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

Anonymous ACL submission

Abstract

Surprisal, defined as the negative log-probability of a word given its context, has been advocated as a metric for modeling human sentence-processing difficulty. While surprisal from recent neural language models (LMs) generally captures human processing difficulty on naturalistic corpora, it severely underestimates processing difficulty on syntactically ambiguous sentences (garden-path effects), leading to the claim that the processing difficulty of such sentences cannot be reduced to surprisal. In this paper, we investigate whether it is truly impossible to construct an LM that can explain garden-path effects via surprisal. Specifically, while previous work has evaluated off-the-shelf pretrained neural LMs, we fine-tune these LMs on garden-path sentences to align surprisal-based reading-time estimates with actual human reading times, and evaluate both the success of this approach and its impact on predictive power for reading times on naturalistic corpora. Our results show that fine-tuning succeeds without degrading (and in fact improves) predictive power for human reading times on naturalistic corpora, providing an existence proof for an LM that can explain both garden-path effects and naturalistic reading times via surprisal.

1 Introduction

Surprisal theory (Hale, 2001; Levy, 2008) hypothesizes that one goal of the human sentence processing system is *prediction*, and that processing difficulty increases linearly with the negative log-probability of a word given its context. Computational psycholinguistics has experimentally tested this hypothesis under the principle that "Frequency affects performance" (Hale, 2001, Principle 2), using Language Models (LMs) to estimate next-word probabilities $p_\theta(\text{word} \mid \text{context})$ as a proxy for human predictability $p_{\text{human}}(\text{word} \mid \text{context})$. LM surprisal has been shown to explain human reading times and neural responses, which presumably reflect the cognitive load in sentence processing, providing empirical support for surprisal theory (Smith and Levy, 2013; Frank et al., 2015; Wilcox et al., 2023, *inter alia*).

However, while surprisal from recent neural LMs generally captures human processing difficulty on naturalistic corpora with predominantly simple sentences, it severely underestimates processing difficulty on syntactically ambiguous sentences—that is, garden-path effects observed in sentences like "the horse raced past the barn fell" (Bever, 2013; van Schijndel and Linzen, 2021; Arehalli et al., 2022). There are two possible reasons for this failure of LM surprisal (Timkey et al., 2025). The first possible reason lies in probability estimation. That is, the probabilities that humans and neural LMs assign to words given their context differ in some cases, resulting in the failure of LM surprisal to explain garden-path effects. The other possible reason lies in surprisal theory itself. That is, the processing difficulty of garden-path sentences for humans cannot be reduced to predictability. Several recent studies argue for the second possibility (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024).

In this paper, we pursue the first possibility by asking whether it is possible to construct an LM that can explain garden-path effects via surprisal. Specifically, while previous work has primarily evaluated off-the-shelf pretrained neural LMs, we fine-tune these LMs on garden-path sentences to align surprisal-based reading time estimates with actual human reading times, and evaluate both the success of this approach and its impact on predictive power for reading times on naturalistic corpora (Sections 3 and 4). Our results show that fine-tuning succeeds in explaining garden-path effects without degrading (and in fact improves) predictive power for naturalistic reading times, providing an existence proof for an LM that can explain garden-path effects via surprisal (Section 5). Further anal-

yses demonstrate that fine-tuned LMs generalize to unseen garden-path types, and 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 broader implications for surprisal theory of our existence proof (Section 7).

2 Background

2.1 Surprisal Theory

Surprisal theory states that the processing cost of an input in human sentence processing scales linearly with its negative log-probability given the prior context:

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

Surprisal theory has several theoretical justifications. Most notably, Levy (2008) showed that surprisal is equivalent to the degree of belief updating about the latent sentence structures T:

$$\begin{aligned} &-\log p_{\text{human}}(w_t \mid \boldsymbol{w}_{<t}) \\ &= D_{\text{KL}}\big(p_{\text{human}}(T \mid \boldsymbol{w}_{\leq t}) \,\|\, p_{\text{human}}(T \mid \boldsymbol{w}_{<t})\big). \end{aligned} \tag{1}$$

In general, surprisal theory is a computational-level hypothesis in Marr's three levels of description (Marr, 1982), providing a characterization of the goal that the human sentence processing system should achieve (Hale, 2014). It is important to note that this theory posits that one goal of the human sentence processing system is prediction, which simultaneously corresponds to belief updating about latent sentence structures, while remaining neutral about the representations, algorithms, and implementations that realize this goal.

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.[1] 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.

[1]The flow of this subsection largely follows the literature review of Levy (2013).

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 processing (Hale, 2001, Principle 2). Empirical results have shown that LM surprisal indeed correlates strongly with human reading times and neural responses, providing substantial support for surprisal theory (Smith and Levy, 2013; Frank et al., 2015; Wilcox et al., 2023, *inter alia*).

However, recent findings have revealed systematic discrepancies. For instance, larger and more sophisticated LMs exhibit *worse* psychometric fit to 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 corpus-level objectives does not necessarily improve alignment with human processing difficulty. Another particularly striking manifestation of LM-human misalignment, observed consistently across models of varying sophistication, is the failure to capture garden-path effects, which we discuss in the following subsection.

2.3 Garden-Path Effect

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) (Frazier, 1979; Bever, 2013):

(1) a. The horse raced past the barn *fell*...
b. The horse <u>that was</u> 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, or (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 difficulty—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 Levy's formulation of surprisal theory (Equation 1), this processing cost should be modeled as belief updating about sentence structure triggered by syntactic disambiguation, 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 (van Schijndel and Linzen, 2021; Huang et al., 2024). This has led researchers to argue not only that LM next-word probabilities fail to serve as proxies for human predictability but also that syntactic disambiguation difficulty may be irreducible to predictability.

3 Methods

We apply the fine-tuning method of Kiegeland et al. (2024) to off-the-shelf pretrained neural LMs on garden-path sentences.

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: pair ID $s(d)$, ambiguity type $g(d) \in \{\text{MVRR}, \text{NPS}, \text{NPZ}\}$,[2] 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 $r(d) \in \{0, 1, 2, \text{null}\}$, 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 region of interest.

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 share no overlap in the verbs that induce syntactic ambiguity (e.g., *raced* in Example (1)) or in words within the region of interest. This ensures that the evaluation tests whether the fine-tuned LMs generalize to unseen data points.

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 region of interest. $D^{(-)}$ serves as the target for reading time estimation, as spillover variables are undefined for the first two words of a sentence and sentence-final words may reflect wrap-up effects (Just and Carpenter, 1980). $D^{(--)}$ is used for regression coefficient estimation. Under the assumption of surprisal theory that negative log-probability is linearly related to reading times, coefficients estimated on "ordinary" reading times—those unaffected by syntactic disambiguation—should also account for reading times in the region of interest where disambiguation occurs (Smith and Levy, 2013; van Schijndel and Linzen, 2021).

Fine-Tuning 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 ($k = 1, 2$, for 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 pairs from each ambiguity type. For each B, we estimate regression coefficients $\boldsymbol{\beta}_{\theta,B^{(--)}}$ via ordinary least squares:

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

where $X_{\theta,B^{(--)}}$ denotes the design matrix with rows $\mathbf{x}_\theta(d)^\top$ for $d \in B^{(--)}$, and $\boldsymbol{\psi}_{B^{(--)}}$ denotes the vector of reading times RT_d for $d \in B^{(--)}$.

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

We then compute the following loss:

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

The first term is the primary objective, measuring the squared difference between actual reading times and their estimates. The second term is a regularization penalty that prevents the regression coefficients from deviating excessively from the initial coefficients estimated using the initial LM parameters θ_0 on $D_{\text{gp}}^{\text{train}(--)}$.[3]

Evaluation We evaluate from two perspectives:

Garden-Path Effect Alignment We compute regression coefficients $\boldsymbol{\beta}_{\theta, D_{\text{filler}}^{(-)}}$ on $D_{\text{filler}}^{(-)}$ via ordinary least squares as in the fine-tuning procedure. We then evaluate the alignment between the estimated reading time difference

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

for ambiguous versus unambiguous sentences in $D_{\text{gp}}^{\text{test}}$, and the actual reading time difference $\Delta RT_{g,r}$ (van Schijndel and Linzen, 2021). Here, S_g denotes the set of test pairs for ambiguity type g, and $d(s, c, r)$ denotes the data point corresponding to pair s, condition c, and region r.

Impact on Naturalistic Sentences Using regression coefficients estimated on $D_{\text{nat}}^{(-)}$, we evaluate the per-datapoint log-likelihood improvement of a regression model including surprisal as a predictor over a baseline model with control variables only (Wilcox et al., 2020):

$$\Delta\text{llh}(\theta) = \frac{1}{|D_{\text{nat}}^{(-)}|} \times \sum_{d \in D_{\text{nat}}^{(-)}} \left[\log f(RT_d \mid \mathbf{x}_\theta(d); \boldsymbol{\beta}_{\theta, D_{\text{nat}}^{(-)}}) - \log f(RT_d \mid \mathbf{z}(d); \boldsymbol{\beta}_{\emptyset, D_{\text{nat}}^{(-)}})\right].$$

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

Here, $f(\cdot \mid \cdot; \boldsymbol{\beta})$ denotes the probability density function of the regression model with coefficients $\boldsymbol{\beta}$, and the subscript $\emptyset$ indicates coefficient estimation using control variables only.

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 (Wolf et al., 2020) implementation.[4] Prior work on naturalistic corpora has shown that 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).[5] This dataset contains 24 pairs for the following three ambiguity types.

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 ambiguity type 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.

[4] https://huggingface.co/openai-community/gpt2
[5] https://github.com/caplabnyu/sapbenchmark

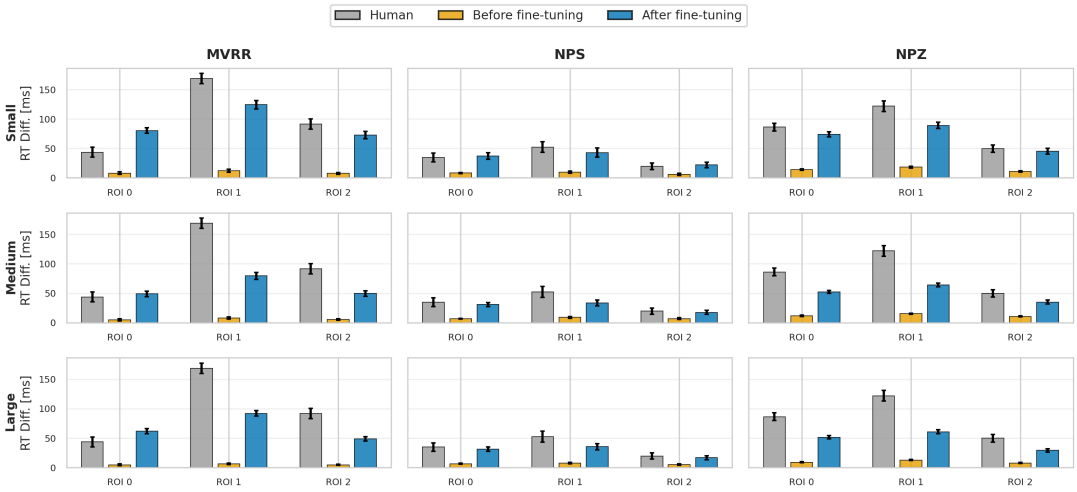

Figure 1: Garden-path effect alignment before and after fine-tuning. Rows indicate model sizes and columns indicate ambiguity types, with the x-axis representing the region of interest position and the y-axis representing the reading time difference (ms) between ambiguous and unambiguous conditions. 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.

The word *remained* disambiguates ($r = 0$).

Each word is annotated with self-paced reading times from 220–440 anonymized native English speakers. We exclude observations below 100 ms or above 3000 ms, and use the mean reading time across subjects as the representative value. 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 ambiguity types (three pairs total) as test data and construct the training set from the remaining pairs such that it satisfies the non-overlap constraint (Section 3). After excluding one pair containing data errors, we perform 23 evaluations and report the average across folds.[6]

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

Fine-Tuning 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. 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, it substantially underestimates their magnitude, replicating previous work (van Schijndel and Linzen, 2021; Huang et al., 2024). For example, at ROI=1 (the primary focus of analysis in prior work as the "Effect of Interest"), 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.

[6]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.

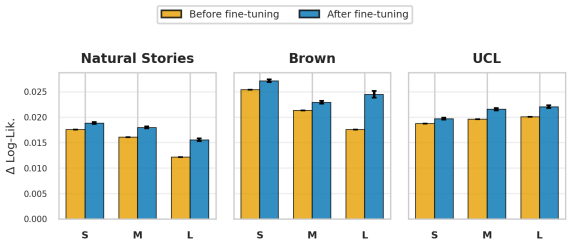

Figure 2: Impact of fine-tuning on predictive power for naturalistic corpora. Each graph corresponds to a naturalistic dataset, with the x-axis representing model size and the y-axis representing the log-likelihood improvement over the baseline regression model with control variables only.

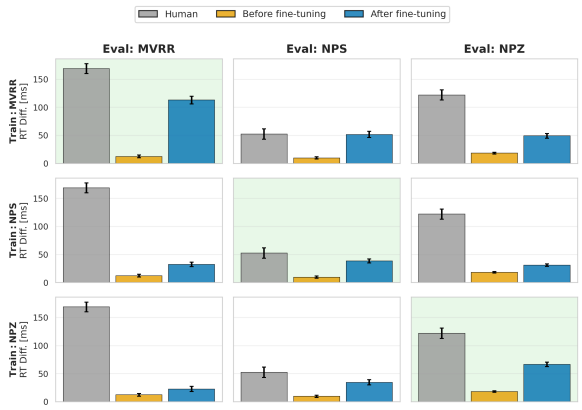

Figure 3: Cross-phenomenon transfer for GPT-2 small at ROI=1. Rows indicate the phenomenon used for fine-tuning, and columns indicate the phenomenon used for evaluation, with a green background highlighting in-domain evaluation.

In contrast, surprisal from fine-tuned LMs shows substantially improved alignment with human reading time slowdowns, demonstrating generalizability to unseen data points. 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 relative magnitude of slowdowns across garden-path phenomena, while pre-fine-tuned LMs failed to match the human ordering (MVRR>NPZ>NPS) at the Effect of Interest (ROI=1), instead showing NPZ>MVRR>NPS, fine-tuned LMs exhibited slowdown magnitudes consistent with human data.

Impact on Naturalistic Sentences Figure 2 shows the results for the impact on naturalistic sentences. Across all corpora and all model sizes, 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 naturalistic corpora that predominantly contain simple sentences.

6 Analysis

6.1 Cross-Phenomenon Transfer

In Section 5, we fine-tuned LMs using all three ambiguity types. In this subsection, we fine-tune LMs on a single ambiguity type and evaluate them on all types, to explore whether the models acquire phenomenon-specific patterns or learn more general mechanisms underlying garden-path effects.

Figure 3 shows the results for GPT-2 small at the Effect of Interest (ROI=1). First, regarding in-domain performance, LMs fine-tuned on a single phenomenon 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-tuning on all three phenomena (Figure 1).[7]

Crucially, regarding cross-phenomenon transfer, LMs fine-tuned on one ambiguity type better captured human reading time slowdowns on other ambiguity types compared to the pre-fine-tuned baseline. For example, the LM fine-tuned on MVRR predicted slowdowns of 51.5ms (baseline: 9.6ms) for NPS and 48.9ms (baseline: 18.1ms) for NPZ. While the transfer is not perfect—with predictions for phenomena different from the training target consistently smaller than those from LMs fine-tuned on that phenomenon—this result shows that fine-tuned LMs demonstrate generalizability to unseen garden-path types.

Regarding log-likelihood on naturalistic corpora, single-phenomenon fine-tuning generally yielded improvements or maintained performance on Natural Stories and Brown across most model sizes; slight decreases were observed on UCL for medium and large models (Appendix B).

6.2 An Unsuccessful Example: Subject/Object Relative Clauses

One potential concern is that the current method simply allows the model to learn to simulate any

[7] Medium and large models showed broadly similar trends (Appendix B).

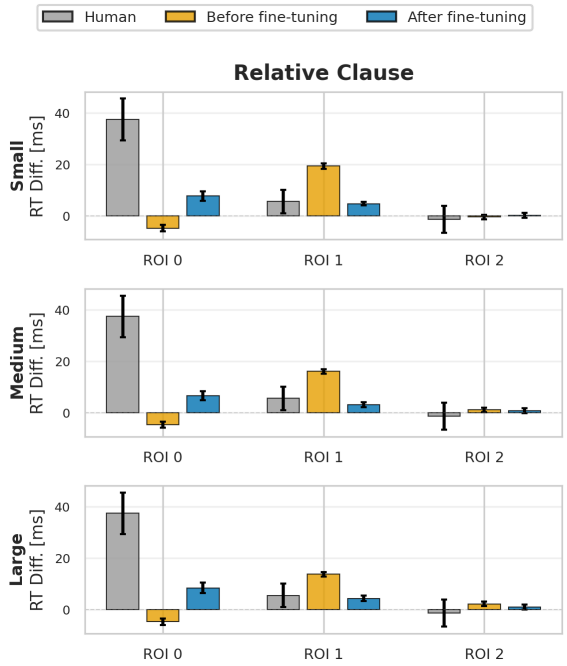

Figure 4: Subject/object relative clause asymmetry alignment before and after fine-tuning

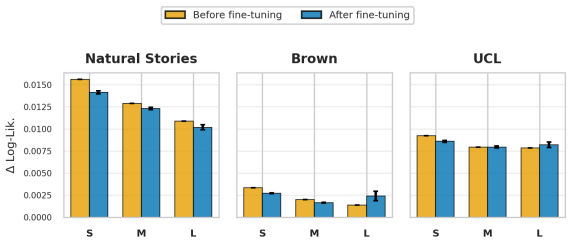

Figure 5: Impact of subject/object relative clause asymmetry alignment fine-tuning on predictive power for naturalistic corpora

kind of processing difficulty. If so, that would undermine the interpretation that there is an 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.

One phenomenon considered difficult to explain under surprisal theory is the processing difficulty pattern observed in English subject relative clauses (SRCs) versus object relative clauses (ORCs) (Levy, 2008, 2013; Levy and Gibson, 2013):

(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 the longer reading time at the ORC verb compared to the SRC verb since the ORC subject provides additional context that makes the verb more predictable. 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 (Gibson, 2000; Grodner and Gibson, 2005).

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. 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 under the same LOO setting (Section 4).

Figure 4 shows the results. First, consistent with prior work, pre-fine-tuned LMs predict a *speed-up* in reading time (-14% across all 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). Fine-tuned LMs successfully 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,[8] 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, which showed relatively high variance.

These results demonstrate that, under conditions in which surprisal theory is considered inadequate, the fine-tuning method struggles to reproduce hu-

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

man reading time differences and comes at the cost of predictive power for naturalistic reading times.

7 Discussion

Relating Reanalysis and Belief Updating Our results show that fine-tuning succeeds without degrading (and in fact improves) predictive power for human reading times on naturalistic corpora, providing an existence proof for an LM that can explain both garden-path effects and naturalistic reading times via surprisal.

Some prior studies challenging surprisal theory have argued that (i) the fully parallel maintenance of the latent sentence structures assumed by surprisal theory (Equation 1) lacks psychological reality (Huang et al., 2024), and consequently, that (ii) processing difficulty in garden-path sentences stems from a selective mechanism of reanalysis to structures that were not maintained at the point of disambiguation (Fodor and Ferreira, 1998), inducing a cost that cannot be reduced to predictability.

Does our finding—the existence of an LM that can explain garden-path effects via its surprisal—reject the reanalysis account? We argue it does not; rather, reanalysis and surprisal can compatibly coexist as descriptions at different levels of analysis (Marr, 1982). Regarding point (i), Levy (2008) noted from the outset that surprisal theory does not necessarily commit to the psychological reality of fully parallel representations. Regarding point (ii), we take Levy's view further and argue that, given that surprisal theory is a computational-level hypothesis, it is possible to assign non-zero probability to structures that will serve as candidates during reanalysis in advance, without committing to the psychological reality of such representations—as long as this provides a sufficient description of the human sentence processing system. Under this view, reanalysis can implement belief updating when the parse distribution undergoes substantial change, operating as an algorithmic-level process.

Therefore, we believe that the critical question going forward is to investigate under what conditions reanalysis occurs and how it relates to belief updating. This requires moving beyond purely computational-level claims to examine the algorithmic-level mechanisms.

Low Falsifiability of Surprisal Theory Our results show that there exists a *probability distribution* that can explain both garden-path effects and naturalistic reading times. Does this finding provide proof-of-concept for surprisal theory, a *predictability*-based account? From an optimistic perspective, yes; from a more critical perspective, our findings instead highlight that surprisal theory cannot be easily falsified within a simple correct-or-incorrect binary framework. This is because (i) unless the theory specifies how to determine the probability distribution, surprisal theory itself merely posits the existence of *some* probability distribution that describes human processing difficulty, and (ii) the modern LM framework allows one to construct a desired probability distribution (Bowers and Mitchell, 2025)—even though our additional analysis on SRC/ORC asymmetry fortunately reveals some limitations to this flexibility.

Given this theoretical landscape, we believe the productive research direction is not to test whether to reject surprisal theory as a binary hypothesis but rather to investigate what probability distributions best capture human sentence processing and what properties characterize such distributions. This shift in focus leads to a reinterpretation of prior findings.

Previous work showing that surprisal from various LMs fails to explain garden-path effects has often been framed as counterevidence to surprisal theory (van Schijndel and Linzen, 2021; Huang et al., 2024; Timkey et al., 2025). However, given the low falsifiability discussed above, such results may be better interpreted as reflecting divergence between different probability distributions: those describing human sentence processing versus those learned from corpora. Importantly, we believe this reframing does not diminish the value of such research. Investigations of the differences between these probability distributions should provide valuable insights into human sentence comprehension, regardless of whether they are positioned as tests of surprisal theory.

8 Conclusion

In this paper, we asked whether it is possible to construct an LM that can explain garden-path effects via surprisal. Our results provide an existence proof for an LM that can explain both garden-path effects and naturalistic reading times via surprisal. Based on our results, we discussed the possibility of relating reanalysis and belief updating, encouraging the community to move beyond computational-level claims to examine algorithmic-level mechanisms.

Limitations

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

Our study focuses on demonstrating the existence of an LM that can explain both garden-path effects and naturalistic reading times via surprisal, but does not investigate what changes occur in the internal mechanisms of LMs 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 model representations and processing mechanisms could provide valuable insights into the algorithmic-level implementations of surprisal theory.

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. 2013. The cognitive basis for linguistic structures. In *Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures*, page 0. Oxford University Press.

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

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.

Lyn Frazier. 1979. ON COMPREHENDING SENTENCES: SYNTACTIC PARSING STRATEGIES. *Doctoral Dissertations*, pages 1–243.

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.

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.

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.

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.

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.

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.

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.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

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. For hyperparameters with two values separated by a slash, the first value corresponds to training with three ambiguity types, and the second value corresponds to training with a single ambiguity type. The total computational cost required for all experiments was approximately 40 GPU hours (NVIDIA RTX 6000 Ada, 48GB).

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 regularization term λ	100

Table 1: Fine-tuning hyperparameters

B Full Results of Cross-Phenomenon Transfer

Figure 6 shows cross-phenomenon transfer results for all GPT-2 model sizes at ROI=1.

C Licenses

Table 2 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.

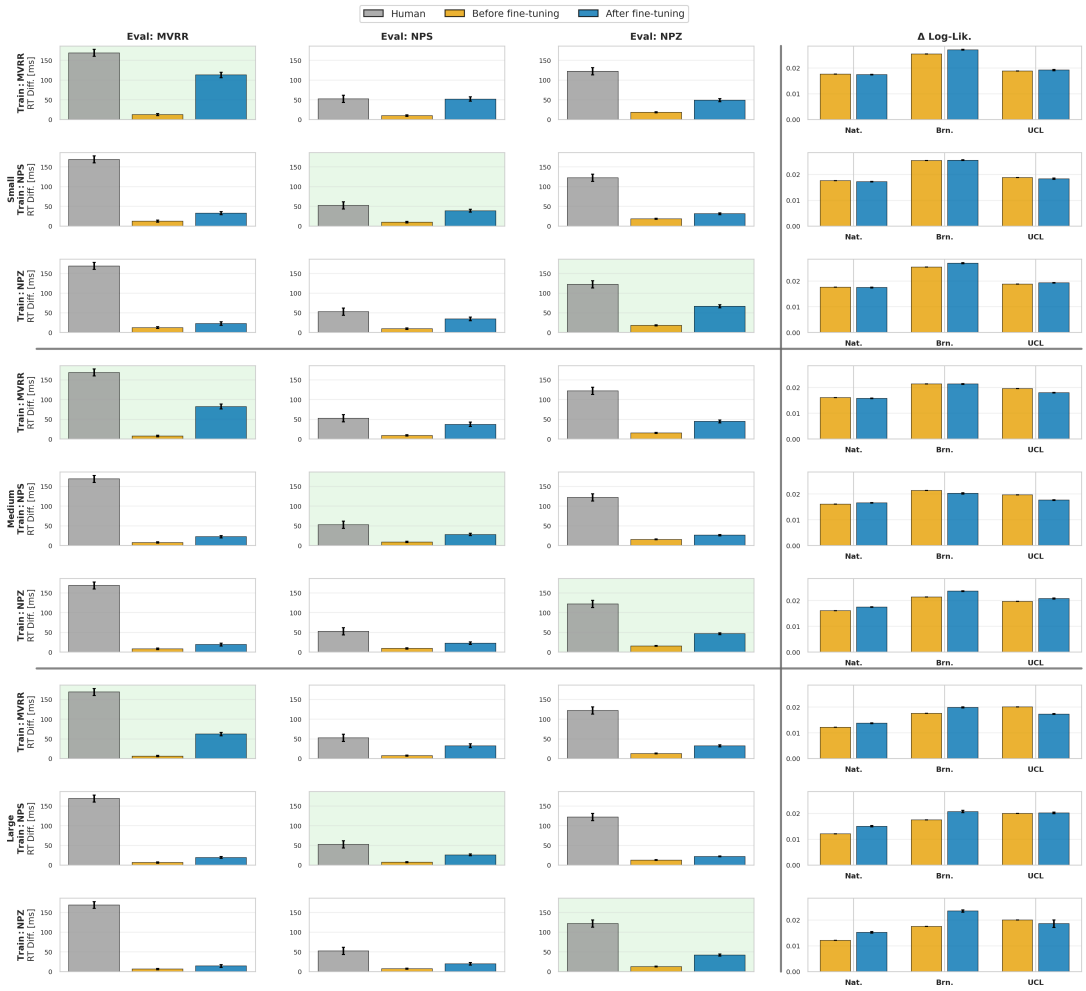

Figure 6: Cross-phenomenon transfer results for all GPT-2 model sizes at ROI=1. Left: Each panel corresponds to a model size (`small`, `medium`, `large`). Within each panel, rows indicate the phenomenon used for fine-tuning, and columns indicate the phenomenon used for evaluation, with a green background highlighting in-domain evaluation. The y-axis shows estimated reading time differences (ms) between ambiguous and unambiguous conditions. Right: Changes in log-likelihood (Δ Log-Lik.) on naturalistic corpora before and after fine-tuning (Nat.: Natural Stories, Brn.: Brown, UCL).

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

Table 2: Licenses of datasets and tools