An Existence Proof for Neural Language Models That Can Explain Garden-Path Effects via Surprisal

Ryo Yoshida[♠] **Shinnosuke Isono**[♡] **Taiga Someya**[♠]
Yohei Oseki[♠,◇*] **Tatsuki Kuribayashi**[♣*]
[♠]The University of Tokyo [♡]NINJAL [◇]NII LLMC [♣]MBZUAI
{yoshiryo0617,tsomeya,oseki}@g.ecc.u-tokyo.ac.jp
s-isono@ninjal.ac.jp tatsuki.kuribayashi@mbzuai.ac.ae

Abstract

Surprisal theory hypothesizes that the difficulty of human sentence processing increases linearly with *surprisal*, the negative log-probability of a word given its context. Computational psycholinguistics has tested this hypothesis using language models (LMs) as proxies for human prediction. While surprisal derived from recent neural LMs generally captures human processing difficulty on naturalistic corpora that predominantly consist of simple sentences, it severely underestimates processing difficulty on sentences that require syntactic disambiguation (*garden-path effects*). This leads to the claim that the processing difficulty of such sentences cannot be reduced to surprisal, although it remains possible that neural LMs simply differ from humans in next-word prediction. In this paper, we investigate whether it is truly impossible to construct a neural LM that can explain garden-path effects via surprisal. Specifically, instead of evaluating off-the-shelf neural LMs, we fine-tune these LMs on garden-path sentences so as to better align surprisal-based reading-time estimates with actual human reading times. Our results show that fine-tuned LMs do not overfit and successfully capture human reading slowdowns on held-out garden-path items; they even improve predictive power for human reading times on naturalistic corpora and preserve their general LM capabilities. These results provide an existence proof for a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal, but also raise a theoretical question: what kind of evidence can truly falsify surprisal theory?

1 Introduction

Surprisal theory (Hale, 2001; Levy, 2008) hypothesizes that human sentence processing involves *prediction*, and that processing cost increases linearly with *surprisal*, the negative log-probability of a word given its context. Computational psycholinguistics has empirically tested this hypothesis using language models (LMs), which are trained for next-word prediction on text corpora. Specifically, next-word probabilities $p_\theta(\text{word} \mid \text{context})$ from LMs are used as proxies for human predictability $p_{\text{human}}(\text{word} \mid \text{context})$, under the principle that "frequency affects performance" (Hale, 2001, Principle 2). Surprisal from various LMs, including Probabilistic Context-Free Grammars (PCFGs), n-grams, and neural LMs, has been shown to explain human reading times and neural responses, which presumably reflect processing difficulty, providing empirical support for surprisal theory (Hale, 2001; Smith and Levy, 2013; Frank et al., 2015, *inter alia*).

However, while surprisal from recent neural LMs generally captures human sentence processing difficulty on naturalistic corpora that consist predominantly of simple sentences (Goodkind and Bicknell, 2018; Wilcox et al., 2020), it severely underestimates processing difficulty on sentences that require syntactic disambiguation (van Schijndel and Linzen, 2021; Huang et al., 2024)—*garden-path effects* observed in sentences like "the horse raced past the barn fell" (Bever, 1970). There are two possible reasons for this failure of neural LM surprisal (Huang et al., 2024, page 13). The first possible reason lies in probability estimation: the probabilities that humans and neural LMs assign to words given their context differ in some cases, resulting in the failure of neural LM surprisal to explain garden-path effects. The second possible reason lies in surprisal theory: the processing difficulty of garden-path sentences cannot be reduced to surprisal. Several recent studies argue for the second possibility (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024; Timkey et al., 2025) (for a review see Staub, 2025, Section 5).

In this paper, we pursue the first possibility by

*Corresponding authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36556–36569

investigating whether it is truly impossible to construct a neural LM that can explain garden-path effects via surprisal. While previous work has primarily evaluated off-the-shelf neural LMs, we analyze whether these LMs can be made to capture garden-path effects without overfitting or sacrificing general LM capabilities. Specifically, we fine-tune them on garden-path sentences so as to better align surprisal-based reading-time estimates with actual human reading times (Sections 3 and 4). Our results show that fine-tuned LMs do not overfit and successfully capture human reading slowdowns on held-out garden-path items; they even improve predictive power for human reading times on naturalistic corpora and preserve their general LM capabilities (Section 5). These results provide an existence proof for a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal. Further analyses demonstrate that LMs fine-tuned on a single garden-path construction also capture human processing difficulty of other unseen garden-path constructions, but that the current method *does not* allow the models to explain processing difficulties that are likely accounted for by memory-based theories rather than surprisal theory (Section 6). Finally, we discuss the theoretical implications for surprisal theory raised by our existence proof (Section 7).[1]

2 Background

2.1 Surprisal Theory

Surprisal theory (Hale, 2001; Levy, 2008) states that the processing cost of an input in human sentence processing is determined by *predictability*, specifically, scaling linearly with its negative log-probability given the context:

$$\text{Cost}_{\text{human}}(w_t \mid \boldsymbol{w}_{<t}) \propto -\log p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t}).$$

Surprisal theory is a computational-level hypothesis in Marr's (1982) three levels of description, providing a characterization of the goal of human sentence processing, while remaining neutral about the representations, algorithms, and implementations that realize this goal (Hale, 2014).

Surprisal theory has several theoretical justifications. Hale (2001) showed that, given a specific grammar, surprisal equals the degree to which syntactic structures defined by that grammar are disconfirmed upon observing each new word:

$$\begin{aligned} &-\log p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t}) \\ &= -\log \frac{\sum_{\tau \in \mathcal{T}(\boldsymbol{w}_{\leq t})} p_{\text{human}}(\tau)}{\sum_{\tau \in \mathcal{T}(\boldsymbol{w}_{<t})} p_{\text{human}}(\tau)}, \end{aligned} \tag{1}$$

where $\mathcal{T}(\boldsymbol{w}_{\leq t})$ denotes the set of syntactic structures consistent with the word sequence up to position t. Under the assumption that "the relation between the parser and grammar is one of strong competence" (Hale, 2001, Principle 1) (see also Chomsky, 1965), this degree of disconfirmation can be interpreted as processing cost—if there were processing costs distinct from those postulated in the grammar, strong competence would be violated (Hale, 2001, Footnote 1). This formulation requires $p_\theta(\tau)$ to be computed over explicit syntactic structures, such as those defined by a PCFG, to approximate $p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t})$.

Levy (2008) offered a more general interpretation, showing that without assuming a specific grammar, surprisal equals the KL divergence between posterior and prior beliefs about the latent structures T:

$$\begin{aligned} &-\log p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t}) \\ &= D_{\text{KL}}\big(p_{\text{human}}(T \mid \boldsymbol{w}_{\leq t}) \,\|\, p_{\text{human}}(T \mid \boldsymbol{w}_{<t})\big). \end{aligned} \tag{2}$$

Under the assumption that probabilities are represented as activation levels of relevant neural structures within the brain—such that larger belief updates correspond to larger physical changes (Levy, 2008, Footnote 8)—this KL divergence can be interpreted as the cost of reallocating cognitive resources upon observing each new word. Crucially, this interpretation allows surprisal computed from string-based LMs such as n-grams and recent neural LMs to capture human belief updating about latent structures, as surprisal functions as a *causal bottleneck* between structural representations and processing cost (Levy, 2008, Section 2.3).

Note that while both formulations are intended as computational-level theories, they may implicitly commit to algorithmic-level assumptions: Hale interpreted his formulation as presupposing *total-parallelism parsing* (Hale, 2001, Section 3), and Levy's interpretation is based on a specific mechanism for cognitive resource reallocation. We return to this point in Section 7.2.

[1] Code for reproducing our results is available at https://github.com/osekilab/RE-GPE.

2.2 Language Models as Proxies for Human Prediction

A fundamental challenge in empirically testing surprisal theory is operationalizing human prediction $p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t})$, which is not directly observable.[2] Traditionally, researchers have employed *cloze tasks* (Taylor, 1953), in which participants are presented with a sentence fragment and asked to predict the next word, with the proportion of participants producing each word serving as an estimate of its probability.

However, this approach suffers from a critical limitation: it cannot reliably estimate low probabilities. This is particularly problematic for surprisal theory, which assumes that processing cost scales with the *logarithm* of probability, meaning that the difference between probabilities 0.0001 and 0.0099 should have the same impact as the difference between 0.01 and 0.99. Cloze tasks cannot capture such distinctions with finite sample sizes.

To address this limitation, recent work has used LMs as proxies for human prediction, assuming that corpus statistics approximate human linguistic experience and that frequency affects language processing performance (Hale, 2001, Principle 2). Empirical results have shown that LM surprisal indeed correlates strongly with human reading times and neural responses (Smith and Levy, 2013; Frank et al., 2015, *inter alia*), even outperforming cloze probability (Shain et al., 2024), providing substantial support for surprisal theory.

However, recent findings have revealed systematic discrepancies. For instance, larger and more sophisticated neural LMs exhibit *worse* predictive power for human reading times despite achieving lower perplexity (Oh and Schuler, 2023; Shain et al., 2024; Kuribayashi et al., 2022, 2024). This inverse relationship suggests that optimizing for next-word prediction does not necessarily improve alignment with human prediction. Another striking manifestation of LM-human misalignment, observed consistently across neural LMs of varying scale and architecture, is the failure to capture garden-path effects. The latter discrepancy is the main focus of this paper; we review it further in the following subsection.

2.3 Garden-Path Effects

When reading sentences like Example (1-a) from left to right, humans exhibit substantially longer reading times at the word *fell* (and subsequent regions reflecting spillover effects, Mitchell, 1984), compared to unambiguous control sentences like Example (1-b) (Bever, 1970):

(1) a. The horse raced past the barn *fell*...
b. The horse that was raced past the barn *fell*...

In psycholinguistics, this phenomenon is explained as follows: at the point of reading *the horse raced past the barn* in Example (1-a), a syntactic ambiguity arises between two interpretations: (i) *raced* is the main verb with *the horse* as its subject, and (ii) *raced* forms a passive reduced relative clause, with *the horse* as the modified noun. Readers prefer interpretation (i), but the appearance of *fell* forces them to abandon this analysis, resulting in increased processing cost, a phenomenon known as the *garden-path effect*.

Traditionally, this difficulty has been attributed to a selective reanalysis mechanism that reconstructs syntactic structures (Fodor and Ferreira, 1998). However, under Hale's formulation and Levy's interpretation of surprisal theory (Equations 1 and 2), this processing cost should be modeled as structure disconfirmation and belief updating, respectively, and thus falls within the scope of surprisal theory.

Recently, multiple studies have shown that surprisal from neural LMs consistently underestimates garden-path effects to a severe degree; for example, it predicts only approximately $1/10$ to $1/30$ of the slowdown observed in self-paced reading times (Huang et al., 2024). This has led researchers to argue not only that next-word probabilities from neural LMs fail to serve as proxies for human predictability but also that syntactic disambiguation difficulty may be irreducible to surprisal.

3 Methods

We adopt a fine-tuning method to align surprisal-based reading-time estimates with actual human reading times (Kiegeland et al., 2024).[3] Whereas Kiegeland et al. fine-tuned neural LMs on naturalistic corpora, we fine-tune neural LMs on garden-path sentences, and evaluate the resulting models on three criteria: (i) whether they generalize to held-out garden-path items without overfitting, (ii) whether they maintain predictive power for hu-

[2] This subsection largely follows Levy (2013, page 164).

[3] https://github.com/samuki/reverse-engineering-the-reader

man reading times on naturalistic corpora, and (iii) whether they preserve general LM capabilities.

Data Let D_{gp} denote the dataset of garden-path sentences, where each data point $d \in D_{\text{gp}}$ consists of a word w_d and its self-paced reading time RT_d (Just et al., 1982). Each data point d is annotated with the following attributes: sentence pair ID $s(d)$, garden-path construction $g(d) \in \{\text{MVRR}, \text{NPS}, \text{NPZ}\}$,[4] sentence ambiguity condition $c(d) \in \{\text{amb}, \text{unamb}\}$ (corresponding to Examples (1-a) and (1-b), respectively), position in sentence $t(d)$, and region of interest (ROI) $r(d) \in \{0, 1, 2, \text{null}\}$, where $r = 0$ denotes the disambiguating position (e.g., *fell* in Example (1)), $r \in \{1, 2\}$ denotes the two subsequent positions potentially reflecting spillover effects, and $r = \text{null}$ denotes positions outside the ROI.

D_{gp} consists of a training set $D_{\text{gp}}^{\text{train}}$ and a test set $D_{\text{gp}}^{\text{test}}$. The training and test sets have no overlap in the verbs that induce syntactic ambiguity (e.g., *raced* in Example (1)) or in words within the ROI. This ensures that the evaluation assesses whether LMs generalize to unseen data points without overfitting. We also use D_{filler} to denote the dataset of naturalistic filler sentences whose reading times were collected in the same experiment as D_{gp}, and D_{nat} to denote a naturalistic corpus whose reading times were independently collected.

For any dataset D, we use $D^{(-)}$ to denote the subset excluding data points corresponding to the first two words of a sentence and sentence-final words, and $D^{(--)}$ to denote the subset further excluding data points in the ROIs ($r = 0, 1, 2$). $D^{(-)}$ can serve as the target for reading time estimation, as spillover variables are undefined for the first two words of a sentence, while sentence-final words may reflect wrap-up effects (Just and Carpenter, 1980). $D^{(--)}$ is used for regression coefficient estimation. This is motivated by surprisal theory: coefficients estimated on "ordinary" reading times, i.e., those unaffected by syntactic disambiguation, should also account for reading times in the ROI, which are affected by disambiguation (Smith and Levy, 2013; van Schijndel and Linzen, 2021).

Loss Function For each data point $d \in D_{\text{gp}}^{\text{train}}$, let the feature vector be $\mathbf{x}_\theta(d) = [\boldsymbol{\iota}_\theta(d)^\top, \mathbf{z}(d)^\top]^\top$, where

$$\boldsymbol{\iota}_\theta(d) = [-\log p_\theta(w_d^{(k)} \mid \text{ctx}_d^{(k)})]_{k=0}^{2}$$

denotes surprisal at the current position ($k = 0$) and two preceding positions ($k = 1, 2$) to capture spillover effects, with $w_d^{(k)}$ denoting the word at position $t(d) - k$ and $\text{ctx}_d^{(k)}$ its preceding context, and $\mathbf{z}(d)$ denotes control variables such as word length.

Each batch $B \subseteq D_{\text{gp}}^{\text{train}}$ is sampled such that it contains equal numbers of sentence pairs from each garden-path construction. For each B, we estimate regression coefficients $\boldsymbol{\beta}_{\theta,B^{(--)}}$ via ridge regression:

$$\boldsymbol{\beta}_{\theta,B^{(--)}} = (X_{\theta,B^{(--)}}^\top X_{\theta,B^{(--)}} + \rho I)^{-1} X_{\theta,B^{(--)}}^\top \boldsymbol{\psi}_{B^{(--)}}.$$

Here, $X_{\theta,B^{(--)}}$ denotes the design matrix with rows $\mathbf{x}_\theta(d)^\top$ for $d \in B^{(--)}$, ρI denotes a regularization term where I is the identity matrix and $\rho > 0$, and $\boldsymbol{\psi}_{B^{(--)}}$ denotes the vector of reading times RT_d for $d \in B^{(--)}$. We then compute the following loss:

$$\mathcal{L}_B(\theta) = \frac{1}{|B^{(-)}|} \sum_{d \in B^{(-)}} \left(\text{RT}_d - \mathbf{x}_\theta(d)^\top \boldsymbol{\beta}_{\theta,B^{(--)}}\right)^2 + \lambda \|\boldsymbol{\beta}_{\theta,B^{(--)}} - \boldsymbol{\beta}_{\theta_0,D_{\text{gp}}^{\text{train}(--)}}\|^2. \tag{3}$$

The first term minimizes the squared residuals between actual and estimated reading times. The second term penalizes deviation of the regression coefficients from the initial coefficients estimated on $D_{\text{gp}}^{\text{train}(--)}$ using the initial LM parameters θ_0.[5]

Evaluation We evaluate based on three criteria:

Garden-Path Effect Alignment We compute regression coefficients $\boldsymbol{\beta}_{\theta,D_{\text{filler}}^{(-)}}$ on $D_{\text{filler}}^{(-)}$ via ridge regression as in the fine-tuning procedure. We then evaluate how well the estimated reading time difference for ambiguous versus unambiguous sentences in $D_{\text{gp}}^{\text{test}}$,

$$\widehat{\Delta\text{RT}}_{g,r}(\theta) = \frac{1}{|S_g|} \times \sum_{s \in S_g} \left[\mathbf{x}_\theta(d(s, \text{amb}, r))^\top \boldsymbol{\beta}_{\theta,D_{\text{filler}}^{(-)}} - \mathbf{x}_\theta(d(s, \text{unamb}, r))^\top \boldsymbol{\beta}_{\theta,D_{\text{filler}}^{(-)}}\right],$$

[4] Main Verb/Reduced Relative clause ambiguity, Noun Phrase/Sentential complement ambiguity, and Noun Phrase/Zero ambiguity, respectively. See Section 4 for concrete examples.

[5] Preliminary experiments revealed that without this term, the LM would artificially inflate estimated reading times in the ROIs by reducing surprisal outside this region to increase the regression coefficients.

aligns with the actual reading time difference $\Delta\mathrm{RT}_{g,r}$ (van Schijndel and Linzen, 2021). Here, S_g denotes the set of test pairs for garden-path construction g, and $d(s, c, r)$ denotes the data point corresponding to pair s, condition c, and region r.

Impact on Naturalistic Corpora We evaluate the per-datapoint log-likelihood improvement of a Gaussian linear regression model including surprisal as a predictor over a baseline model with control variables only (Wilcox et al., 2020):

$$\Delta\mathrm{llh}(\theta) = \frac{1}{|D_{\mathrm{nat}}^{(-)}|} \times \sum_{d \in D_{\mathrm{nat}}^{(-)}} \Big[\log f(\mathrm{RT}_d \mid \mathbf{x}_\theta(d); \boldsymbol{\beta}_{\theta, D_{\mathrm{nat}}^{(-)}}) - \log f(\mathrm{RT}_d \mid \mathbf{z}(d); \boldsymbol{\beta}_{\emptyset, D_{\mathrm{nat}}^{(-)}})\Big],$$

where $f(\cdot \mid \cdot; \boldsymbol{\beta})$ denotes the probability density function of a Gaussian linear regression model with coefficients $\boldsymbol{\beta}$ and the subscript $\emptyset$ indicates the baseline regression model using control variables only. The two regression models are fitted separately on $D_{\mathrm{nat}}^{(-)}$.

Language Model Capabilities To assess whether the fine-tuned LMs preserve general LM capabilities, we additionally evaluate perplexity on naturalistic corpora and grammatical knowledge using BLiMP (Warstadt et al., 2020).

4 Experimental Settings

Language Models We use GPT-2 (Radford et al., 2019) `small` (S), `medium` (M), and `large` (L) as θ_0, using the Hugging Face implementation (Wolf et al., 2020).[6] Prior work on naturalistic corpora has shown that neural LM surprisal from models around the size of GPT-2 small exhibits the best fit to human reading times (Oh and Schuler, 2023; Shain et al., 2024).

Data For D_{gp}, we use the Syntactic Ambiguity Processing (SAP) dataset (Huang et al., 2024).[7] This dataset contains 24 pairs for the following three garden-path constructions.

Main Verb/Reduced Relative Clause (MVRR)

(2) a. The girl fed the lamb *remained* relatively calm...
b. The girl who was fed the lamb *remained* relatively calm...

This garden-path construction is similar to Example (1): whether *fed* is the main verb with *the girl* as its subject, or *fed* introduces a passive reduced relative clause modifying *the girl*. The word *remained* disambiguates ($r = 0$).

Noun Phrase/Sentential Complement (NPS)

(3) a. The girl found the lamb *remained* relatively calm...
b. The girl found that the lamb *remained* relatively calm...

The ambiguity is whether *the lamb* is the direct object of *found* or the subject of a sentential complement. The word *remained* disambiguates ($r = 0$).

Noun Phrase/Zero (NPZ)

(4) a. When the girl attacked the lamb *remained* relatively calm...
b. When the girl attacked, the lamb *remained* relatively calm...

The ambiguity is whether *the lamb* is the direct object of *attacked* or the subject of the main clause. The word *remained* disambiguates ($r = 0$).

Each word is annotated with self-paced reading times from 220–440 anonymized native English speakers, with the number of participants varying by sentence. We exclude observations below 100 ms or above 3000 ms, following Futrell et al. (2018). We use the average reading time across subjects as the representative value, following recent practices (Pimentel et al., 2023; Oh and Schuler, 2023; Kuribayashi et al., 2024). The same preprocessing is applied to all subsequent datasets.

Since this dataset is relatively small for LM fine-tuning, we adopt leave-one-out (LOO) cross-validation. In each fold, we hold out one pair from each of the three garden-path constructions (three pairs total) as test data and construct the training set from the remaining pairs such that it satisfies the non-overlap constraint (see Section 3). After excluding one pair containing data errors, we perform 23 folds and report the average across folds.[8]

For D_{filler}, we use the filler sentences from the same dataset (39 sentences extracted from the

[6] https://huggingface.co/openai-community/gpt2
[7] https://github.com/caplabnyu/sapbenchmark
[8] The training set contains an average of 1645 words across folds, comparable in size to the Provo corpus (Luke and Christianson, 2018) (1113 words), on which Kiegeland et al. (2024) reported the effectiveness of this method.

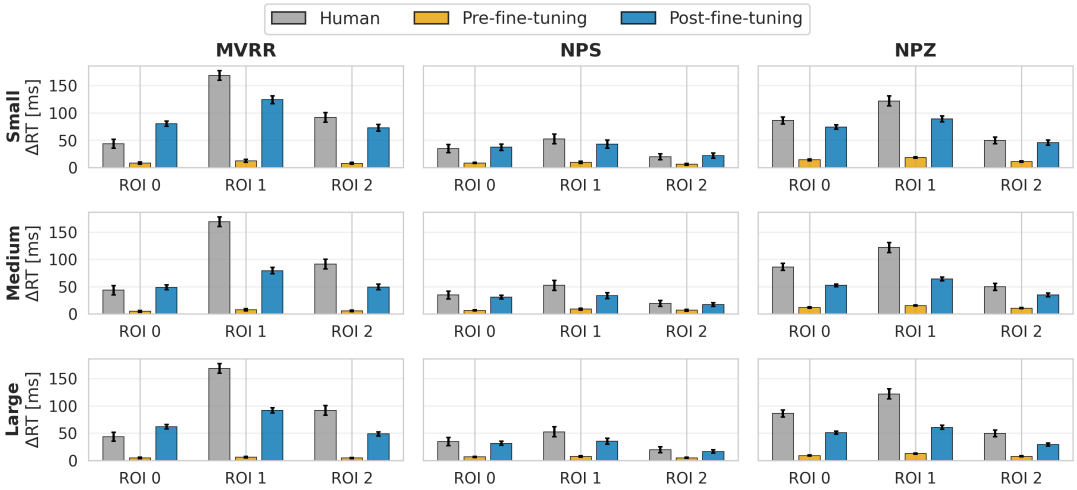

Figure 1: Garden-path effect alignment for pre- and post-fine-tuned LMs. Rows indicate model sizes and columns indicate garden-path constructions, with the x-axis representing the ROI position and the y-axis representing the reading time difference (ms) between ambiguous and unambiguous conditions. Black bars show actual human reading time differences, while orange and blue bars show estimates from pre- and post-fine-tuned LMs, respectively. Error bars represent standard errors across folds.

Provo corpus, Luke and Christianson, 2018). For D_{nat}, we use three corpora: Natural Stories (10 stories with syntactically diverse sentences, 485 sentences, 181 participants, Futrell et al., 2018), Brown (35 passages from written American English, 449 sentences, 35 participants, Smith and Levy, 2013), and UCL (3 unpublished novels from an online fiction platform, 361 sentences, 117 participants, Frank et al., 2013). All corpora are annotated with self-paced reading times from anonymized native English speakers.

Regression Variables The control variable vector $\mathbf{z}(d)$ includes unigram surprisal, word length, and position in sentence. To account for spillover effects, we also include values from one and two words prior for unigram surprisal and word length. Unigram surprisal is estimated using the `wordfreq` library (Speer, 2022), and data points with missing frequency values are excluded from regression. For surprisal, we use the corrected sum of subword surprisals (Oh and Schuler, 2024; Pimentel and Meister, 2024). The details of fine-tuning hyperparameters are provided in Appendix A.

5 Results

Garden-Path Effect Alignment Figure 1 shows the results for garden-path effect alignment. First, while surprisal from pre-fine-tuned LMs qualitatively captures the existence of garden-path effects ($\widehat{\Delta\text{RT}} > 0$), it substantially underestimates their magnitude, consistent with previous work (van Schijndel and Linzen, 2021; Huang et al., 2024). For example, at ROI 1 (the primary focus of analysis in prior work, Huang et al., 2024, Figure 4), even GPT-2 small, which shows the most substantial effect estimates, captures only approximately 7 %, 19 %, and 15 % of the human reading time slowdown for MVRR, NPS, and NPZ, respectively. In contrast, surprisal from post-fine-tuned LMs shows substantially improved alignment with human reading time slowdowns on held-out test set $D_{\text{gp}}^{\text{test}}$. Among the LMs of different sizes, GPT-2 small achieves the best alignment, capturing approximately 73 %, 83 %, and 73 % of the human reading time slowdown at ROI 1 for MVRR, NPS, and NPZ, respectively. Furthermore, regarding the ordering of slowdown magnitudes across constructions, pre-fine-tuned LMs failed to match the human ordering (MVRR $>$ NPZ $>$ NPS) at ROI 1, instead showing NPZ $>$ MVRR $>$ NPS, whereas post-fine-tuned LMs exhibited slowdown magnitudes consistent with human data.

Impact on Naturalistic Corpora Figure 2 shows the results for the impact on naturalistic corpora. Across all corpora and all model sizes, post-fine-tuned LMs demonstrated higher predictive power for human reading times than pre-fine-tuned LMs. Interestingly, this result demonstrates that fine-tuning on garden-path sentences enhances predictive power for human reading times on naturalis-

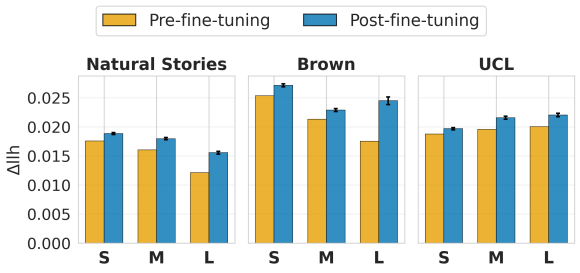

Figure 2: Impact of fine-tuning on predictive power for naturalistic corpora. Each panel corresponds to a naturalistic corpus, with the x-axis representing model size and the y-axis representing the log-likelihood improvement over the baseline regression model with control variables only.

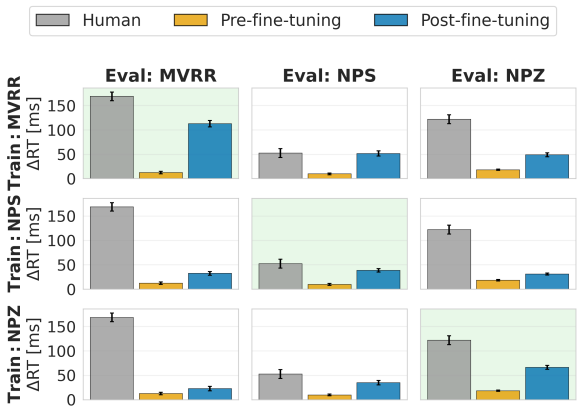

Figure 3: Cross-construction transfer for GPT-2 small at ROI 1. Rows indicate the construction used for fine-tuning, and columns indicate the construction used for evaluation, with a green background highlighting in-domain evaluation.

tic corpora that predominantly contain simple sentences.

Language Model Capabilities Table 2 in Appendix B shows the results for general LM capabilities. As expected, perplexity increases after fine-tuning (e.g., 53.0 → 80.0 for GPT-2 small on Natural Stories), as the training objective primarily increases surprisal. BLiMP accuracy also degrades slightly (e.g., 0.82 → 0.80 for GPT-2 small). Nevertheless, the magnitudes of degradation in both perplexity and BLiMP accuracy are comparable to those observed when fine-tuned on naturalistic reading times (Kiegeland et al., 2024), and the fine-tuned LMs remain well below the uniform perplexity baseline (50257 for all corpora) and well above chance BLiMP accuracy (0.50), indicating that general LM capabilities are largely preserved.

These results provide an existence proof for a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal.

6 Analysis

6.1 Cross-Construction Transfer

In Section 5, we fine-tuned neural LMs using all three garden-path constructions. In this subsection, we fine-tune them on a single garden-path construction and evaluate on all three to investigate whether the LMs learn construction-specific patterns or general mechanisms underlying garden-path effects.

Figure 3 shows the results for GPT-2 small at ROI 1. First, regarding in-domain performance, LMs fine-tuned on a single construction showed substantial improvement over the pre-fine-tuned LM, capturing 67 %, 73 %, and 54 % of the human slowdown for MVRR, NPS, and NPZ, respectively, though performance remained lower than when fine-tuned on all three constructions (see Figure 1).

Crucially, regarding cross-construction transfer, LMs fine-tuned on one garden-path construction better captured human reading-time slowdowns on other garden-path constructions compared to the pre-fine-tuned baseline. For example, the LM fine-tuned on MVRR predicted slowdowns of 51.5 ms (baseline: 9.6 ms) for NPS and 48.9 ms (baseline: 18.1 ms) for NPZ. While the transfer is not perfect—with predictions for constructions different from the training target smaller than those from LMs fine-tuned on that construction in most cases—this result suggests that fine-tuned LMs learn general mechanisms underlying garden-path effects.[9]

6.2 An Unsuccessful Example: Subject/Object Relative Clauses

One potential concern is that the current method allows neural LMs to simulate any kind of processing difficulty. If so, this would undermine the claim that there exists a neural LM that explains garden-path effects via *predictability*. We address this concern by checking whether the current method also allows the models to explain processing difficulties that are unlikely to be due to predictability.

A phenomenon considered difficult to explain under surprisal theory is the processing asymmetry between English subject relative clauses (SRCs) and object relative clauses (ORCs) (Levy, 2008, 2013; Levy and Gibson, 2013):

[9]Medium and large models show broadly similar trends (Appendix C).

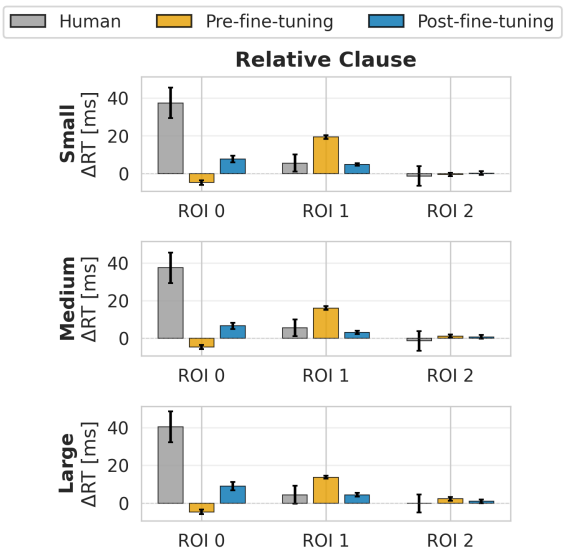

Figure 4: Subject/object relative clause asymmetry alignment for pre- and post-fine-tuned LMs

(5) a. The reporter that the senator *attacked*...
b. The reporter that *attacked* the senator...

Humans exhibit longer reading times at the verb position (*attacked*) in ORCs like Example (5-a) compared to SRCs like Example (5-b). Traditional surprisal theory fails to predict this pattern: the ORC subject provides additional context that makes the verb more predictable, resulting in lower surprisal at the ORC verb than at the SRC verb. Although the increased reading time at the verb could be interpreted as a spillover effect from the unpredictable noun phrase *the senator*, the dominant explanation attributes it to the increased distance to the noun phrase *the reporter* that must be accessed at the point of *attacked*, as posited by memory-based accounts such as Dependency Locality Theory (DLT, Gibson, 2000; Grodner and Gibson, 2005) or Category Locality Theory (CLT, Isono, 2024).

In this subsection, we examine whether fine-tuning succeeds under conditions in which surprisal theory is considered inadequate, specifically for SRC/ORC asymmetry without including spillover variables, and assess its impact on predictive power for naturalistic reading times and general LM capabilities. We use 24 SRC/ORC pairs from the SAP dataset, following the original study in which the verb is designated ROI 0, the determiner ROI 1, and the noun ROI 2. The training and test sets contain completely different words at ROI 0 and ROI 2

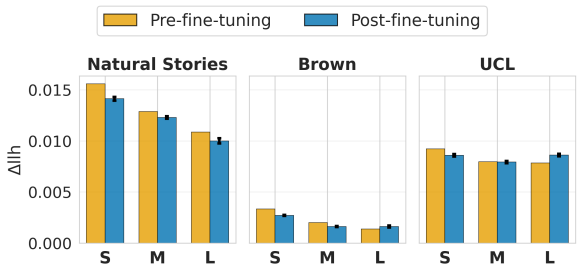

Figure 5: Impact of subject/object relative clause asymmetry fine-tuning on predictive power for naturalistic corpora

under the same LOO setting (see Section 4).

Figure 4 shows the results.[10] First, consistent with prior work, pre-fine-tuned LMs predict a *speed-up* in reading time (-13 % on average across model sizes) at the verb position (ROI 0), where humans show a reading time slowdown for ORCs compared to SRCs, while predicting a slowdown at the determiner position (ROI 1). Post-fine-tuned LMs learn that ORCs exhibit longer reading times than SRCs at the verb position rather than at the determiner position. However, in contrast to garden-path effects, the magnitude of the effect remains limited even for the best-performing GPT-2 large, capturing only 22 % of the human effect. Furthermore, as shown in Figure 5,[11] and again in contrast to garden-path effects, fine-tuning degraded predictive power for human reading times on naturalistic corpora in most conditions except for GPT-2 large on Brown and UCL. As for general LM capabilities (Table 2 in Appendix B), the LMs fine-tuned on SRC/ORC asymmetry show smaller perplexity degradation than the ones fine-tuned on garden-path effects (e.g., 50 $\rightarrow$ 70 for GPT-2 small on Natural Stories)—as expected, given that fine-tuning on SRC/ORC asymmetry less directly increases surprisal than fine-tuning on garden-path effects—but larger degradation in BLiMP accuracy (e.g., 0.82 $\rightarrow$ 0.78 for GPT-2 small).

These results demonstrate that, under conditions in which surprisal theory is considered inadequate, the fine-tuned LMs struggle to reproduce human reading-time differences and show degraded predictive power for naturalistic reading times.

[10] Two of 24 folds for GPT-2 large are excluded from all evaluations due to convergence failures.

[11] The Brown corpus, in particular, showed a substantial decrease in Δllh compared to Figure 2, as this corpus may benefit greatly from spillover predictors.

7 Discussion

7.1 Implications of the Existence Proof

Several recent studies (e.g., Huang et al., 2024) have argued that given the failure of neural LM surprisal to capture garden-path effects, the processing cost of syntactic disambiguation cannot be reduced to surprisal, and consequently additional mechanisms such as reanalysis (Fodor and Ferreira, 1998) are necessary to account for garden-path effects. While these studies were aware that this inference is not deductively valid, noting that the space of possible LMs is unbounded (van Schijndel and Linzen, 2021, Section 4.4), they treated the failure of current neural LMs as strong inductive evidence against surprisal theory. Our existence proof shows that this inductive argument does not hold even in practice. Note that this finding does not logically rule out the reanalysis account. Taken literally, reanalysis to a structure τ at position t requires $p_{\text{human}}(T = \tau | \boldsymbol{w}_{<t}) = 0$. This may give the impression that surprisal and reanalysis are mutually inconsistent explanations of garden-path effects. However, surprisal as a computational-level theory need not commit to the psychological reality of all the parses assigned non-zero probabilities. Under this view, surprisal and reanalysis can be parallel explanations of the garden-path effects.

More fundamentally, the unboundedness of the LM space is problematic not just for surprisal-based accounts of garden-path effects but for surprisal theory itself. Surprisal theory, at its core, merely posits the existence of *some* probability distribution that log-linearly predicts human processing cost without specifying constraints that distribution must satisfy to be interpreted as human predictability. Our proof raises a troubling possibility: if the choice of training procedure is left unconstrained, the expressive power of modern neural LMs may make any such distribution empirically constructible (Bowers and Mitchell, 2025), rendering surprisal theory unfalsifiable in practice.[12]

7.2 Toward a Falsifiable Surprisal Theory

We argue that a promising direction is to improve the falsifiability of surprisal theory (Popper, 1959). We consider two directions:

Direction (i): Constraining the Probability Distribution One direction is to specify constraints that the probability distribution must satisfy, building upon the spirit of Hale's (2001) Principles. For example, consider making Principle 2, "frequency affects performance", a strict constraint. One possibility is to constrain the training procedure: to what extent are interventions on the training distribution permissible? For instance, is fine-tuning with a regularization term (see Equation 3) admissible, or must the distribution be estimated solely from a naturalistic corpus? Another is to constrain the quantity and quality of training data: given Principle 2 appeals to frequencies that humans have experienced, should training be restricted to human-scale corpora (Warstadt et al., 2023)?

Direction (ii): Requiring Psychological Reality A second direction is to abandon the purely computational-level stance and to require the psychological reality of the parse distribution posited in Equations 1 and 2, as Hale (2001) and Levy (2008) appear to have implicitly done (Section 2.1).[13] Under this view, the neural realization of each structure and its probability, posited in the right-hand sides of these equations, becomes an empirical question, thereby grounding the falsifiability of surprisal theory in implementational evidence (see Mangalam, 2025, for a parallel argument regarding the falsifiability of the *Bayesian brain hypothesis*, Friston 2010). Note that under this direction, the reanalysis account and surprisal theory become mutually incompatible: if competing structures are mentally represented in parallel, there is no role for a selective reanalysis mechanism that constructs such structures only post-hoc.

8 Conclusion

In this paper, we provide an existence proof for a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal, while highlighting that surprisal theory may be too flexible to falsify. We propose two directions to make surprisal theory falsifiable: by imposing strict constraints on the probability distribution, or by requiring the psychological reality of the posited parse distribution.

[12] Our additional analysis on SRC/ORC asymmetry (Section 6.2) suggests some empirical bounds on this flexibility, though whether this limitation persists with larger or more diverse training data remains an open question.

[13] Note that this does not require always maintaining full parallelism over all theoretically possible structures. As Levy (2008, Section 2.4) argues, it suffices to commit to parallelism over analyses that are realistically competing in a given context—such as the main verb and reduced relative analyses in garden-path sentences.

Limitations

This study evaluates LMs using leave-one-out cross-validation on a relatively small dataset containing three garden-path constructions. While this dataset represents the largest collection of garden-path sentences with human reading time annotations available to our knowledge, future work should validate our findings on larger-scale data. Additionally, extending this investigation to other languages and garden-path constructions would be valuable for assessing the cross-linguistic generalizability of our findings.

Our study focuses on demonstrating the existence of a neural LM that can explain both garden-path effects and naturalistic reading times via surprisal, but does not investigate the internal mechanisms that change as a result of fine-tuning. While examining such internal mechanisms falls outside the scope of our current research question, future investigations into how fine-tuning modifies internal mechanisms could provide valuable insights.

Acknowledgements

We thank Masaki Kumakawa, Ryo Ueda, Shunsuke Kando, and Taiga Ishii for valuable discussions on this paper. This work used a generative AI tool (Claude) for language editing and coding assistance. This work was supported by JSPS KAKENHI Grant Number JP24H00087, Grant-in-Aid for JSPS Fellows JP24KJ0800, JST BOOST Grant Number JPMJBY24B2, JST CREST Grant Number JPMJCR2565, JST PRESTO Grant Number JPMJPR21C2, JST ACT-X Grant Number JPMJAX25CS, and JST SPRING Grant Number JPMJSP2108.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thomas G. Bever. 1970. The cognitive basis for linguistic structures. In *Cognition and the Development of Language*, pages 279–362. Wiley, New York.

Jeffrey S. Bowers and Jeff Mitchell. 2025. Studies with impossible languages falsify LMs as models of human language. *Preprint*, arXiv:2511.11389.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Janet Dean Fodor and Fernanda Ferreira. 1998. *Reanalysis in Sentence Processing*. Studies in Theoretical Psycholinguistics ; v.21. Kluwer Academic Publishers, Dordrecht ;.

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.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Karl Friston. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press, Cambridge, MA, US.

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.

Daniel Grodner and Edward Gibson. 2005. Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity. *Cognitive Science*, 29(2):261–290.

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*.

John T. Hale. 2014. *Automaton Theories of Human Sentence Comprehension*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, CA.

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.

Shinnosuke Isono. 2024. Category Locality Theory: A unified account of locality effects in sentence comprehension. *Cognition*, 247:105766.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.

Marcel A. Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228–238.

Samuel Kiegeland, Ethan Wilcox, Afra Amini, David Robert Reich, and Ryan Cotterell. 2024. Reverse-Engineering the Reader. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9367–9389, Miami, Florida, USA. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric Predictive Power of Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence Processing*, Current Issues in the Psychology of Language, pages 78–114. Psychology Press, New York, NY, US.

Roger Levy and Edward Gibson. 2013. Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, 4.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Madhur Mangalam. 2025. The myth of the Bayesian brain. *European Journal of Applied Physiology*, 125(10):2643–2677.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, San Francisco.

D. C. Mitchell. 1984. An Evaluation of Subject-Paced Reading Tasks and Other Methods for Investigating Immediate Processes in Reading 1. In *New Methods in Reading Comprehension Research*. Routledge.

Byung-Doh Oh and William Schuler. 2023. 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.

Byung-Doh Oh and William Schuler. 2024. Leading Whitespaces of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

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.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Karl R. Popper. 1959. *The Logic of Scientific Discovery*. Routledge, London.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

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

Robyn Speer. 2022. Rspeer/wordfreq: V3.0. Zenodo.

Adrian Staub. 2025. Predictability in Language Comprehension: Prospects and Problems for Surprisal. *Annual Review of Linguistics*, 11(Volume 11, 2025):17–34.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

William Timkey, Kuan-Jung Huang, Byung-Doh Oh, Grusha Prasad, Suhas Arehalli, Tal Linzen, and Brian Dillon. 2025. Eye movements reveal a dissociation between prediction and structural processing in language comprehension.

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.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan G. Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-TimeComprehension Behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Hyperparameters

Fine-tuning hyperparameters are shown in Table 1. The total computational cost for all experiments was approximately 40 GPU hours on an NVIDIA RTX 6000 Ada (48 GB).

B Language Model Capabilities

Table 2 shows perplexity and BLiMP results for pre- and post-fine-tuned LMs.

C Full Results of Cross-Construction Transfer

Figure 6 shows cross-construction transfer results for GPT-2 medium and large at ROI 1.

D Licenses

Table 3 summarizes the licenses of the datasets and tools employed in this paper. All datasets and tools were used in accordance with their respective license terms.

Optimizer	AdamW (Loshchilov and Hutter, 2019)
LR scheduler	Cosine annealing with warm restarts (Loshchilov and Hutter, 2017)
Batch size	66/44
Training steps	500
Warm-up steps	3
Max learning rate	5.25×10^{-5}/3.5×10^{-5}
Min learning rate	7.8×10^{-8}/5.2×10^{-8}
Decrease rate of max LR	0.01
Weight of loss regularization λ	100
Weight of ridge regularization ρ	1.0×10^{-5}

Table 1: Hyperparameters used for fine-tuning. For parameters with two values separated by a slash, the first value corresponds to training with all three garden-path constructions and the second to training with a single construction.

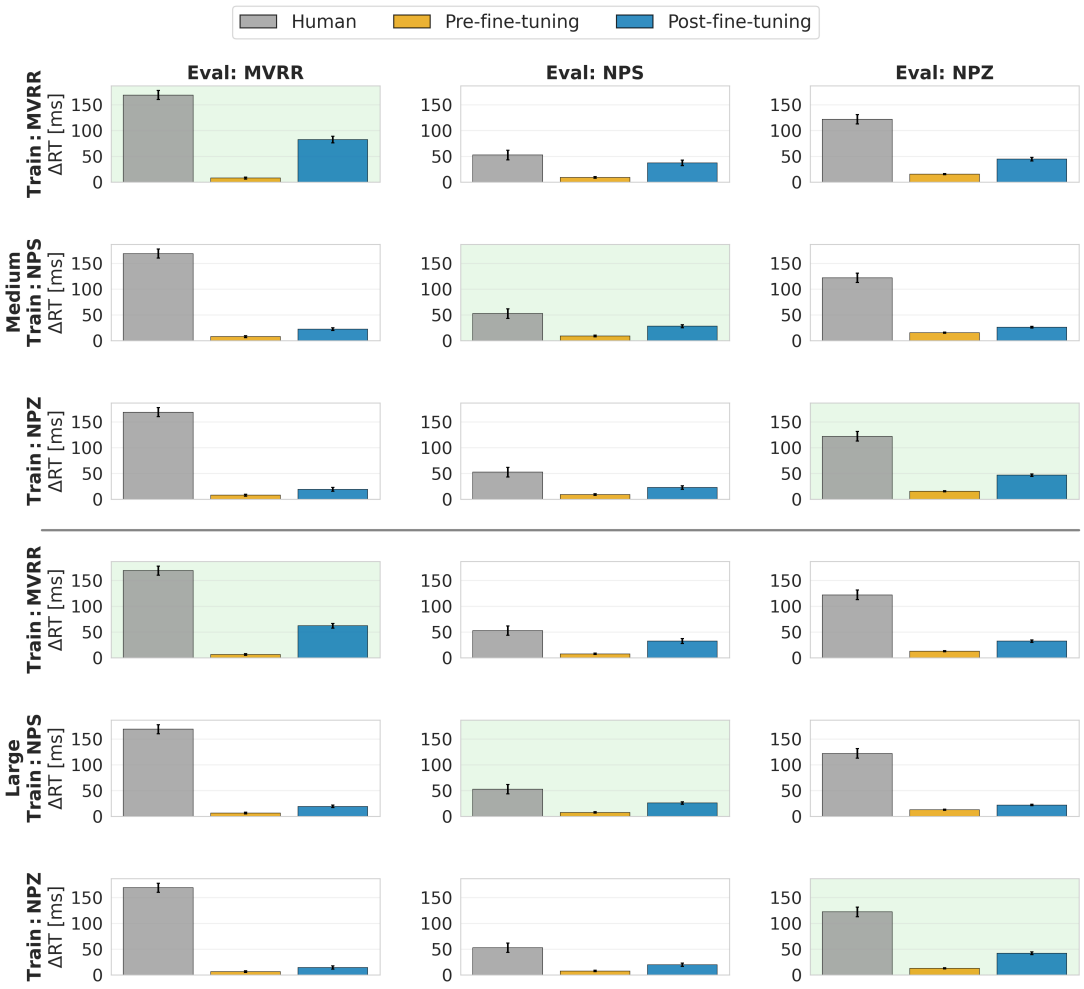

Figure 6: Cross-construction transfer results for GPT-2 medium and large at ROI 1. Within each panel, rows indicate the construction used for fine-tuning and columns indicate the construction used for evaluation, with a green background highlighting in-domain evaluation. The y-axis shows estimated reading time differences (ms) between ambiguous and unambiguous conditions.

Size	Condition	Perplexity (↓) Natural Stories	Brown	UCL	Accuracy (↑) BLiMP
Small	Pre	53.04	78.44	60.95	0.821
	Post (GP)	80.08 ± 0.93	130.17 ± 1.62	100.34 ± 1.62	0.803 ± 0.002
	Post (RC)	70.10 ± 0.47	107.26 ± 0.85	80.29 ± 1.00	0.787 ± 0.002
Medium	Pre	43.82	64.94	50.63	0.827
	Post (GP)	58.83 ± 0.56	93.14 ± 0.97	73.48 ± 1.18	0.822 ± 0.001
	Post (RC)	55.47 ± 0.61	84.88 ± 0.98	65.81 ± 1.09	0.788 ± 0.002
Large	Pre	39.17	59.52	48.42	0.836
	Post (GP)	76.72 ± 14.25	120.95 ± 24.41	81.85 ± 10.52	0.807 ± 0.014
	Post (RC)	62.20 ± 1.65	92.20 ± 2.16	79.10 ± 2.70	0.772 ± 0.005
Uniform baseline		50257	50257	50257	0.500

Table 2: Perplexity (lower is better) on three naturalistic corpora and BLiMP overall accuracy (higher is better) for pre- and post-fine-tuned LMs (GP: garden-path effects, RC: SRC/ORC asymmetry). Values for post-fine-tuned LMs are means ± standard errors across folds. The uniform baseline assumes a uniform distribution over the vocabulary.

Dataset/Tool	License
Datasets	
SAP dataset (Huang et al., 2024)	MIT License
Natural Stories corpus (Futrell et al., 2018)	CC BY-NC-SA 4.0
Brown corpus (Smith and Levy, 2013)	CC BY 3.0
UCL corpus (Frank et al., 2013)	CC BY 3.0
Tools	
`transformers` (Wolf et al., 2020)	Apache 2.0
`wordfreq` (Speer, 2022)	Apache 2.0

Table 3: Licenses of datasets and tools used in this paper