditions. The intervening material between subject and verb is either a subject relative clause (SRC) or an object relative clause (ORC), singly embedded (4-word dependency) or doubly embedded (9-word dependency), yielding four core conditions. Two additional conditions anchor the design: a 0-word dependency baseline with no intervening material, and a prepositional phrase (PP) condition at 4 words. The PP condition is critical to the logic of the design: it matches the singly embedded relatives in surface dependency length while introducing no additional clause (i.e., no embedded subject and verb) between subject and verb, isolating linear distance as a factor independently of clausal embedding. Critical items thus comprise six conditions in total (an example set is shown in Table 1).

Control sentences extend the linear distance manipulation introduced by the PP condition in the critical items. They augment dependency length through prepositional phrases of increasing length (0, 4, and 9 words), without embedding an additional clause; an example set can be seen in Table 1. Matched to critical items on dependency length but not hierarchical complexity, they allow direct dissociation of linear distance from structural depth effects, with the PP conditions serving as the common reference point between the two sets.

Stimulus construction. Sentences were generated with LLaMA-3.2 (Grattafiori et al., 2024) and Perplexity.ai (2024) under tight constraints (identical subject-verb pairs, minimal lexical variation, semantic plausibility), then manually reviewed and extensively corrected for naturalness. Complete construction details (constraints, correction criteria, and item statistics) are provided in Appendix B. LLM output served only as an initial drafting aid to vary semantic domains while preserving structural consistency across conditions; final acceptance required independent manual judgments of plausibility, semantic equivalence, and naturalness across all six condition-mates in a set. The final dataset comprises 60 sentences per condition across the six critical item conditions (360 sentences total) and 30 sentences per condition across the three control conditions (90 sentences total), for a grand total of 450 sentences.⁴

⁴Sentences were matched in length (13–15 words) to prevent reading-strategy differences across conditions, since RTs were collected via non-cumulative moving-window self-paced reading. Shorter sentences were extended to match the double-

Data collection. Reading times were collected from native English speakers recruited via Prolific, an online crowdsourcing platform widely used in cognitive research, and using a self-paced reading (SPR) procedure. Participants read one list of 90 counterbalanced sentences (details in Appendix B); comprehension was verified using manually curated comprehension questions and only participations with a score $\geq 70\%$ were kept, yielding a total of 475 participants. Sentences with word reading times below 100 ms or above 3,000 ms are excluded from the dataset as likely reflecting inattention or technical errors ($<5\%$ of the data), yielding a final dataset of 27,103 critical sentences and 13,663 control sentences. After the reading task, participants performed a working memory capacity test using the operation span test. Full participant details, exclusion criteria, task parameters, participant instructions, consent procedure, and recruitment details are provided in the Datasheet in Appendix B.

4 Human Processing Behavior Analysis

Before examining the ability of LMs to predict reading times, we first report a series of control tests to assess the quality of the collected data, where we observe that human reading times follow the patterns described by sentence processing theories. We apply a Region of Practical Equivalence (ROPE) analysis, comparing RTs between different sentence groups and considering differences smaller than 0.1 SD of the RT distribution (35 ms in our data) as negligible. This helps us determine whether the observed differences are meaningful, rather than merely detectable due to our large sample size.

Structure drives difficulty beyond linear distance.

We first observe the effect of dependency length on verb reading times, collapsing across clause types. Reading times increase from 0-word ($M = 490$ ms) to 4-word dependencies ($M = 544$ ms), and decrease slightly at 9 words ($M = 528$ ms). This pattern aligns partially with sentence processing theories, which argue that longer dependencies lead to higher processing costs, but already suggests that distance alone does not drive difficulty: the ROPE analysis confirms that both the 0- vs. 4-word ($d = -0.15$, 95 % CI $[-0.19, -0.12]$) and 0- vs. 9-word contrasts ($d = -0.11$, 95 % CI

embedded ones by appending a coordinated clause in the post-critical region.

Condition	Dep. (words)	N	Mean RT	SD	Median RT	Cohen's d (vs. baseline)	
						d	95% CI
Baseline	0	4,574	490	308	422	—	—
PP	4	4,562	512	321	435	−0.08	[−0.12, −0.04]
SRC	4	4,549	539	370	439	−0.15	[−0.19, −0.12]
ORC	4	4,496	583	408	467	−0.19	[−0.24, −0.15]
2×SRC	9	4,545	504	302	436	−0.11	[−0.15, −0.08]
2×ORC	9	4,377	553	383	438	−0.14	[−0.18, −0.10]

Table 2: RTs (ms) at the main verb by condition, with Cohen's d relative to the baseline (0-word dependency).

[−0.15, −0.08]) exceed the ROPE, while the 4- vs. 9-word contrast is negligible ($d = 0.05$, 95 % CI [0.02, 0.07]). Collapsing across clause types, however, confounds distance with syntactic structure, so we turn to condition-level comparisons to investigate the impact of syntactic structure on processing difficulty.

Figure 1 displays average log-transformed RTs at each word position for all six conditions and shows a consistent spike at the main verb across all conditions⁵. All conditions slow readers down relative to baseline, but to varying degrees (Table 2). Crucially, at the same 4-word dependency length, the three conditions pattern differently: the PP produces a negligible effect relative to the ROPE ($d = -0.08$, 95 % CI [−0.12, −0.04]), the SRC a small but meaningful one ($d = -0.15$, 95 % CI [−0.19, −0.12]), and the ORC the largest one ($d = -0.19$, 95 % CI [−0.24, −0.15]). That three conditions sharing identical dependency length produce graded slowdowns suggests that the structural source of the dependency contributes to processing difficulty independently of its surface length. At 9-word dependencies, the same ordering holds: 2×SRC ($d = -0.11$, 95 % CI [−0.15, −0.08]) is slower than baseline but faster than 2×ORC ($d = -0.14$, 95 % CI [−0.18, −0.10]). Comparing across embedding depths, the increase from single to double embedding is modest: SRC and 2×SRC differ by $\Delta d = 0.04$, and ORC and 2×ORC by $\Delta d = 0.05$, both negligible relative to the ROPE. Clause type thus appears to be a stronger determinant of processing cost than embedding depth – and, in fact, the direction of this difference runs counter to expectation: doubly-embedded conditions are consistently *faster* than their singly-embedded counterparts for both clause types, a reversal we return to in Section 6, a reversal al-

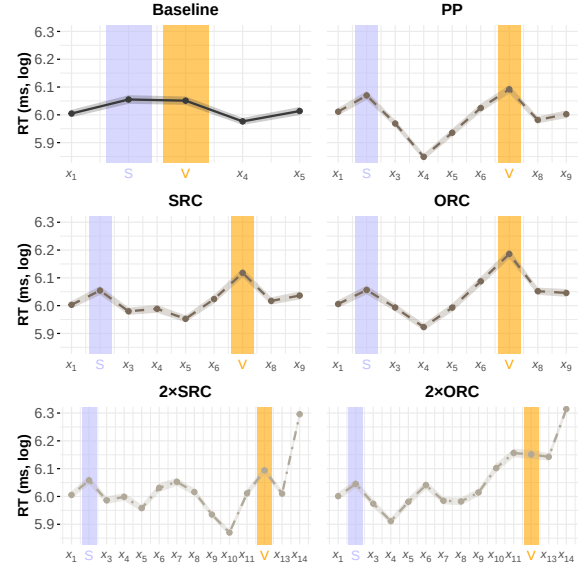


Figure 1: Average log-transformed RTs by condition with 95 % CIs (shaded ribbon). Curve colour indicates dependency length between **subject noun** (S) and **main verb** (V): dark grey = 0-word; medium grey = 4-word; light grey = 9-word.

ready visible in Table 2.

Mixed-effects models confirm that dependency length contributes to reading times even after accounting for standard psycholinguistic confounds (i.e., word length and frequency). In the following inferential analyses, we log-transform reading times (base 2), as is standard in psycholinguistics, which reduces skewness, stabilizes variance across conditions, and facilitates comparisons across studies. All models predict log-transformed RTs at the main verb with random intercepts for participants, items, and presentation order, fitted using the lme4 package (Bates, 2010). A baseline model (M0) including word length and log-transformed lexical frequency from WikiText-103 (Merity et al., 2016) was compared via likelihood ratio test to a model additionally including dependency length

⁵The end-of-sentence rise in the nested sentences (light gray curves) follows the expected SPR wrap-up effect; the other conditions end flat because their final word is the onset of a coordinated clause, not the sentence boundary.

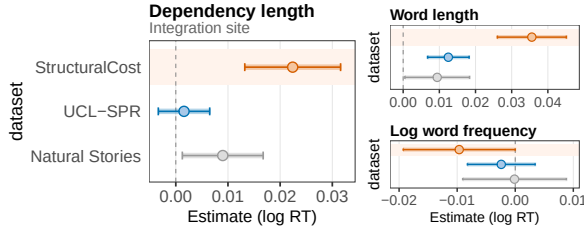


Figure 2: Fixed-effect estimates from M1 fitted separately to Natural Stories, UCL-SPR, and STRUCTURALCOST (at the integration site). All predictors are z -scored; estimates are on the log-RT scale. Error bars are 95 % CIs.

(M1)⁶. The extended model significantly improves fit $\chi^2(1) = 21.6$, $p < .001$,⁷ and all three predictors reach significance: word length ($\beta = 0.012$, $t = 5.04$, $p < .001$), log frequency ($\beta = -0.009$, $t = -4.10$, $p < .001$), and dependency length ($\beta = 0.0065$, $t = 4.73$, $p < .001$). Crucially, the dependency length effect corresponds to an increase of approximately 3 ms per additional intervening word, a stable estimate in our large sample collected outside controlled laboratory conditions.

Comparison with other datasets. To situate STRUCTURALCOST against related resources with less control, we fit the same model M1 to Natural Stories (Futrell et al., 2021) and UCL-SPR (Frank et al., 2013) at the integration site of the main verb and report the resulting fixed-effect coefficients in Figure 2 (implementation details in Appendix D). The dependency length effect decreases at the main verb in both naturalistic corpora, while it increases in STRUCTURALCOST. Back-transforming from an example 500 ms baseline, this corresponds to a ~ 1 –5ms difference per SD of dependency length in naturalistic datasets versus ~ 11 ms in STRUCTURALCOST – a difference our controlled design can isolate because dependency length is varied orthogonally to lexical properties. While word frequency effects are relatively similar across datasets, word length is a stronger RT predictor in STRUCTURALCOST than in the naturalistic corpora, which is expected given that our stimuli contain longer verbs by design (7.5 characters on average vs. 5 in the naturalistic datasets). These comparisons

⁶ $\log(RT) \sim \log(freq) + len + dep + (1|part.) + (1|item)$

⁷The likelihood ratio test compares two nested models by asking whether the gain in fit, measured as the drop in negative log-likelihood, analogous to a drop in training loss, is larger than what chance alone would produce. The χ^2 distribution provides the null reference for that drop.

suggest that STRUCTURALCOST complements naturalistic corpora by providing experimental control over dependency length effects.

5 Probing Surprisal as a Cognitive Proxy

We evaluate the extent to which LM surprisal captures the graded cost of integrating a subject with its verb across an intervening dependency as opposed to surface-level predictability. This requires dissociating two levels of alignment, which we term *global* and *local*.

Global alignment asks whether surprisal correlates with RTs across all word positions. This is a meaningful and well-established result (Smith and Levy, 2013), but a weak test of cognitive plausibility: a model that has learned general word frequency and co-occurrence statistics will score well here, because most RT variance reflects lexical predictability rather than structural cost. Global alignment is therefore a necessary baseline but is not enough to model human processing: averaging over all positions dilutes the structural signal present at the main verb, which we evaluate separately as local alignment.

Local alignment measures the same correlation restricted to a theoretically motivated position: the main verb, where the subject-verb dependency is resolved and humans show graded difficulty as a function of dependency length and syntactic structure. This is a much stricter test: fitting RTs well requires assigning higher surprisal when the dependency is harder to integrate, a sensitivity to structural cost that goes beyond surface statistics. A model that passes the global test but fails the local one has learned lexical predictability without encoding the memory-based integration cost that drives human processing difficulty.

5.1 Models and Setup

Language models. We selected models spanning a memory constraint spectrum: two KenLM n -gram models (Heafield, 2011)⁸ as a Markovian baseline with a hard local window of 5 tokens; three GPT-2 models (small, large, XL; Radford et al., 2019) and three PYTHIA models (70M, 1.4B, 12B; Biderman et al., 2023) as representatives of the currently dominant transformer architecture with unconstrained self-attention; and three MAMBA models (130M, 1.4B, 2.8B; Gu and Dao, 2024) im-

⁸Trained on WikiText-103, tokenized with GPT-2 and PYTHIA tokenizers.

plementing selective state-space compression as a proxy for recency-biased memory and incremental recurrent processing at scale. Within each family, size variation tests whether scaling improves behavioral fit while controlling for training data and vocabulary.

Surprisal extraction. Word-level surprisal was extracted for every sentence, aggregating subword tokens following [Pimentel and Meister \(2024\)](#)⁹. Preprocessing details are given in Appendix C.

Statistical analyses We assess alignment between model surprisal and human RTs using two complementary methods. The first captures broad co-variation: for each word position, we compute Spearman correlations between model surprisal and mean RT across participants, evaluated both *globally* (all positions) and *locally* (main verb only). The second tests whether surprisal contributes independently of known low-level predictors: we fit a sequence of nested linear mixed-effects models ([Bates, 2010](#)) predicting log RT for each word read by each participant, adding predictors incrementally to isolate the unique contribution of surprisal and dependency length:

- M0** : $\log \text{RT} \sim \log \text{freq} + \text{length} + (1|\text{part.}) + (1|\text{item})$
- M1** : $\dots + \text{surprisal}$
- M2** : $\dots + \text{dependency length}$

Model comparison via likelihood ratio tests quantifies whether each added predictor genuinely improves fit. Note that in the global model, dependency length is a per-sentence condition-level predictor tested as a general robustness check across all word positions; it is only at the local (main-verb) level that dependency length functions as a theoretically motivated integration-cost predictor, since this is where the subject-verb dependency is resolved.

5.2 Results

Global alignment: surprisal as a general RT predictor. We first establish whether LM surprisal predicts human RTs across all word positions, replicating prior work and setting a baseline against which local alignment can be assessed. Spearman correlations with word-level RTs are moderate and strikingly uniform: all models cluster at $r = 0.51$ – 0.53 with tight CIs (± 0.02) far from zero (Figure 3, left panel), regardless of architecture or size.

The mixed-effects analysis confirms this picture: adding surprisal to the baseline model improves fit for all neural LMs ($p < .001$, $\chi^2(1) > 846$; Table 3), indicating that surprisal captures variance in RTs beyond lexical frequency and word length. Further adding dependency length yields additional improvement ($p < .001$), suggesting that structural information contributes independently of surprisal. Global fit is thus robust, but as argued above, not sufficient: whether models encode the specific structural cost of subject-verb integration is the question we turn to next.

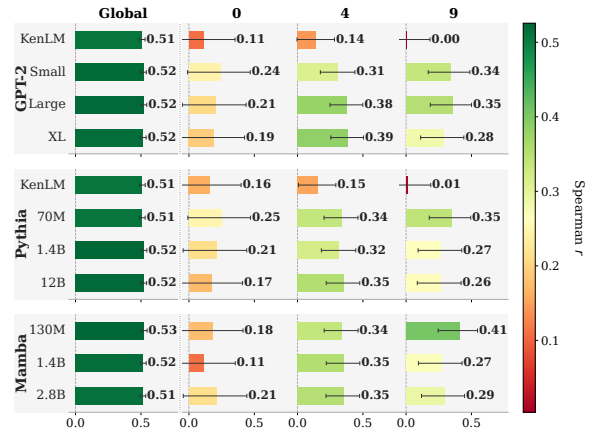


Figure 3: Spearman correlations between model surprisal and reading times, by dependency length (0, 4, 9 words) and fit level: all words (left) vs. verb only (right). Error bars show 95% CIs. See Appendix F for the table.

Local alignment: surprisal at the integration site. Restricting the analysis to the main verb reveals a consistent and informative drop (Figure 3): at dependency length 0, no model yields reliable correlations, with all CIs including zero – likely because the main verb follows immediately after the subject, leaving minimal left context for surprisal to track. At dependency length 4, correlations fall from $r = 0.51$ – 0.53 globally to $r = 0.31$ – 0.39 at the verb for neural models, and collapse to $r = 0.14$ – 0.15 for n -gram models, whose CIs include zero ($[-.01, .29]$). Neural models show significant surprisal effects at dependency length 4 (all $p < 10^{-5}$, $\chi^2(1) = 18$ – 43), with CIs entirely above zero, and, when comparing model surprisal at the main verb itself across conditions matched for linear distance, reproduce the syntactic ordering $\text{PP} < \text{SRC} < \text{ORC}$ at fixed linear distance (all $p < .001$), confirming sensitivity to hierarchical structure beyond raw word count. Dependency length contributes further be-

⁹Using the implementation of [Wisniewski \(2025\)](#).

Model	M0→M1		M1→M2	
	Global	Local	Global	Local
<i>N-gram baselines</i>				
KenLM (GPT-2)	940***	0.02	44***	20.5***
KenLM (Pythia)	993***	0.08	44***	20.6***
<i>GPT-2</i>				
Small	1082***	27.8***	33***	8.0**
Large	1257***	43.4***	29***	5.0*
XL	1010***	31.7***	31***	8.1**
<i>Pythia</i>				
70M	930***	25.3***	42***	10.6***
1.4B	1165***	20.8***	29***	9.4**
12B	1005***	20.3***	29***	9.2**
<i>Mamba</i>				
130M	1045***	23.5***	32***	9.7**
1.4B	906***	18.2***	31***	11.3***
2.8B	846***	18.7***	31***	9.4**

Table 3: $\chi^2(1)$ from likelihood ratio tests, globally and locally at the main verb. M0: baseline (word log frequency + word length); M1: + surprisal; M2: + dependency length. * $p < .05$, ** $p < .01$, *** $p < .001$. Global here tests dependency length as a sentence-level covariate across all positions; Local restricts to the integration site.

yond surprisal (all $p < .05$, $\chi^2(1) = 5\text{--}11$), indicating that structural distance is not fully captured by surprisal alone. At dependency length 9, neural models hold up but unevenly: most settle at $r = 0.26\text{--}0.35$ with wider CIs reflecting reduced power, the SRC < ORC verb-surprisal contrast attenuates, and no model approaches the sharp 70 ms ORC penalty humans incur at fixed linear distance (PP vs. ORC in Table 2), an indication that the failure is qualitative, not merely quantitative. *N-gram* models fail entirely: with the subject noun beyond the 5-gram window at dependency 4 and 9, surprisal adds nothing over baseline predictors at any dependency length ($p > 0.5$, $\chi^2(1) < 0.1$). This global-to-local reduction is the central finding of the present work: models that appeared equivalent under global alignment diverge sharply when tested at the integration site.

Crucially, scaling within families has no consistent benefit: larger models do not fit human integration profiles better, and the largest variants lose correlation points relative to their smaller counterparts. This suggests the gap is not a matter of capacity but a structural one (e.g., inductive bias or training objective). Consistent with this, MAMBA-130M reaches $r = 0.41$ at dependency length 9, the only model whose CI does not overlap with any 0-word dependency interval, approaching the global baseline and outperforming larger attention-based

models. This suggests that memory-constrained architectures produce integration profiles closer to human behavior, though the advantage remains modest and requires further investigation. Reader profile analyses (Appendix E) point to a further limitation independent of architecture or scale: surprisingly, LMs fit slow and medium readers better than fast ones, who display high working-memory readers and are simultaneously faster and more accurate. The fast readers being the hardest to model suggests efficient integration strategies remain outside what surprisal captures.

5.3 Comparison of predicted and observed RTs

If surprisal captured the full integration cost of the dependency, a surprisal-to-RT conversion factor estimated independently should predict RT effects of the observed magnitude. Following Van Schijndel and Linzen (2021) and Huang et al. (2024)¹⁰, we test this by estimating, for each model, a conversion factor (ms per bit of surprisal) on an independent corpus (Natural Stories, Futrell et al., 2021), regressing RT on surprisal at the current and three preceding words (spillover), word length, and log-frequency¹¹. Only four models yielded a significant spillover coefficient at $p < .01$ (Mamba-130M, Pythia-70M, Pythia-1.4B, KenLM/GPT-2) and were retained as conversion factors; we applied their conversion factors to surprisal on STRUC-TURALCOST critical items to obtain predicted Effects of Interest (EOIs)¹².

Table 4 compares predicted and observed EOIs. Predicted effects were between one and three orders of magnitude too small across all four models: for KenLM-GPT2, predicted EOIs ranged 0.02–1.99 ms against observed effects of 14–93 ms, while Pythia-70M predicted a speedup (negative EOI) in every condition. No model recovered the correct direction of the effect in the 2×SRC condition, where predictions were near-zero or negative despite a small observed slowdown. Even calibrated on an independent corpus and estimated via a completely separate method, surprisal did not reproduce the scale of the observed integration cost.

¹⁰Originally developed to test whether surprisal explains syntactic disambiguation difficulty.

¹¹Pythia-12B and KenLM_Pythia were excluded due to technical constraints.

¹²Predicted RT minus predicted Baseline RT.

Condition	Observed EOI (ms)	Predicted EOI (ms)				
		See Table 2	KenLM (GPT2)	Mamba-130M	Pythia-1.4B	Pythia-70M
PP	22		1.04	0.10	−0.07	−1.84
SRC	49		0.02	−1.18	−1.25	−2.13
ORC	93		1.99	0.10	0.03	−1.42
2×SRC	14		−0.21	−1.49	−1.54	−3.20
2×ORC	63		1.69	0.31	0.41	−0.89

Table 4: Predicted versus observed EOIs (relative to Baseline) at the main verb. Predicted values apply spillover-adjusted conversion factors, estimated on Natural Stories, to each model’s surprisal on STRUCTURALCOST critical items; negative values indicate a predicted speedup. Models without a significant conversion factor are omitted.

6 Discussion

Our results both replicate and qualify prior work (Futrell et al., 2021; Oh and Schuler, 2023; Shain et al., 2024): global surprisal-RT alignment is robust but dominated by lexical frequency and word length, explaining why correlations cluster at $r \approx 0.5$ regardless of parameter count, training data, or memory architecture. Any model with basic word statistics will track these signals, which is precisely why global alignment is a weak diagnostic for cognitive plausibility that does not generalize to structural difficulty. Neural models significantly improve fit at the main verb over n -gram baselines and partially mirror the syntactic ordering of conditions, yet they underestimate how processing cost scales with dependency length. The gap persists across architectures and sizes, pointing to a specific limitation: current models appear to capture the predictive component of processing but not the memory-based integration cost that working memory imposes at structural dependencies. The one partial exception is MAMBA-130M, which achieves the highest verb-level correlations at long dependencies, consistent with the hypothesis that selective state-space compression introduces a recency bias that partially mimics distance-dependent memory decay. Whether these advantages reflect genuine architectural sensitivity or confounds in training data and tokenization remains to be established. An unexpected pattern emerges at double embedding: 2×SRC and 2×ORC are consistently read faster than their singly-embedded counterparts. No model predicts this reversal. One possible explanation is within-sentence syntactic adaptation: by the second embedding, readers have already processed the relevant structure once earlier in the same sentence, potentially lowering marginal integration cost even as cumulative memory load increases. We flag this as an open question for future work rather than a settled mechanism. Together,

these findings motivate a reorientation away from global evaluation and toward controlled datasets that isolate memory-sensitive positions. On the modeling side, they favor architectural interventions that introduce explicit memory constraints such as limited context (Kuribayashi et al., 2022), recency-biased attention (De Varda and Marelli, 2024; Clark et al., 2025), or capacity-constrained recurrence (Timkey and Linzen, 2023) over further scaling. STRUCTURALCOST provides the evaluation framework needed to test such interventions directly at the integration site.

7 Conclusions

We introduced STRUCTURALCOST, a controlled self-paced reading dataset isolating the processing cost of dependency length from lexical confounds. Human reading times at the main verb increase with dependency length, driven by syntactic embedding beyond linear distance, confirming that the dataset captures memory-based difficulty. Using STRUCTURALCOST, we evaluated LMs spanning n -gram, SSM, and transformer architectures. Two findings emerge consistently. First, global surprisal-RT alignment is architecturally uninformative: models cluster at similar correlations regardless of architecture or scale, reflecting the dominance of lexical frequency and word length in RT variance. Second, this global alignment does not extend to structural difficulty: neural models partially mirror the graded difficulty profile at the main verb, yet underestimate the integration cost humans incur, and this gap persists across architectures and sizes. These findings suggest that global surprisal-RT alignment is not a sufficient criterion for cognitive plausibility. Current models seem to capture the predictive component of human processing but not the memory-based integration cost. STRUCTURALCOST provides the controlled evaluation data needed to make progress on this gap.

Limitations

STRUCTURALCOST prioritizes internal validity at the cost of ecological generalizability: sentences are controlled constructions in English only, collected online from native speakers, limiting cross-linguistic and cross-population conclusions. The online setting also means the reading environment and equipment were not controlled. A denser condition gradient (e.g., a PP+SRC combination) would further help dissociate linear and hierarchical contributions, and would clarify the embedding-depth reversal noted in Section 6; we leave this to future work. Working memory capacity scores, though collected, are not yet fully leveraged here; a fuller individual-differences analysis is a natural extension. The 0.7 comprehension accuracy cutoff is conventional; participants just below this threshold may have been engaged readers. Due to an implementation error, the dataset only contains each participant’s overall comprehension accuracy, not their per-item comprehension results; the latter would have allowed a finer-grained comparison between reading times and comprehension accuracy at the item level. On the modeling side, LM surprisal captures predictive processing cost but not the integration cost arising from working memory constraints: a limitation that motivates the dataset rather than undermining it but that questions nonetheless the reliability of surprisal as a complete proxy for human processing difficulty. Finally, we evaluate a representative but non-exhaustive set of architectures; other models may behave differently, and their evaluation on STRUCTURALCOST remains open.

Ethical considerations

This study has required an interaction with humans. As such, it has been reviewed and approved by an Institutional Review Board (IRB No. 00012024-128).

Acknowledgments

This research has been financially supported by the Agence Nationale de la Recherche (project COMPO, ANR-23-CE23-0031-01). We also want to thank the EMNLP reviewers for their constructive suggestions, as well as those from participants of the *Computational Psycholinguistics Meeting 2025* where the dataset had been initially presented.

References

- Douglas M Bates. 2010. lme4: Mixed-effects modeling with r.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usven Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.
- Christian Clark, Byung-Doh Oh, and William Schuler. 2025. [Linear recency bias during training improves transformers’ fit to reading times](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747, Abu Dhabi, UAE. Association for Computational Linguistics.
- Andrew RA Conway, Michael J Kane, Michael F Bunting, D Zach Hambrick, Oliver Wilhelm, and Randall W Engle. 2005. Working memory span tasks: A methodological review and user’s guide. *Psychonomic bulletin & review*, 12(5):769–786.
- Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.
- Andrea De Varda and Marco Marelli. 2024. Locally biased transformers better align with human reading times. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 30–36.
- Fernanda Ferreira and Zoe Yang. 2019. The problem of comprehension in psycholinguistics. *Discourse Processes*, 56(7):485–495.
- Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.
- Richard Futrell, Edward Gibson, Harry J. Tily, Ian Blank, Alex Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. [The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions](#). *Language Resources and Evaluation*, 55(1):63–77.
- Edward Gibson. 2001. [The dependency locality theory: A distance-based theory of linguistic complexity](#). In Alec P. Marantz, Yukiyasu Miyashita, and William O’Neil, editors, *Image, Language, Brain*, pages 95–126. The MIT Press.
- Adam Goodkind and Klinton Bicknell. 2018. [Predictive power of word surprisal for reading times is a linear function of language model quality](#). In *Proceedings*

- of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.
- Peter C Gordon, Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, 27(6):1411.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daniel Grodner and Edward Gibson. 2005. [Consequences of the serial nature of linguistic input for sentential complexity](#). *Cognitive Science*, 29(2):261–290.
- Albert Gu and Tri Dao. 2024. [Mamba: Linear-time sequence modeling with selective state spaces](#). *Preprint*, arXiv:2312.00752.
- John Hale. 2001. [A probabilistic earley parser as a psycholinguistic model](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*, pages 1–8.
- Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.
- Nora Hollenstein, Jonathan Rotsztein, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):180291.
- Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.
- Marcel A Just and Patricia A Carpenter. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122.
- Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.
- Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2):153–168.
- Jonathan King and Marcel Adam Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602.
- Victor Kuperman, Sascha Schroeder, and Daniil Gnetov. 2024. [Word length and frequency effects on text reading are highly similar in 12 alphabetic languages](#). *Journal of Memory and Language*, 135:104497.
- Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Robert L. Lewis, Shravan Vasishth, and J. A. Victor Dyke. 2006. [Computational principles of working memory in sentence comprehension](#). *Trends in Cognitive Sciences*, 10(10):447–454.
- Jinlu Liu, Ming Wu, and Haitao Liu. 2026. [Beyond clause counting: a processing perspective on syntactic complexity in english](#). *Corpus Linguistics and Linguistic Theory*.
- Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. [Pointer sentinel mixture models](#). *Preprint*, arXiv:1609.07843.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.
- Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.
- Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make llms less superhuman. *Trends in cognitive sciences*.
- Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1915–1921.
- Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. [Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Perplexity AI. 2024. [Perplexity ai](#).

Tiago Pimentel and Clara Meister. 2024. [How to compute the probability of a word](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Christopher Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. [Large-scale evidence for logarithmic effects of word predictability on reading time](#). In *Proceedings of the National Academy of Sciences*, volume 121, page e2307876121.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, and 1 others. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (mecor). *Behavior research methods*, 54(6):2843–2863.

Nathan J. Smith and Roger Levy. 2013. [The effect of word predictability on reading time is logarithmic](#). *Cognition*, 128(3):302–319.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipes: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 8705–8720.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Titus von der Malsburg. 2015a. [Py-Span-Task: A software for testing working memory span](#).

Titus von der Malsburg. 2015b. [Self-hosted linguistic experiments](#).

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Guillaume Wisniewski. 2025. [Probability of a word](#). GitLab. Personal implementation with bugfixes of Pimentel and Meister, 2024.

Appendices

A Dataset comparison

Table 5 compares STRUCTURALCOST against existing reading time datasets across the four criteria we argue are necessary to study LLM replication of human processing difficulty.

B STRUCTURALCOST datasheet

B.1 Design & conditions

Critical conditions. Each of 60 item sets contains 6 sentences sharing the same main subject-verb (NV) pair. Conditions vary dependency length and syntactic structure as shown in Table 1.

Control conditions. Control sentences isolate linear distance effects without clause embedding. They do not share NV pairs with critical items and use only prepositional phrases to increase distance (Table 1).

Sentence construction constraints. All critical and control sentences were subject to the following constraints during generation and manual review:

- All six sentences in a set share the same main subject and main verb, with intervening words held constant or maximally similar across conditions.
- Main verbs are never sentence-final to avoid spillover and end-of-sentence effect ambiguity.
- Embedded nouns match the number of the main subject to prevent agreement-attraction confounds (e.g., “The *babysitters*_{PL} who the *children*_{PL} loved prepare the diner”).
- Embedded clauses use past tense while main clauses use present tense.
- All sentences are 13–15 words; shorter conditions are padded with a post-critical coordinated clause to match length (see Table 1).
- Cloze-likely verbs were replaced to avoid predictability artifacts (e.g., “The baker bakes” or “The violinist plays” where “prepares” or “leaves” were preferred).
- Each sentence was judged semantically correct and plausible by a proficient speaker.

Dataset		Method	Uniq. Sent.	Subj.	Subj./Item	C1	C2	C3	C4
<i>Naturalistic datasets</i>									
Natural Stories	Futrell et al., 2021	SPR	485	181	~100	✗	✗	✓	✓
Dundee	Kennedy et al., 2003	ET	2,368	10	10	✗	✗	✓	✗
Provo	Luke and Christianson, 2018	ET	~138	84	~8	✗	✗	—	✗
UCL Corpus	Frank et al., 2013	SPR + ET	361 + 205	117 + 43	*	✓	✗	✓	*
MECO	Kuperman et al., 2024	ET	104	50	*	✗	✗	—	*
GECO	Cop et al., 2017	ET	5,300	14	14	✗	✗	✓	✗
ZuCo	Hollenstein et al., 2018	ET + EEG	1,107	12	12	—	✗	✓	✗
<i>Controlled datasets</i>									
Sent. Complexity	Grodner and Gibson, 2005	SPR	16 + 30	42+49	21 & 49	✓	✓	✗	✗
Syn. Ambiguity	Huang et al., 2024	SPR	18-24	2,000	220-440	✓	✓	✗	✓
<i>This work</i>									
STRUCTURALCOST		SPR	450	475	80	✓	✓	✓	✓

Table 5: Comparison of reading time datasets against four criteria required to study LLM replication of human processing difficulty (non-exhaustive list). **C1**: isolated sentences; **C2**: systematic dependency length manipulation; **C3**: sufficient items for model evaluation; **C4**: sufficient participants per item. ET = eye-tracking; SPR = self-paced reading. ✓ = satisfied; ✗ = not satisfied; — = partially satisfied; * = information not found.

The revision pipeline involves:

- Proportion of sentence sets substantially modified or replaced: 60% for the critical sentences and 90% for the control sentences;
- Number of reviewers: 1 for the critical sentences and 2 for the control sentences;
- Total hours: ± 70 .

Comprehension assessment. True/false comprehension questions followed more than half (56%) of sentences to ensure active processing ([Ferreira and Yang, 2019](#)). Each participant answered 30 questions on control sentences and 20 on critical ones, with an equal number of true and false answers, ensuring consistent reading strategy across sentence types. Participants received feedback via a progress bar showing their running accuracy. Questions were not created for doubly embedded conditions (2×SRC and 2×ORC), which are already taxing to process, and never targeted the main verb to avoid drawing attention to the critical region.

Items per participant. Table 6 summarizes the number of items seen per participant. The 30 critical 4-word dependency length items comprise 10 PP + 10 SRC + 10 ORC. The 20 critical 9-word dependency length items comprise 10 2×SRC + 10 2×ORC.

B.2 Item counts & list structure

Latin square counterbalancing. Items were distributed across 6 lists using a Latin square design so

	Dep. 0	Dep. 4	Dep. 9	Total
Critical	10	30	20	60
Control	10	10	10	30
Total	20	40	30	90

Table 6: Items per participant by dependency length.

that each participant saw exactly one sentence per item set, with condition assignment rotating across lists. Each list contains 60 critical items (1 per set), a 50%/50% singular/plural number balance, and 30 control sentences.

B.3 Participants: recruitment and payment

To avoid confounding structural difficulty with proficiency effects, we restrict our analysis to native speaker data exclusively. Participants were recruited via Prolific, restricted to self-reported native English speakers with U.S. nationality or longest residence in the U.S., and no diagnosed or suspected reading disorder. The experiment lasted approximately 30 minutes and participants were compensated £5 (£10/hour, approximately \$13.50/hour), consistent with Prolific’s fair pay guidelines.

B.4 Consent procedure

Before the experiment began, participants were presented with a full information letter describing the study objectives, procedure, risks, and their rights (including the right to withdraw at any time without consequence). They provided explicit consent to

participate in a study on reading times, acknowledging that no directly identifying data would be reported. Only consenting participants were admitted. Data was pseudonymized upon collection and anonymized before analysis.

B.5 Instructions given to participants

General instructions.

This experiment consists of two tasks: a reading comprehension task and a memory test. In the first part, you will read several sentences, one word at a time, and answer a comprehension question if one follows. In the second part, you will evaluate the correctness of simple math equations while memorizing letters which will appear in between. Please make sure you are in a quiet environment and have a stable internet connection before you begin.

Reading task instructions.

You will now read a series of sentences and answer a simple question about some of them. The sentences will appear one word at a time. Press the SPACE bar to reveal each new word. You cannot go back to previous words, so read each sentence carefully and naturally at a pace that feels normal to you. Your goal is to read for comprehension in order to score 80% accuracy or higher on the comprehension questions.

Memory task instructions.

In this task, you will do two things: evaluate the correctness of math equations and memorize the letters that appear after each item. Read each math equation aloud and press Y if the equation is correct, or N if it is incorrect. After a series of items, you will be asked to type in the list of letters in the order that you read them.

Debrief.

In this study, we measured whether the length of dependencies between words in a sentence increases reading time. Some sentences were more complex and difficult to process — they were included to

test the limits of human processing capacity. These measures will serve to constrain language models to process language more like humans.

B.6 Data collection procedure

Self-paced reading task. In self-paced reading, sentences are presented word by word, with each word appearing only when the participant presses a key; reading times are recorded as the time between keypresses. The task was implemented in jsPsych, adapted from [von der Malsburg \(2015b\)](#) for online administration via Prolific.

Working memory capacity. Working memory capacity (WMC) was assessed using an operation span task ([Conway et al., 2005](#)), in which participants solve arithmetic operations while maintaining sequences of letters in memory. Scores ranged from 0 to 6 (Table 7), with concentration in mid-range values consistent with healthy adult populations. The task was adapted from [von der Malsburg \(2015a\)](#).

OSpan score	N	%
0	38	8
1	0	0
2	32	7
3	84	18
4	131	28
5	116	24
6	74	16
Total	475	100.0

Table 7: Distribution of operation-span (OSpan) scores across participants. Protocol: [Conway et al. \(2005\)](#); implementation adapted from von der Malsburg’s Python version.

B.7 Stimulus generation

Generation pipeline. Critical sentences were generated using the LLaMA-3.2 API ([Grattafiori et al., 2024](#)), prompting the model with a semantic domain (e.g., *school, health, weather*) and grammatical number (singular/plural), and requesting all six structural variants simultaneously with a shared subject-verb pair and maximal lexical overlap. Control sentences were generated one set at a time using Perplexity.ai ([Perplexity AI, 2024](#)), specifying the PP-only constraint, three dependency lengths (0, 4, 9 words), and maximal lexical overlap. All items then underwent manual review for plausibility, verb predictability, number

agreement, tense consistency, and length matching (13-15 words).

C Preprocessing

The first and last word of each sentence are excluded from all analyses: SPR paradigms produce elevated RTs at sentence boundaries that reflect task mechanics rather than linguistic processing. As a byproduct, this removes the beginning-of-sequence token whose surprisal is undefined or uninformative across LMs. These exclusions are applied uniformly to RT and surprisal files.

Reader profiles are computed on trimmed data: each participant’s mean RT is calculated across all retained word positions, and participants are partitioned into fast, medium, and slow profiles at the 33rd and 66th percentiles. All predictors are z-scored to make estimates comparable across datasets.

D Dependency parsing

Dependency length is operationalized as the linear distance in tokens between the main verb and its subject, extracted automatically via UDPipe (Straka et al., 2016) using the English Universal Dependencies model (Nivre et al., 2016) for UCL corpus. Natural Stories already provide parse files.

E Reader profile alignment

Human variability is non-trivial and theoretically informative. Thus we compute reader profiles on the previously trimmed data: each participant’s mean RT is calculated across all retained word positions, and participants are partitioned into fast, medium, and slow profiles based on the 33rd and 66th percentiles of this distribution. These profiles are used throughout as secondary diagnostics.

We observe an asymmetry concerning which readers LMs approximate most closely. Contrary to the prediction that surprisal, a prediction-based signal, would align with fast, anticipation-driven readers, the opposite holds: fast readers show the lowest verb-level fit across all models ($r = 0.15$ - 0.22 for neural models; near zero for KenLM), while medium and slow readers show consistently higher fit ($r = 0.36$ - 0.40). Cognitive profiling suggests this is not an artifact of inattention: high working-memory readers are simultaneously faster and more accurate, yet the ones LMs fit least well. One interpretation is that efficient readers deploy an integration strategy that bypasses the retrieval

operations surprisal tracks, a dimension of human processing that current LMs, regardless of architecture or scale, do not capture.

F Surprisal-RT correlation

Spearman is used instead of Pearson as the relation between surprisal and RT is not fully linear. Table 8 shows Spearman correlations between model surprisal and mean per-word RT, globally and locally at the main verb by dependency length.

Model	Spearman r [95% CI]			
	Global	Dep=0	Dep=4	Dep=9
<i>N-gram</i>				
KenLM_G	.51 [.49,.53]	.11 [-.14,.35]	.14 [-.01,.28]	.00 [-.18,.19]
KenLM_P	.51 [.49,.53]	.16 [-.09,.39]	.15 [.01,.29]	.01 [-.17,.19]
<i>GPT-2</i>				
Small	.52 [.50,.54]	.24 [-.02,.47]	.31 [.17,.45]	.34 [.16,.51]
Large	.52 [.50,.55]	.21 [-.05,.44]	.38 [.25,.51]	.35 [.18,.52]
XL	.52 [.50,.54]	.19 [-.05,.42]	.39 [.26,.51]	.28 [.12,.45]
<i>Pythia</i>				
70M	.51 [.49,.53]	.25 [-.00,.50]	.34 [.19,.47]	.35 [.17,.51]
1.4B	.52 [.50,.55]	.21 [-.05,.45]	.32 [.18,.46]	.27 [.08,.44]
12B	.52 [.50,.55]	.17 [-.09,.41]	.35 [.21,.48]	.26 [.08,.43]
<i>Mamba</i>				
130M	.53 [.50,.55]	.18 [-.07,.42]	.34 [.21,.47]	.41 [.24,.57]
1.4B	.52 [.50,.54]	.11 [-.14,.35]	.35 [.21,.48]	.27 [.09,.44]
2.8B	.51 [.49,.54]	.21 [-.04,.43]	.35 [.21,.49]	.29 [.10,.46]

Table 8: Spearman correlations between model surprisal and mean per-word RT, globally and locally at the main verb by dependency length. Bold marks the single best local score. CIs computed via bootstrap ($k = 1000$).