Garden Path Recovery in Causal and Masked Language Models

Sanjan Baitalik and Rajashik Datta
Institute of Engineering & Management, Kolkata, West Bengal, India
{sanjan.baitalik2022, rajashik.datta2022}@iem.edu.in

Abstract

Garden-path sentences offer a controlled probe of English incremental sentence processing because they require a reader to revise an initially plausible parse when a later region disambiguates the structure. We present an architecture-aware comparison of garden-path recovery in causal and masked language models using 100 English garden-path/control pairs (200 sentences) spanning three constructions: NP/Z, where a noun phrase is initially read as a direct object but must be reanalyzed as the subject of a zero-complement clause; NP/S, where a noun phrase must be reanalyzed as the subject of an embedded sentence; and MV/RR, where an apparent main verb must be reanalyzed as a reduced relative modifier. Causal models are evaluated with left-to-right word surprisal, whereas masked models are evaluated with pseudo-surprisal derived from masked language model scoring. Beyond the disambiguating word, we analyze cumulative excess surprisal, area-under-curve recovery summaries, and layer-wise hidden-state divergence between each garden-path sentence and its minimally different control. Across the audit-valid model set, causal models show larger within-model disambiguation effects than masked models overall, with the clearest family-level difference on NP/Z constructions. We interpret this difference cautiously because surprisal and pseudo-surprisal are not numerically commensurable across architectures or tokenizers. The results nevertheless show that architecture changes which recovery signals are observable: decoder-only models exhibit sharper online disruption at the point of syntactic revision, while bidirectional encoders appear comparatively buffered at the disambiguator due to right-context access. More broadly, the findings argue that garden-path evaluation should emphasize *recovery dynamics*, not merely end-state plausibility or task accuracy.

1 Introduction

Garden-path sentences remain a central probe in psycholinguistics because they expose the interaction between structural expectation, lexical information, and reanalysis during incremental comprehension (Frazier and Rayner, 1982; Trueswell et al., 1994; MacDonald et al., 1994; Levy, 2008). This paper studies English garden-path sentences. The language matters: the syntactic configurations that produce garden-path effects are not universal across languages, and our claims are therefore restricted to English constructions and English pretrained models.

To make the phenomenon concrete, consider the NP/S sentence *The girl knew the answer was wrong all along*. Immediately after *knew the answer*, a reader can adopt a transitive-object analysis: the girl knew *the answer*. The later word *was* disrupts that analysis, because *the answer* must instead be reinterpreted as the subject of an embedded sentence: the girl knew *that the answer was wrong*. The phrase *all along* then reinforces the reinterpreted meaning, namely that the girl's knowledge concerned the wrongness of the answer throughout, not merely that she possessed the answer. This temporary commitment, disruption, and repair is the process we refer to as garden-path recovery.

Recent work has shown that masked models such as BERT can exhibit garden-path-like effects in comprehension-oriented settings (Irwin et al., 2023), and that autoregressive models such as GPT-2 display informative hidden-state geometry around disambiguation points (Jurayj et al., 2022). Recent studies further investigate lingering misinterpretations and reanalysis in LLMs through question answering, parse probes, attention, and mechanistic analyses (Li et al., 2024; Hanna and Mueller, 2025; Amouyal et al., 2025). At the same time, newer evidence suggests that larger autoregressive systems do not necessarily resolve temporary am-

biguity more successfully (Cao and Schuler, 2025), and broader psycholinguistic work has warned that lower perplexity does not automatically yield more human-like processing signatures (Oh and Schuler, 2023b,a; Liu et al., 2024). Taken together, these findings motivate a more careful question than whether language models merely "get the sentence right": how do different architectures *recover* once the sentence forces a structural revision?

This paper addresses that question by comparing causal and masked language models under architecture-appropriate scoring. For decoder-only models, we use left-to-right word surprisal at and after the disambiguating word. For encoder-only models, we use pseudo-surprisal from masked language model scoring, which avoids treating bidirectional encoders as if they were incremental left-to-right parsers (Salazar et al., 2020). We then compare models using *paired garden-path/control sentences*: each garden-path sentence is matched to a minimally different control sentence, and the score difference between the two forms the main effect size. We call the single-word difference at the disambiguator the *disambiguation spike*; we call the sum and area-under-curve of downstream differences *post-disambiguation recovery summaries*; and we use *hidden-state divergence* to measure how far the internal representations of the paired sentences separate at each layer. A model is described as *audit-valid* only if its outputs pass non-finite, repeated-value, token-alignment, and batch-invariance checks on the canonical dataset.

Empirically, the clearest result in our final audited run is a robust family-level separation on NP/Z constructions. Across audit-valid model–item observations, causal models show larger disambiguation spikes than masked models overall, and this difference is driven primarily by NP/Z items. GPT-2-medium and GPT-2 produce the largest causal-model spikes, while RoBERTa-base and BERT-base-uncased show clear but smaller NP/Z effects. DeBERTa-v3-base shows smaller and more variable evidence of disruption on the same construction. In addition, hidden-state divergence peaks very shallowly for the masked models but deeper for the causal models once the embedding layer is excluded.

Our claim is therefore not that one family is universally more syntactically competent, nor that raw surprisal values are directly comparable across masked and causal objectives. Rather, the results indicate that architecture changes what counts as a meaningful recovery signal. Decoder-only models appear to experience sharper online disruption at the disambiguator, whereas bidirectional encoders often appear more stable at that point, plausibly because right-context access helps retrospectively regularize the sentence representation. This framing, we argue, is a more faithful basis for comparing language model behavior on garden-path phenomena.

2 Related Work

Research on garden-path processing in humans has long emphasized that temporary ambiguity is resolved through an interaction between structural bias and lexical-semantic constraints. Early eye-tracking and self-paced reading work established the classical garden-path effect and showed that readers incur measurable costs when their initial parse must be revised (Frazier and Rayner, 1982). Subsequent work demonstrated that these effects are modulated by thematic-role expectations and lexical information rather than by purely syntax-first mechanisms alone (Ferreira and Henderson, 1990; Trueswell et al., 1994; MacDonald et al., 1994). Expectation-based theories later formalized these ideas in probabilistic terms, connecting processing difficulty to the surprisal of disconfirming incoming material (Hale, 2001; Levy, 2008; Smith and Levy, 2013).

In NLP, a large literature has used controlled syntactic tests to study whether neural language models encode grammatical regularities. Early targeted evaluations examined subject–verb agreement and other syntax-sensitive dependencies (Marvin and Linzen, 2018; Wilcox et al., 2018). Challenge sets such as BLiMP extended this evaluation paradigm to broader classes of minimal syntactic contrasts (Warstadt et al., 2020), and subsequent work refined how targeted syntactic evaluations should be interpreted (Newman et al., 2021). Probing studies have also highlighted that interpretability claims require caution, since probe success can partly reflect semantic shortcuts rather than cleanly isolated syntax (Maudslay and Cotterell, 2021; Stańczak et al., 2022; Bazhukov et al., 2024).

A more directly relevant thread has begun applying psycholinguistic methodology to modern language models. Futrell et al. (2019) explicitly treat neural language models as psycholinguistic subjects, showing that controlled experimental stimuli can reveal incremental syntactic state in model be-

havior. For masked models, Irwin et al. (2023) find that BERT-family systems show garden-path effects in question answering and semantic-role probing. For causal models, Jurayj et al. (2022) show that GPT-2 hidden states undergo measurable geometric shifts around garden-path disambiguation points. Li et al. (2024) study lingering misinterpretations in LLMs using comprehension questions, parse-shift probes, and attention visualization, while Hanna and Mueller (2025) use sparse autoencoders to ask whether autoregressive transformers represent syntactic features, maintain multiple readings, and repair initial representations during garden-path processing. Newer work broadens this picture: Cao and Schuler (2025) show that larger autoregressive models are not necessarily better at resolving temporary ambiguity, while Amouyal et al. (2025) compare humans and language models on garden-path comprehension and report partially overlapping but non-identical difficulty profiles.

Finally, our methodology is informed by recent work on psycholinguistic scoring for neural language models. For masked encoders, pseudo-log-likelihood and pseudo-perplexity provide a principled way to score full sentences without imposing a left-to-right factorization that the architecture does not support (Salazar et al., 2020). At the same time, pseudo-surprisal is not the same measurement as causal next-token surprisal: the conditioning context, tokenization, and normalization differ across model families. Several studies also caution that raw surprisal from larger or better-optimized language models may drift away from human processing signals even as standard NLP performance improves (Oh and Schuler, 2023a,b; Liu et al., 2024; Wang et al., 2025; Liu and Ding, 2025). Our study builds on these insights by combining architecture-aware scoring, matched English garden-path stimuli, and hidden-state analyses in a single comparison.

3 Materials & Methods

3.1 Stimuli

We evaluate models on 100 English garden-path/control pairs, for a total of 200 sentences. The benchmark covers three classical garden-path constructions: 41 NP/Z pairs, 35 NP/S pairs, and 24 MV/RR pairs. Each garden-path item is paired with a minimally matched non-garden-path control, allowing pairwise computation of disruption at the disambiguating word and during the post-disambiguation window. The dataset is not presented as a new standalone psycholinguistic benchmark; rather, it is a controlled stimulus set curated from standard garden-path construction types and validated for the architecture-aware recovery analysis introduced here. The final stimulus table retains all 200 rows after validation, with balanced garden-path and control conditions within each construction type.

Type	Definition	Illustrative example
NP/Z	Noun phrase vs. zero-complement clause.	*While the man hunted the deer ran into the woods.* The phrase *the deer* first looks like the object of *hunted*; *ran* forces it to be the subject of the following clause.
NP/S	Noun phrase vs. sentential complement.	*The girl knew the answer was wrong all along.* The phrase *the answer* first looks like the object of *knew*; *was* forces the reading *knew that the answer was wrong*.
MV/RR	Main verb vs. reduced relative.	*The horse raced past the barn fell.* The word *raced* first looks like the main verb; *fell* forces *raced past the barn* to modify *horse*.

Table 1: Construction labels and linguistic intuition. Examples are illustrative rather than necessarily identical to the experimental items.

Table 1 defines the three construction labels used throughout the paper. NP/Z refers to a noun phrase/zero-complement ambiguity: after a verb, a noun phrase can initially be analyzed as the verb's direct object, but a later finite verb reveals that the noun phrase actually begins a clause without an overt complementizer. NP/S refers to a noun phrase/sentence-complement ambiguity: a noun phrase after a verb can first be analyzed as an object, but later material reveals it as the subject of an embedded sentence. MV/RR refers to a main-verb/reduced-relative ambiguity: a past-tense or participial form is initially read as the main verb of the sentence, but a later verb forces it to be reanalyzed as part of a reduced relative clause.

The analysis focuses on a designated disambiguation region for each pair. For every sentence, we record the word-level disambiguator index, align tokenizer pieces back to whitespace-level words, and aggregate subword surprisals to word-

level values. Alignment diagnostics in the final run show complete success for the audit-valid causal models, ensuring that the disambiguator statistics are computed on the intended lexical region rather than on a tokenization artifact.

3.2 Models

We evaluate seven pretrained English language models. The causal family initially includes GPT-2, GPT-2-medium, Pythia-160M, and Pythia-410M (Radford et al., 2019; Biderman et al., 2023). The masked family includes BERT-base-uncased, RoBERTa-base, and DeBERTa-v3-base (Devlin et al., 2019; Liu et al., 2019; He et al., 2021). In the final audited analysis, GPT-2 and GPT-2-medium pass all causal audits and are retained as audit-valid causal models, while the two Pythia models are quarantined because repeated-value audits indicate unstable fallback-like behavior. The masked family contributes all three models to the main analyses.

3.3 Architecture-aware scoring

A central design choice of this paper is that causal and masked models are not forced into the same token scoring scheme. Figure 2 summarizes the full pipeline. For causal models, we compute standard left-to-right next-token surprisal and aggregate subword pieces to word-level surprisal. For masked models, we use pseudo-surprisal obtained by masking each token in turn and evaluating the log-probability assigned to the original token under full bidirectional context (Salazar et al., 2020). This makes the evaluation architecture-aware: both model families are scored according to the conditioning structure they actually implement.

This design does not make causal surprisal and masked pseudo-surprisal numerically interchangeable. Decoder-only scores condition only on the left context; masked-model scores condition on both left and right context; and tokenization differs across model families. Consequently, we do not interpret a raw score of, for example, 4.0 in a causal model as the same quantity as a pseudo-surprisal of 4.0 in a masked model. The main inferential object is instead the within-model garden-path-minus-control difference over matched items. Cross-family statistics are reported as aggregate contrasts over these within-model deltas, and should be read as evidence about relative recovery profiles under architecture-appropriate scoring, not as direct comparisons of absolute probability scales.

For each garden-path/control pair and model, we compute a disambiguation spike defined as the garden-path-minus-control word-level score at the disambiguator. We also derive post-disambiguation summaries from the garden-path-minus-control trace in a fixed downstream window: cumulative excess surprisal and an area-under-curve summary. Because absolute scales differ across architectures and scoring objectives, our main comparisons emphasize within-model pairwise deltas and family-level contrasts over matched items rather than raw sentence probabilities.

3.4 Hidden-state divergence

To move beyond output probabilities, we also compare internal representations for ambiguous and control sentences. At each layer, we compute cosine-distance-based divergence at the disambiguator between the ambiguous sentence representation and its matched control counterpart. We summarize both the value at the disambiguator and the layer of maximum divergence. Since embedding-level distances can reflect static lexical differences rather than processing dynamics, we treat layer 0 separately and report a second summary excluding the embedding layer.

3.5 Statistics and audit criteria

For structure-specific disambiguation spikes, we report means, bootstrap 95% confidence intervals, and two-sided tests on pairwise deltas. For family-level comparisons, we report both Welch tests and Mann–Whitney tests over audit-valid model–item observations. We treat the family-level statistics as evidence about aggregate architectural tendencies, while the pairwise model summaries remain the primary basis for interpretation.

An audited analysis means that model outputs are screened before they can support main claims. The non-finite check rejects missing, NaN, or infinite surprisals at required target words. The repeated-value audit detects fallback-like constants that could arise from a scoring bug rather than model behavior. The alignment check verifies that the disambiguating word maps to the intended tokenizer span. The batch-invariance check verifies that scoring the same sentence alone or in a padded batch yields equivalent word-level scores within numerical tolerance. A model that fails any required audit remains available for debugging but is excluded from paper-facing tables and figures.

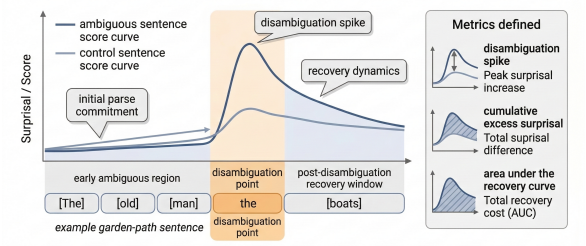

Figure 1: Conceptual token-level view of a garden-path sentence. The highlighted region marks the disambiguation point, and the shaded segment indicates the post-disambiguation recovery window used by our metrics.

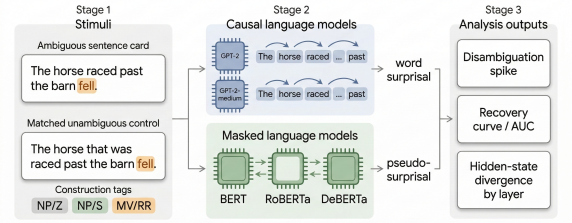

Figure 2: Conceptual overview of the evaluation pipeline. Causal models are scored with left-to-right surprisal, masked models with pseudo-surprisal, and both are compared through matched garden-path/control deltas, recovery summaries, and hidden-state divergence.

4 Algorithm(s)

Although our contribution is empirical rather than algorithmic, the analysis relies on a concrete architecture-aware procedure. Algorithm 1 summarizes the pipeline.

The audit step is important because garden-path analyses are highly sensitive to token alignment and scoring artifacts. In our final pipeline, a model contributes to main claims only if it passes non-finite checks, repeated-value audits, alignment checks, and batch-invariance checks on the canonical 200-row dataset. Figure 1 gives an intuitive view of the recovery quantities used later in the paper: the disambiguation spike is measured at the highlighted revision point, while the cumulative and AUC-based summaries are computed over the post-disambiguation window.

5 Experiments & Results

5.1 Main disambiguation effects

Figure 3 gives the most important empirical result. Across audit-valid model–item observations, causal models exhibit larger disambiguation spikes than masked models overall, with a mean of 3.09 for the causal family versus 1.32 for the masked family (Welch's $t = 5.68$, $p < 0.001$; Mann–Whitney $p < 10^{-8}$). This overall difference is driven primarily by NP/Z constructions, where causal models average 6.18 and masked models average 2.51 (Welch's $t = 7.52$, $p < 0.001$; Mann–Whitney $p < 10^{-9}$). By contrast, the family gap is small and statistically unstable for MV/RR, and more modest for NP/S.

These family-level numbers are not interpreted as direct comparisons of absolute model probabilities. They compare within-model garden-path-minus-control deltas aggregated by architecture family. The result therefore supports a cautious

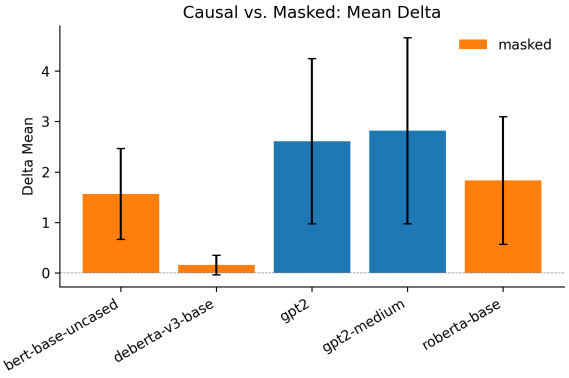

Figure 3: Family-level comparison of mean disambiguation spikes for audit-valid models. The clearest separation occurs on NP/Z constructions, where causal models show substantially larger disruption than masked models.

claim: under their respective scoring regimes, decoder-only models show larger relative disruption at the English disambiguator than bidirectional encoders, especially on NP/Z items.

At the model level, GPT-2-medium has the largest NP/Z disambiguation spike with a mean disambiguation spike of 6.50 (95% CI [5.71, 7.33]), followed by GPT-2 at 5.86 (95% CI [5.14, 6.57]). Among masked models, RoBERTa-base and BERT-base-uncased also show robust NP/Z effects, with means of 4.33 (95% CI [3.01, 5.64]) and 3.33 (95% CI [2.31, 4.38]), respectively. DeBERTa-v3-base behaves differently: its NP/Z mean is slightly negative at -0.15 with a confidence interval that comfortably spans zero, suggesting little reliable garden-path disruption on this construction in our setup. On NP/S and MV/RR, all retained causal models remain positive, but the effects are much smaller than on NP/Z. The masked models again show smaller and less stable patterns, especially RoBERTa on NP/S and MV/RR and DeBERTa

Algorithm 1 Architecture-aware garden-path recovery scoring

Require: Paired stimulus set $\mathcal{D}$ with garden-path/control sentences, model set $\mathcal{M}$, disambiguator index d, post-disambiguation window W

for each pair $(x^{gp}, x^{ctl}) \in \mathcal{D}$ **do**
  for each pair $(x^{amb}, x^{ctl}) \in \mathcal{D}$ **do**
    if m is causal **then**
      compute token log-probabilities with next-token shifting
      aggregate subword scores into word surprisal values
    else
      compute pseudo-surprisal by masking each token in turn
      aggregate masked-token scores into word-level values
    end if
    align tokenizer pieces to the word-level disambiguator d
    compute disambiguation spike $\Delta_d = S(x_d^{gp}) - S(x_d^{ctl})$
    compute post-disambiguation trace $\Delta_{d:d+W}$
    summarize recovery using cumulative excess and AUC statistics
    extract hidden states for garden-path and control sentences
    compute layer-wise cosine divergence at the disambiguator
  end for
  run audit checks for non-finite values, alignment, repeated-value artifacts, and batch invariance
  quarantine model m if any required audit fails
end for
aggregate statistics over audit-valid models and matched pairs

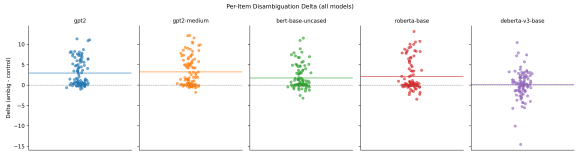

Figure 4: Per-item disambiguation deltas for all audit-valid models. The figure is useful for showing that the strongest NP/Z effects are not driven by a single outlier item but recur across the dataset.

across the board.

Figure 4 complements the family plot by showing item-level variability. The largest NP/Z effects are broadly distributed rather than being driven by a handful of pairs. This is especially clear for GPT-2 and GPT-2-medium, whose NP/Z deltas are consistently above zero across most items. The item-level plots also make the DeBERTa pattern visually salient: unlike the other audit-valid models, it shows a mixture of positive and negative responses on NP/Z.

5.2 Recovery summaries

We next examine whether the initial disruption extends beyond the disambiguator. Using our post-disambiguation summaries, GPT-2 and GPT-2-medium maintain mean recovery deltas of 4.79 and 4.87, respectively, on the full canonical dataset. Among masked models, RoBERTa-base and BERT-base-uncased remain clearly positive, whereas DeBERTa-v3-base again shows a comparatively small recovery summary, with a mean recovery delta of only 0.15. Because causal surprisal and masked pseudo-surprisal live on different absolute scales, we treat these recovery summaries mainly as within-family evidence rather than as direct numerical comparisons across architecture families. Even under that caveat, however, the results reinforce the same ranking already visible at the disambiguator: GPT-2-medium and GPT-2 show the largest decoder-side disruption summaries, while RoBERTa-base and BERT-base-uncased have the largest encoder-side recovery summaries and DeBERTa-v3-base shows the smallest and least consistent recovery summary among the audit-valid models.

Figure 5 further isolates the decoder-only comparison. The pattern is clean: GPT-2-medium exceeds GPT-2 on NP/Z and NP/S, while the two models remain relatively close on MV/RR. Since both passed the final non-finite, alignment, repeated-value, and batch-invariance audits, this decoder-only pattern can be interpreted directly rather than being treated as a debugging artifact.

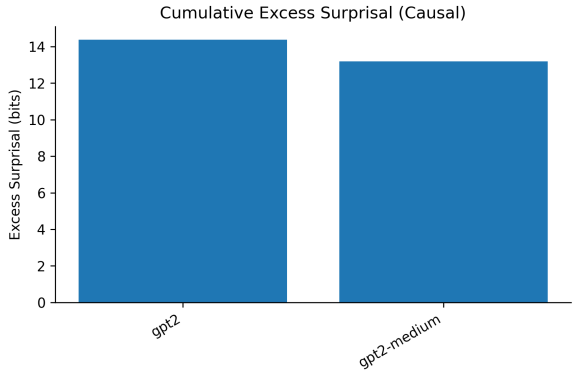

Figure 5: Comparison of disambiguation spikes among audit-valid causal models. GPT-2-medium is consistently strongest, especially on NP/Z.

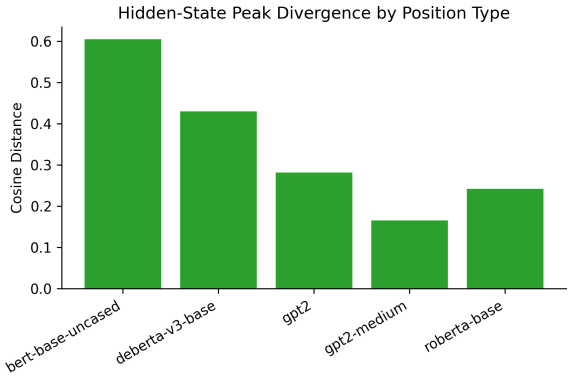

Figure 6: Summary of hidden-state divergence. Excluding the embedding layer, masked models peak very early, whereas the strongest causal models peak deeper in the network.

5.3 Hidden-state evidence

The hidden-state analysis reveals a complementary architectural difference. If layer 0 is included, the point of maximum garden-path/control divergence occurs at the embedding layer for 596 out of 700 model–item combinations, or 85.1% of the dataset. This is unsurprising: even matched garden-path/control sentence pairs retain some lexical differences that can dominate static embedding distance. More informative is the no-embedding summary. Excluding layer 0, BERT-base-uncased, RoBERTa-base, and DeBERTa-v3-base all peak essentially at layer 1, with mean peak layers of 1.18, 1.00, and 1.00. GPT-2 peaks deeper at layer 5.43, while GPT-2-medium peaks much deeper still at layer 17.33.

We interpret Figure 6 cautiously. The result does not imply that deeper is inherently more human-like; rather, it shows that the representational locus of garden-path separation differs by architecture. In our data, bidirectional encoders register the distinction early, while decoder-only models distribute the divergence more deeply across the stack. This pattern is consistent with the broader claim that architecture affects not only final sentence scoring but also *where* revision-like behavior becomes visible internally.

6 Ablation Studies

We report three ablation-style analyses that sharpen the interpretation of the main findings. First, we separate layer 0 from all higher layers in the hidden-state analysis. Without this separation, the representational story is dominated by lexical embedding differences, which are not an adequate proxy for online revision. Excluding the embedding layer reveals the much more interesting encoder–decoder contrast described above, with masked models peaking at the first contextual layer and causal models peaking substantially deeper.

Second, we compare multiple recovery summaries rather than relying only on the single-word disambiguation spike. The core qualitative ranking remains stable under this metric change: GPT-2 and GPT-2-medium retain the largest causal recovery summaries, BERT-base-uncased and RoBERTa-base remain the largest-effect masked models, and DeBERTa-v3-base remains the smallest and least consistent of the audit-valid models. This consistency is important because it suggests that the main claims are not an artifact of one sharply localized scoring decision.

Third, we perform a methodological robustness check through audit-based model exclusion. The initial causal candidate set included the two Pythia models, but both failed repeated-value audits and were therefore quarantined from main claims. We emphasize this exclusion not because it is theoretically interesting in itself, but because it demonstrates that garden-path comparisons are fragile to implementation artifacts if model outputs are not explicitly audited. The paper's main conclusions are therefore intentionally restricted to the final five model runs that passed all validation checks.

7 Discussion & Future Work

The main lesson of this study is that garden-path evaluation should distinguish between *online disruption* and *retrospective regularization*. Decoder-only models are scored under an incremental left-to-right regime and therefore expose sharp changes

when the disambiguator contradicts an initially plausible parse. Bidirectional encoders, by contrast, are scored with access to future context and can therefore appear comparatively buffered at that exact point, even when they still encode substantial garden-path/control differences elsewhere. This distinction helps reconcile why encoder models can appear competent on garden-path comprehension tasks while still yielding a different profile from autoregressive models in token-level analyses.

Our results also suggest that NP/Z constructions remain the clearest stress test for model revision. Both causal models and the largest-effect masked models show their most reliable effects on NP/Z, whereas MV/RR and NP/S yield smaller and more heterogeneous effects. This makes NP/Z a particularly attractive target for future mechanistic studies, intervention experiments, or multilingual extensions.

Several directions follow naturally from this work. One is to integrate the present model-side analysis with human behavioral data, allowing direct tests of whether decoder-only or encoder-only recovery traces better predict reading times or comprehension errors. Another is to extend the benchmark across languages and across broader model families, especially newer audited causal models. A third direction is to move from descriptive divergence to intervention-based analysis, testing whether editing or suppressing specific layers changes garden-path sensitivity. Finally, future work should connect token-level disruption to downstream behavior, for example through paraphrasing, question answering, or multimodal interpretation tasks, in the spirit of recent human-versus-model comparisons (Amouyal et al., 2025).

Ethics Statement

This work studies the linguistic behavior of publicly available pretrained language models on controlled psycholinguistic stimuli. The dataset contains no personal data. The main ethical concern is interpretive: language models should not be anthropomorphized merely because they exhibit some human-like processing signatures. Throughout the paper, we therefore treat model behavior as evidence about representational and probabilistic tendencies, not as evidence of human-equivalent understanding.

Limitations

This study has several limitations. First, the comparison is conducted only in English and only on three garden-path construction families. Although NP/Z, NP/S, and MV/RR are well motivated for English psycholinguistics, syntactic ambiguity is language-specific; the same construction labels and recovery patterns should not be assumed to transfer to languages with different word order, case marking, agreement, or complementizer systems. Second, masked and causal scores are architecture-appropriate but not numerically commensurable. Decoder-only surprisal is a left-to-right next-token quantity, while masked pseudo-surprisal is obtained by conditioning on bidirectional context and masking one token at a time. Tokenization also differs across model families. Our use of matched garden-path/control deltas reduces, but does not eliminate, this confound. Consequently, the causal–masked results should be read as differences in relative disruption profiles under different scoring regimes, not as evidence that one architecture assigns objectively higher processing difficulty on a shared probability scale. Third, the work does not include human reading-time or eye-tracking data in the current experimental loop, so the paper should be read as a model-comparison study rather than a direct human-model fit study. Fourth, the hidden-state divergence analysis is descriptive: it identifies where garden-path and control representations separate, but it does not on its own establish a mechanistic account of reanalysis. Finally, only two causal models survive the final audit. While that is sufficient to support the present architecture-aware comparison, broader architectural claims would benefit from a broader set of decoder-only models that pass the same validation suite.

References

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025. When the lm misunderstood the human chuckled: Analyzing garden path effects in humans and language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8235–8253.

Maxim Bazhukov, Ekaterina Voloshina, Sergey Pletenev, Arseny Anisimov, Oleg Serikov, and Svetlana Toldova. 2024. Of models and men: Probing neural networks for agreement attraction with psycholinguistic data. In *Proceedings of the 28th Confer-*

ence on Computational Natural Language Learning, pages 280–290.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.

Ziyuan Cao and William Schuler. 2025. Are larger language models better at disambiguation? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 155–164.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fernanda Ferreira and John M Henderson. 1990. Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):555.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long and Short Papers)*, pages 32–42.

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*.

Michael Hanna and Aaron Mueller. 2025. Incremental sentence processing mechanisms in autoregressive transformer language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3181–3203, Albuquerque, New Mexico. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. Bert shows garden path effects. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232.

William Jurayj, William Rudman, and Carsten Eickhoff. 2022. Garden path traversal in gpt-2. In *Proceedings of the fifth blackboxnlp workshop on analyzing and interpreting neural networks for nlp*, pages 305–313.

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

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. ArXiv:2405.16042; accepted by CogSci 2024.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the "right reasons"? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9598–9619.

Wei Liu and Nai Ding. 2025. Information integration in large language models is gated by linguistic structural markers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6903–6915.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1192–1202.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 3710–3723.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the association*

for computational linguistics: EMNLP 2023, pages 1915–1921.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2699–2712.

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

Karolina Stańczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598.

John C Trueswell, Michael K Tanenhaus, and Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, 33(3):285–318.

Daphne P Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2025. Extracting structure from an llm-how to improve on surprisal-based models of human language processing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4938–4944.

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 Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 211–221.