Garden Path Traversal in GPT-2

William Jurayj
Brown University
william@jurayj.com

William Rudman
Brown University
william_rudman@brown.edu

Carsten Eickhoff
Brown University
carsten@brown.edu

Abstract

In recent years, large-scale transformer decoders such as the GPT-x family of models have become increasingly popular. Studies examining the behavior of these models tend to focus only on the output of the language modeling head and avoid analysis of the internal states of the transformer decoder. In this study, we present a collection of methods to analyze the hidden states of GPT-2 and use the model's navigation of garden path sentences as a case study. To enable this, we compile the largest currently available dataset of garden path sentences. We show that Manhattan distances and cosine similarities provide more reliable insights compared to established surprisal methods that analyze next-token probabilities computed by a language modeling head. Using these methods, we find that negating tokens have minimal impacts on unambiguous forms of NP/Z and NP/S sentences, but not on such forms of MV/RR sentences. Further, we find that GPT-2 recognizes words that might cause a garden path effect but ultimately do not– surprisal analysis routinely misses this nuance.

1 Introduction

OpenAI's release of GPT-3 marked a major step in the field of massive language models, whose ability to generate news articles indistinguishable from those written by humans provides a salient example of the many social and political implications of these models (Brown et al., 2020; Wallace et al., 2019; Heidenreich and Williams, 2021). Within 2 years of BERT's release, over 150 studies have investigated BERT's structure, exploring how its internal representations enable powerful and flexible language comprehension (Coenen et al., 2019; Kovaleva et al., 2019; Tenney et al., 2019; Rogers et al., 2020). Studies exploring GPT tend to focus on properties of text generated from its language modeling head and do not analyze the internal representations of the model in depth (Heidenreich and Williams, 2021; Brown et al., 2020).

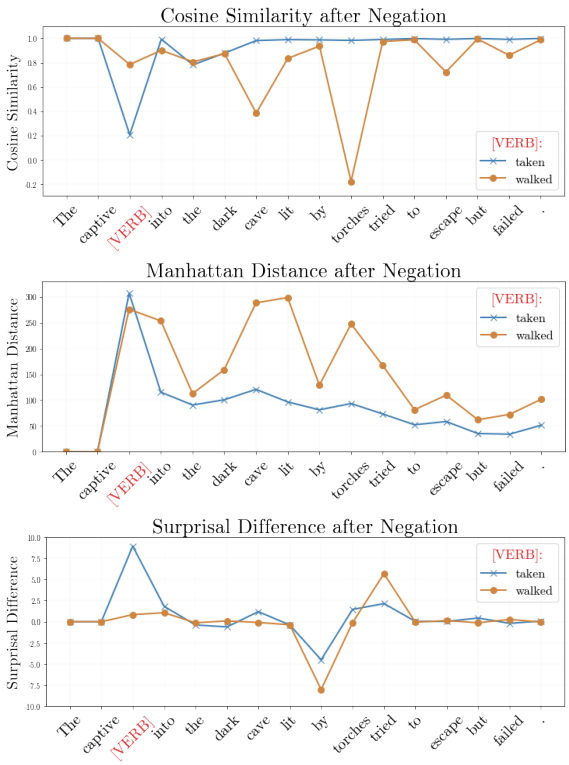

Figure 1: Hidden state relations (Top: cosine similarity, Middle: Manhattan distance, Bottom: surprisal difference) between negated and non-negated forms of garden path and unambiguous sentences. The ambiguous verb "walked" primes the effect later in the sentence, while the unambiguous "taken" avoids it.

The few studies that explore the hidden states of GPT-2 suggest an under-utilization of its massive latent space as representations are dominated by the presence of rogue dimensions (Ethayarajh, 2019; Cai et al., 2021; Rudman et al., 2021; Timkey and van Schijndel, 2021). As massive decoder models become more ubiquitous and powerful, it will become ever more important to understand the internal processes by which they generate content so they can be streamlined and improved upon.

In this paper, we use garden path traversal as a case study to demonstrate the value of directly analyzing properties of the embedding space in

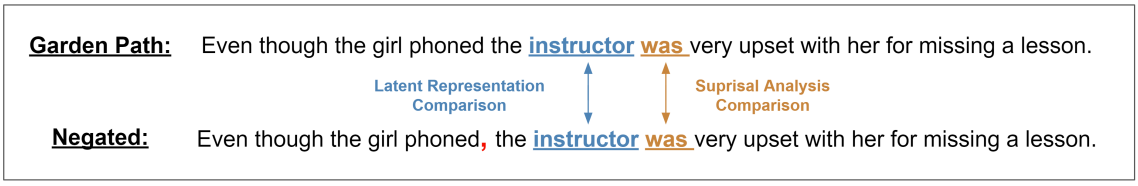

Figure 2: Method for comparing latent space metrics (cosine similarity, Manhattan distance) against surprisal difference

transformer decoder models. A garden path sentence is one where the parse that a reader expects at some point within the sentence is proven incorrect by the end of the sentence. By analyzing how GPT-2 sequentially embeds tokens in space, we are able to identify how GPT-2 internally handles different garden path effects. The contributions of this study are as follows:

- to introduce the largest and most diverse dataset of garden path sentences currently available, along with construction functions to negate or extend the effect within each sentence
- to provide methods of analyzing syntactic properties such as garden path effects by examining geometric relationships between vectors in GPT-2's hidden states using Manhattan distance and cosine similarity
- to motivate further study of the hidden states of transformer decoders as a more thorough alternative to the surprisal-based methods that are typically used to analyze language models.

1.1 Related Work

Many studies into GPT or BERT involve fine-grained analyses of how the model handles specific syntactic phenomena, such as the garden path effect. Consider the sentence:

"Even though the girl phoned[,] the instructor was very upset with her for missing a lesson."

Without the comma, most readers will assume "the instructor" is the direct object of the verb "phoned", rather than the subject of the main clause's verb phrase, "was very upset" (van Schijndel and Linzen, 2019). Adding the comma immediately disqualifies the incorrect parse, nullifying the garden path effect. This method of preventing the effect is referred to as "negation" throughout this paper.

Analysis of garden path traversal is typically done by comparing the surprisal, or negative log likelihood, of the token that would trigger the garden path effect between garden path and negated sentences. Surprisal is calculated using a language modeling head on top of GPT-2, and does not directly analyze the internal representations of the model from which these likelihoods are computed. Previous studies into the navigation of these sentences find that sufficiently large models' relative surprisals at the disambiguating token between garden path and negated sentences show recognition of the garden path effect. However, these models systematically underestimate the magnitude of the effect observed in humans, suggesting that human recovery from an incorrect parse involves more than just the triggering token's lack of predictability (van Schijndel and Linzen, 2021, 2018). Further, using surprisal comparisons, Hu et al. (2020) show that GPT-2 recognizes garden path effects less successfully or consistently than smaller recurrent language models.

OpenAI has not yet released GPT-3's source code and parameters, so we instead analyze its predecessor GPT-2, which uses an almost identical architecture at a much smaller scale (1.5b parameters). Nonetheless, the methods we use to explore GPT-2's traversal of garden path effects can be easily generalized to study any decoder-based model.

2 Methods

2.1 Garden path sentence generation

The dataset used for these experiments builds on the combination of the NP/Z and NP/S sentences from van Schijndel and Linzen (2018) and the NP/Z and MV/RR sentences from Futrell et al. (2019), and consists of 43 NP/Z sentences, 20 NP/S sentence, and 20 MV/RR sentences. Instead of building out side-by-side datasets of each type of sentence, however, we store the components of these sentences in tabular files, and include scripts to construct these sentences in various forms similar to those used by Futrell et al. (2019). Each sentence has a *garden path* and an *unambiguous* form, depending on whether the first verb allows for an ambiguous parse. Each of these forms can be *negated* with the addition of one or two tokens, which nullifies the garden path effect in an ambigu-

Sentence Type	Sentence Form	Sentence
NP/Z	Garden Path	When the dog scratched the vet took off the muzzle.
	Negated	When the dog scratched, the vet took off the muzzle.
	Blocked	When the dog scratched his owner the vet took off the muzzle.
	Unambiguous	When the dog struggled the vet took off the muzzle.
NP/S	Garden Path	The coach discovered the player tried to show off all the time.
	Negated	The coach discovered that the player tried to show off all the time.
	Unambiguous	The coach thought the player tried to show off all the time.
MV/RR	Garden Path	The horses raced past the barn fell into a ditch.
	Negated	The horses that were raced past the barn fell into a ditch.
	Unambiguous	The horses ridden past the barn fell into a ditch.

Table 1: Forms of NP/Z, NP/S, and MV/RR sentences included in our dataset, with the verb that triggers or would trigger the garden path effect underlined in red. Note that all of the perturbations can be combined to avoid the garden path effect, except for the blocked and unambiguous forms of the NP/Z sentence.

ous sentence, but makes no semantic difference in an unambiguous sentence. We provide this as a template to be extended indefinitely to meet the needs of future research. Examples of each sentence type's possible forms can be found along with a detailed description of these effects in Table 1 in the appendix.

2.1.1 NP/Z sentences

The garden path effect in these sentences is caused by ambiguity about whether the verb of the leading subordinate clause has a direct object. These sentences have an additional **blocked** form, which nullifies its garden path effect by adding an explicit direct object to the leading verb. This is considered one of the stronger types of garden path effects, with an average increase in human reading time of 152 ms (Sturt et al., 1999).

2.1.2 NP/S sentences

The garden path effect in these sentences is caused by ambiguity about whether the noun following the main clause's verb is that verb's direct object. This is considered one of the weaker types of garden path effects, with an average increase in human reading times of 50 ms (Sturt et al., 1999).

2.1.3 MV/RR Sentences

The garden path effect in these sentences is caused by ambiguity about whether the past-tense verb of the leading subordinate clause is a past participle or the main verb of the sentence. This effect is considered stronger than that of an NP/S sentence, but reading time data to compare it with the other sentence types is not available.

2.2 Experimental design

The general structure of the tests we run is inspired by Futrell et al. (2019) and Hu et al. (2020). The key difference is that, where previous studies compare the model's surprisal at the disambiguating word, we examine the model's hidden state prior to this word. Figure 4 visualizes our approach.

We compare each sentence to its negated form, computing the vector differences and cosine similarities between each token and its counterpart in the negated form (omitting the token[s] that were added to negate the garden path effect in that sentence type from the pairing process) after re-centering embeddings around the origin. We use Manhattan distance over Euclidean distance to compute scalars from the vector differences between sentences as is generally preferred in high dimensional spaces, where Euclidean distances are sensitive to the dimensions with the largest values (Aggarwal et al., 2001). Cosine similarities are computed after re-centering all vectors so that the distribution has a mean of zero, which prevents the metric from defaulting to near-maximum values and allows it to measure the true directional changes between vectors (Rudman et al., 2021). These side-by-side metrics are generated for all sentences' garden path and unambiguous forms, as well as for the blocked form of the NP/Z sentences.

We expect to see larger distances and lower similarities upon negation in garden path sentences than in unambiguous or blocked sentences. In the garden path sentences the negating tokens help to resolve some ambiguity, whereas in an already unambiguous sentence they will contribute minimally to the sentence's meaning prior to the triggering token.

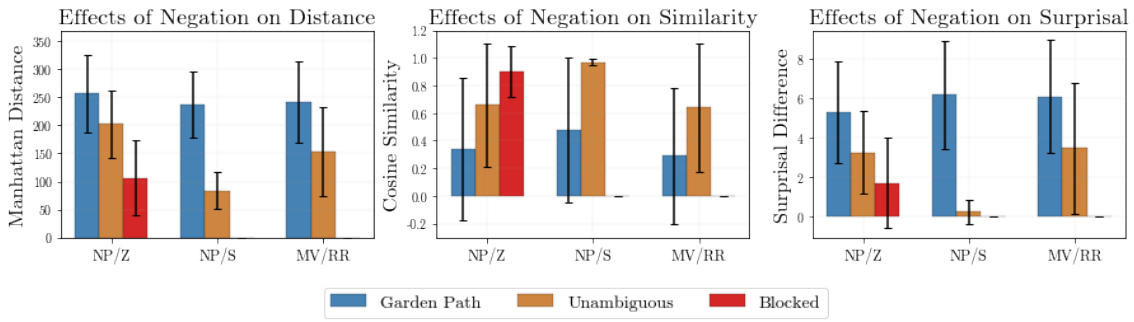

Figure 3: Left: Manhattan distances between sentence types and their negated forms. Center: Cosine similarities between sentence types and their negated forms. Right: Surprisal differences between sentence types and their negated forms

3 Results & Discussion

Our analysis reveals several properties of GPT-2's experience of the garden path effect. Across all sentence types, Manhattan distances and cosine similarities show that the model reacts more heavily reacts to negation of garden path sentences than it does to these sentences' unambiguous counterparts, as is reflected by surprisal analyses done here and in previous studies (Sarti, 2020).

Although our surprisal baselines mirror the trends seen in Manhattan distance, using Manhattan distances provides more consistent results compared to surprisal analysis. Our results demonstrating exceedingly high variance in the surprisal analysis is in line with previous work (Hu et al., 2020). Hu et al. (2020) use surprisal comparisons to score a language model's ability to generalize syntactic features, and find that GPT-2 performs especially poorly and inconsistently on garden path effects. On the other hand, the high-level trends we expected to see are present across all metrics, with negation causing a less pronounced difference in unambiguous and blocked sentences than it does in garden path sentences. Whereas Figure 3 shows Manhattan distances to have relatively low variance compared to the other metrics we examine, cosine similarity and surprisal suffer from very high variances within each sentence form. We believe that this is due to Manhattan distance's resistance to GPT-2's rogue dimensions, which dominate Euclidean distance and cosine similarities (Timkey and van Schijndel, 2021; Aggarwal et al., 2001).

Figure 1 illustrates how these high level trends appear within a given sentence. Here, the verb "tried" triggers the garden path effect, which is directly preceded by a spike in Manhattan distance and a dip in cosine similarity between the two sentence forms as the models anticipate different continuations: in the garden path sentence the model is likely to predict some sort of punctuation or conjunction to end the clause, while in the negated form the model expects a verb to complete the clause that the ambiguous verb is subordinate to.

An interesting property of the specific example, "The captive walked into the dark cave lit by torches tried to escape but failed." is that the verb "lit" also triggers a momentary garden path effect; the sentence could, for instance, simply continue, "The captive walked into the dark cave lit the torches." This possibility is worth considering because it helps explain why there is a spike in Manhattan distance and a dip in cosine similarity at the preceding word, "cave". We believe the relative shallowness of the dip in cosine similarity before "lit" is due to the semantic ambiguity of the word, since even in the garden path case where punctuation can be expected, a verb such as "lit" can easily preempt an adjectival clause as it does here ("lit by torches"). Notably, in the garden path form neither hidden state metric returns to its baseline value until after the verb "lit", because the model expects this verb leads a subordinate clause whereas in the negated form it considers both possibilities.

On the other hand, the verb "tried" is not ambiguous in this way. The clause it might preempt, such as "tried for murder", would be improperly placed and awkward. Thus, the model's hidden states prior to the verb "tried" in the garden path and negated sentences are nearly orthogonal to each other, whereas they bear more similarity right before the verb "lit". Although surprisal spikes at the verb "tried" as well, the surprisal trajectory does not reflect the possible effect at "lit", illustrating the inadequacy of using this metric alone. We believe this type of analysis helps us understand how Manhattan distance and cosine similarity encode relationships between the model's syntactic states.

However, more work is needed to explore how exactly each of these metrics measures this abstract concept, and how sensitive they are to other syntactic and semantic effects. See Figures 5 and 6 for more examples of this phenomenon.

Our analysis revealed a few unexpected results. Most prominent among these is the extent to which the addition of the negating token ("that") to unambiguous NP/S sentences leaves the hidden representation of the sentence unchanged. Across all metrics, the negated and garden path forms of NP/S sentences are closest together, showing that except in cases where it resolves a clear ambiguity, the negating token in these sentences contributes very little to the model's internal representation. The blocked form of NP/Z sentences shows a similar indifference to the negating token (in this case, the addition of a comma), which curiously does not extend to the unambiguous form.

4 Conclusion

This paper presents a suite of methods to analyze the internal representations of transformer decoder language models such as GPT-2, taking advantage of a richer reflection of the model's internal process than can be ascertained from the output of the language modeling head alone. We use Manhattan distance and cosine similarity on the hidden states of GPT-2 to show that GPT-2 is affected by garden path effects in ways that are predictable based on human readers' difficulty with these sentences. Although conventional surprisal analysis mirrors these effects in many cases, it misses certain nuances and exhibits higher variance than Manhattan distances. We hope that these early insights will help inspire deeper exploration of the hidden states of decoder-only language models. Possible directions for future work could more closely examine how information is transformed across different decoder layers within GPT-2, and might explore causes for differences between Manhattan distance and cosine similarity trajectories. The methods introduced in this study can be used to explore decoder models' handling of arbitrary syntactic phenomena beyond garden path effects, such as verb subordination or negative polarity item licensing.

5 Limitations

One weakness of this type of analysis is the necessity of having side-by-side examples, with a single perturbation between them, to compare between. The beam search approach used by Aina and Linzen (2021) avoids this requirement, but relies on the language modeling head, so more work is needed to integrate these benefits Another difficulty is the size of the dataset; although larger than all previous datasets of garden path sentences, it only includes 83 distinct sentences, and while many more variations can be generated with the scripts we include, there is substantial overlap between these that makes training a model on these challenging. Further, since weights for GPT-3 are not available, our analysis is constrained to GPT-2.

Acknowledgments

This research is supported in part by the NSF (IIS-1956221)T32 GM128596). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, NIH, or the U.S. Government.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Laura Aina and Tal Linzen. 2021. The language model understood the prompt was ambiguous: Probing syntactic uncertainty through generation. *CoRR*, abs/2109.07848.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winxter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. *CoRR*, abs/1909.00512.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state.

Hunter Heidenreich and Jake Williams. 2021. The earth is flat and the sun is not a star: The susceptibility of gpt-2 to universal adversarial triggers. pages 566–573.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. *CoRR*, abs/1908.08593.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2021. Isoscore: Measuring the uniformity of vector space utilization. *CoRR*, abs/2108.07344.

Gabriele Sarti. 2020. *Interpreting Neural Language Models for Linguistic Complexity Assessment*. Ph.D. thesis, University of Trieste.

Patrick Sturt, Martin Pickering, and Matther Crocker. 1999. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *CoRR*, abs/1905.05950.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *CoRR*, abs/2109.04404.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. *Cognitive Science*.

Marten van Schijndel and Tal Linzen. 2019. Neural network surprisal predicts the existence but not the magnitude of human syntactic disambiguation difficulty.

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.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp.

A Appendix

A.0.1 NP/Z sentences

The first sentence evokes a garden path effect because the reader initially expects that "the vet" is the direct object of "scratched"; The negated form avoids the effect by using a comma to indicate the separation between the two clauses. The blocked form avoids the effect by adding the direct object "his owner" to block the ambiguity that triggers the effect, while the unambiguous form avoids the effect by replacing the transitive verb "scratched" with the intransitive verb "struggled" to avoid ambiguity around the verb's direct object. Our dataset includes 43 distinct NP/Z sentences, and includes scripts allowing a user to easily transform these into unambiguous or blocked sentences. Moreover, each sentence has the option to include a negation, and an extension so as to increase the duration of the ambiguity.

A.0.2 NP/S sentences

The first sentence evokes a garden path effect because the reader expects that "the player" is the direct object of the verb 'discovered' until the word "tried" reveals that it is her propensity to show off that the coach is discovering. The negated form avoids the effect by adding "that" before "the player" to eliminate the possibility that 'the player' is the verb's direct object. The unambiguous form avoids the effect altogether by using the verb "thought", which could not allow a person to be its direct object. Our dataset includes 20 distinct NP/S sentences, each of which can be negated, unambiguous, extended, or any combination thereof.

A.0.3 MV/RR sentences

The first sentence evokes the garden path effect because the reader assumes "raced" is the main verb of the sentence, while the negated form negates this ambiguity by clarifying that "raced past the barn" is a descriptor for the horses rather than the main clause itself. Note that in some examples, the negating tokens are "who were" instead of "that were" The unambiguous form avoids ambiguity altogether by replacing the ambiguous "raced" with the unambiguously passive "ridden". Our dataset includes 20 distinct MV/RR sentences, each of

which can be negated, rendered unambiguous, extended, or any combination thereof.

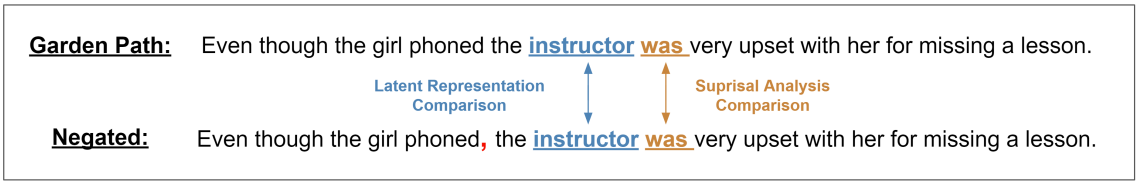

Figure 4: Method for comparing latent space metrics (cosine similarity, Manhattan distance) against surprisal difference

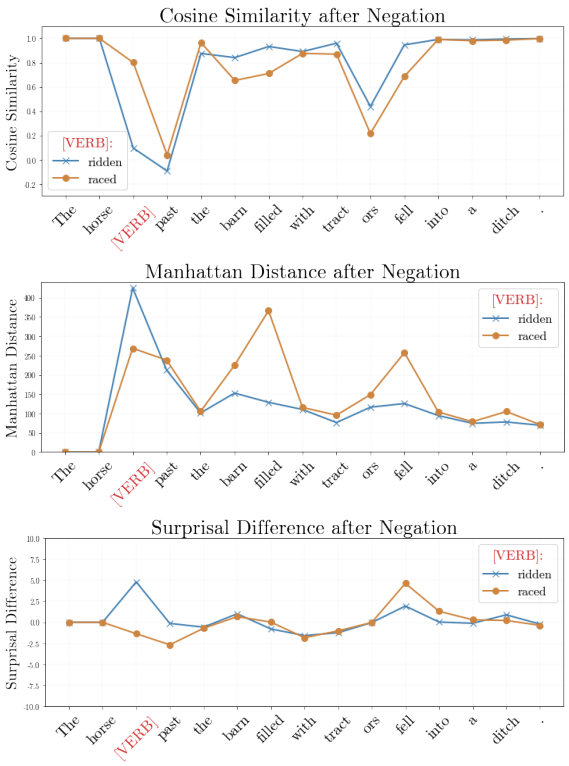

Figure 5: Hidden state relations (Top: cosine similarity, Middle: Manhattan distance, Bottom: surprisal difference) between negated and non-negated forms of garden path and unambiguous sentences. The ambiguous verb "raced" primes the effect later in the sentence, while the unambiguous "ridden" avoids it. Like in Figure 1, all metrics catch the garden path effect at the verb "fell", but only cosine similarity and Manhattan distance anticipate the possible effect at "filled"

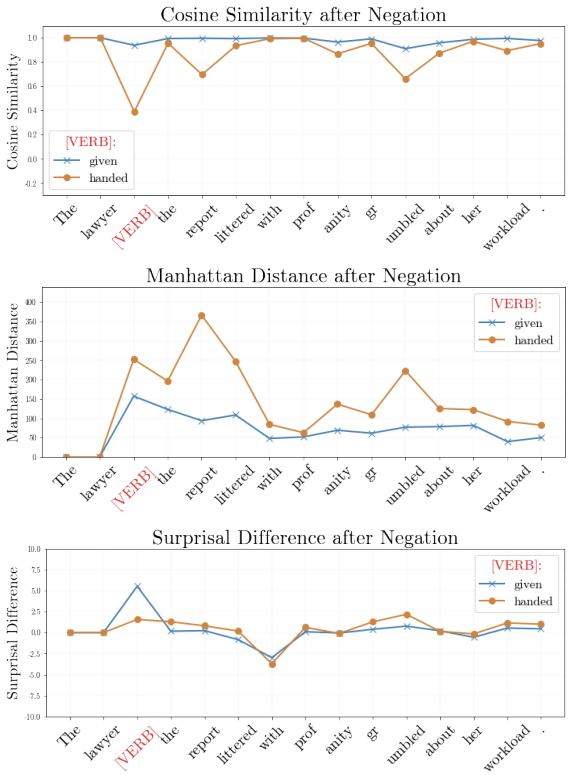

Figure 6: Hidden state relations (Top: cosine similarity, Middle: Manhattan distance, Bottom: surprisal difference) between negated and non-negated forms of garden path and unambiguous sentences. The ambiguous verb "handed" primes the effect later in the sentence, while the unambiguous "given" avoids it. Like in Figure 1, all metrics catch the garden path effect at the verb "grumbled", but only cosine similarity and Manhattan distance anticipate the possible effect at "littered"