Language model surprisal underpredicts garden path effects even with limited syntactic parallelism

William Timkey, Tal Linzen

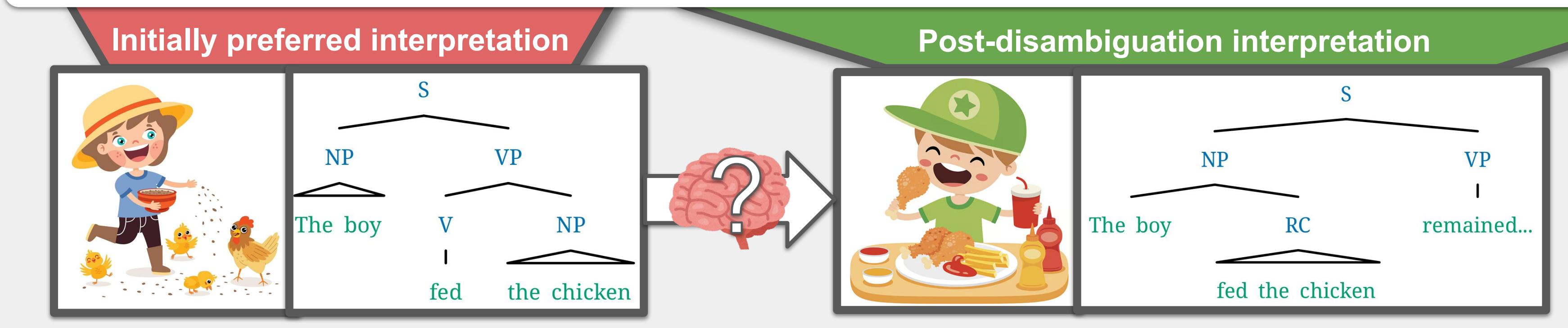
New York University

NYU

Introduction

- **Garden path (GP) sentences** are temporarily ambiguous between multiple syntactic structures
- Readers incur a processing cost when the sentence is disambiguated in an unexpected way:

The boy fed the chicken… remained relatively calm despite having asked for beef.


Initially preferred interpretation
S
NP
VP
The boy
V
NP
fed
the chicken
?
Post-disambiguation interpretation
S
NP
VP
The boy
RC
remained…
fed the chicken

Full-parallel parsing accounts (e.g. Levy 2008):

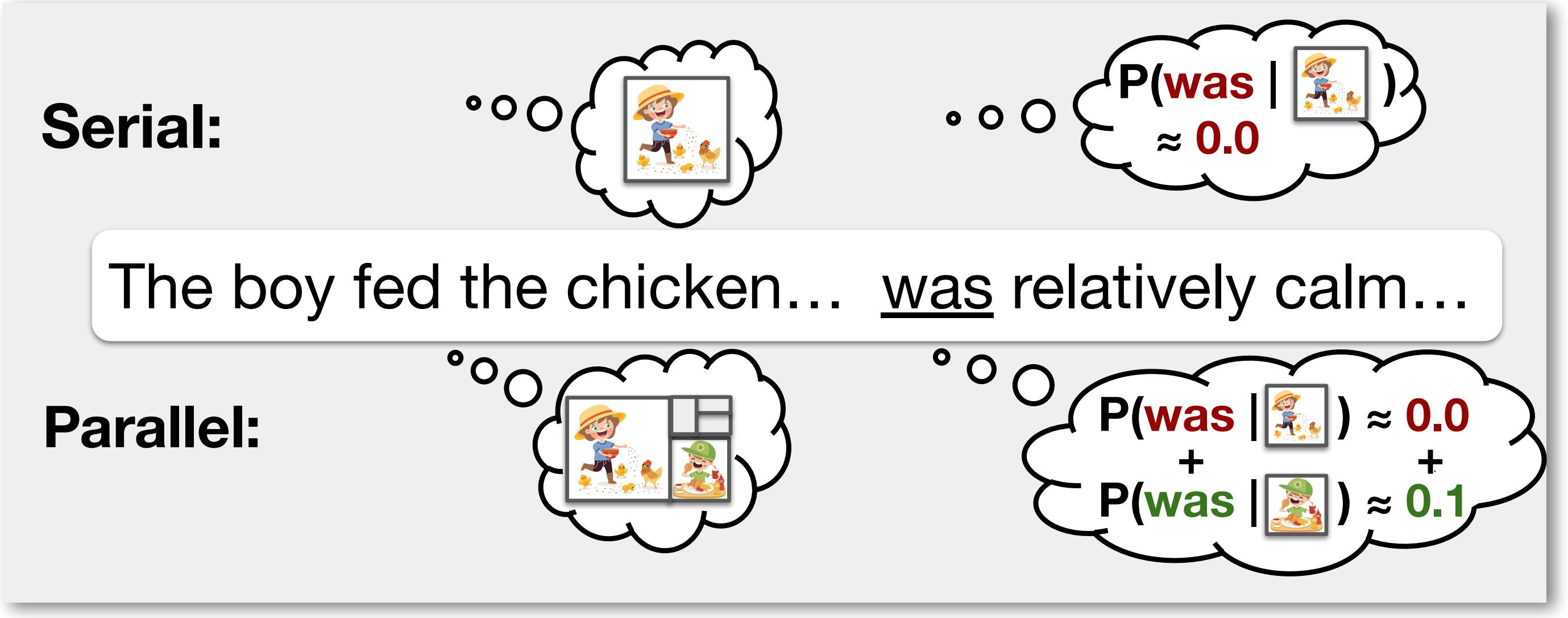
- A word's processing difficulty reflects the cost of updating beliefs over all possible interpretations,
  - Cost is equivalent to a word's surprisal (negative log probability).
- GP sentences are hard because the disambiguating region is unexpected.
- Surprisal-difficulty relationship is linear, and fixed (not specific to garden paths)

Serial parsing + reanalysis accounts:

- The parser can only entertain a single (or very few) interpretations simultaneously,
- GPEs reflect a costly reanalysis process when the initial interpretation is wrong (Frazier & Rayner 1982).

A problem for the full-parallel account:

- Surprisals from neural language models drastically underestimate the magnitude of GP effects in humans (Huang et al. 2024).
- Humans might commit more strongly to a single preferred interpretation, while LMs implicitly entertain many interpretations in parallel.
- **If we limit the syntactic parallelism of an LM, do we see more human like GP effects?**



Serial:

P(was |) ≈ 0.0

The boy fed the chicken… was relatively calm…

Parallel:

P(was |) ≈ 0.0
+
P(was |) ≈ 0.1

Controlling Syntactic Parallelism in LMs

Recurrent Neural Network Grammars:

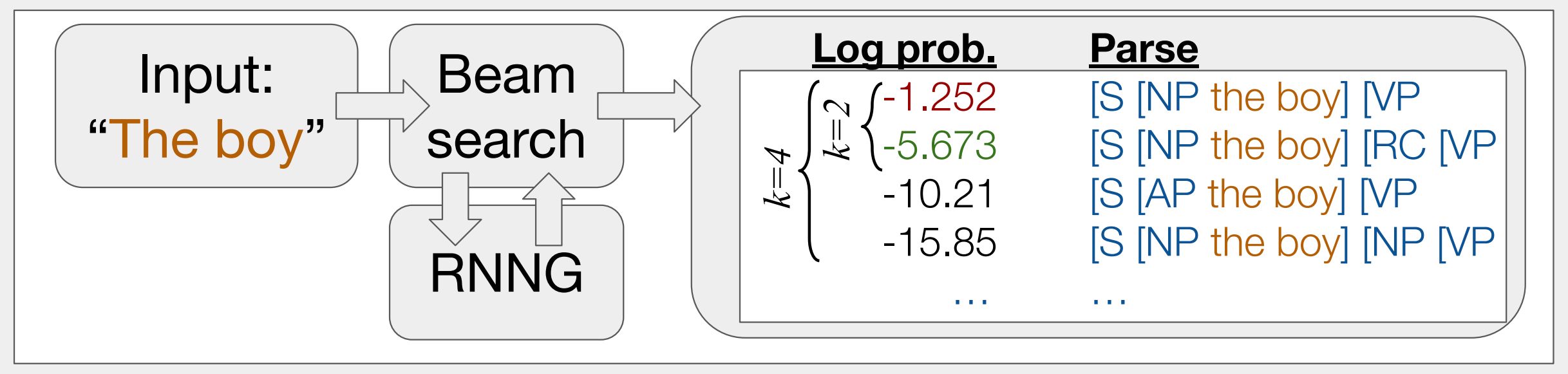
- Unlike standard language models, RNNGs are trained to predict both *the words of a sentence*, and its *structure*.

Standard LMs: P(was | the boy fed the chicken)
(e.g. GPT-2)

RNNGs: P(was | [S [NP the boy [RC [VP fed [NP the chicken]]][VP)

Word Synchronous Beam Search:

- At each word of a sentence, WSBS finds the RNNG's *k* most likely parses.



Input: "The boy"

Beam search

RNNG

		Log prob.	**Parse**
$k=4$	$k=2$	-1.252	[S [NP the boy] [VP
		-5.673	[S [NP the boy] [RC [VP
		-10.21	[S [AP the boy] [VP
		-15.85	[S [NP the boy] [NP [VP
		…	…

- Marginalizing over all *k* structures gives us word level probabilities:

P(fed | the boy) =

$k=5$ {
P(fed | [S [NP the boy] [VP) * P([S [NP the boy] [VP) +
P(fed | [S [NP the boy] [RC [VP) * P([S [NP the boy] [RC [VP) +
P(fed | [S [AP the boy]) * P([S [AP the boy] [VP) +
P(fed | [S [NP the boy] [NP [VP) * P([S [NP the boy] [NP [VP) +
P(fed | [S [AP [NP the boy] [PP) * P([S [AP [NP the boy] [PP) +

surprisal(fed | the boy) = -*log*(P(fed | the boy))

- Larger *k* = more syntactic parallelism in surprisal estimates

Methods:

1. Train RNNGs on a machine parsed version of the 50m token BLLIP news dataset.
2. Get surprisals for experimental stimuli using RNNGs+beam search with various beam widths.
3. Estimate surprisal-to-RT conversion factors on filler sentences.
4. Use conversion factors to predict GPEs in reading times (van Schijndel & Linzen 2021).

Forcing models to garden-path:

- As an upper bound, we can also "force" models to consider only the garden path interpretation of the sentence.
- In the "forced GP" condition, we only include parses that are consistent with the incorrect interpretation (MV/NP complement) when marginalizing.

Materials

1) Main verb / Reduced relative (MVRR)

The little boy (who was) fed the chicken **remained** relatively calm despite having asked for beef.

2) Direct object / Sentential complement (NP/S)

The little boy found (that) the chicken **remained** relatively calm despite the absence of its mother.

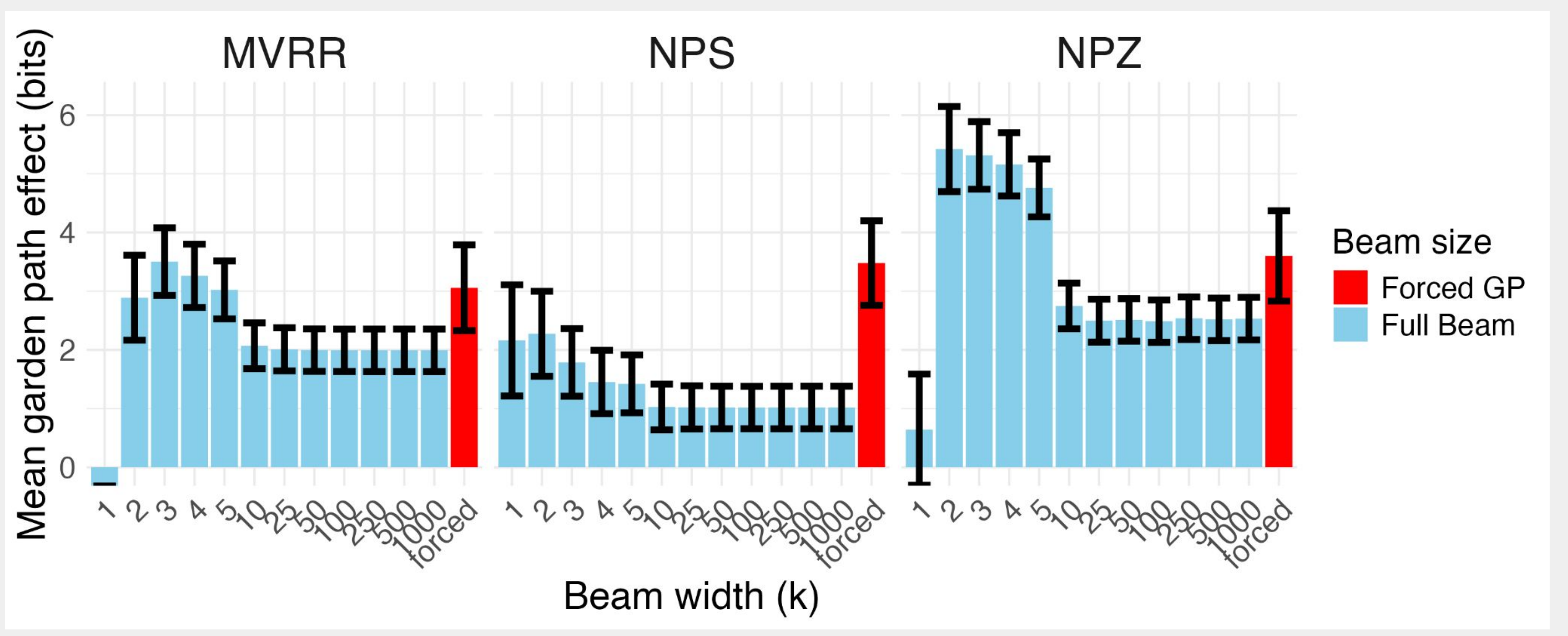
3) Noun phrase / Zero complement (NP/Z)

When the little boy attacked(,) the chicken **remained** relatively calm despite the sudden assault.

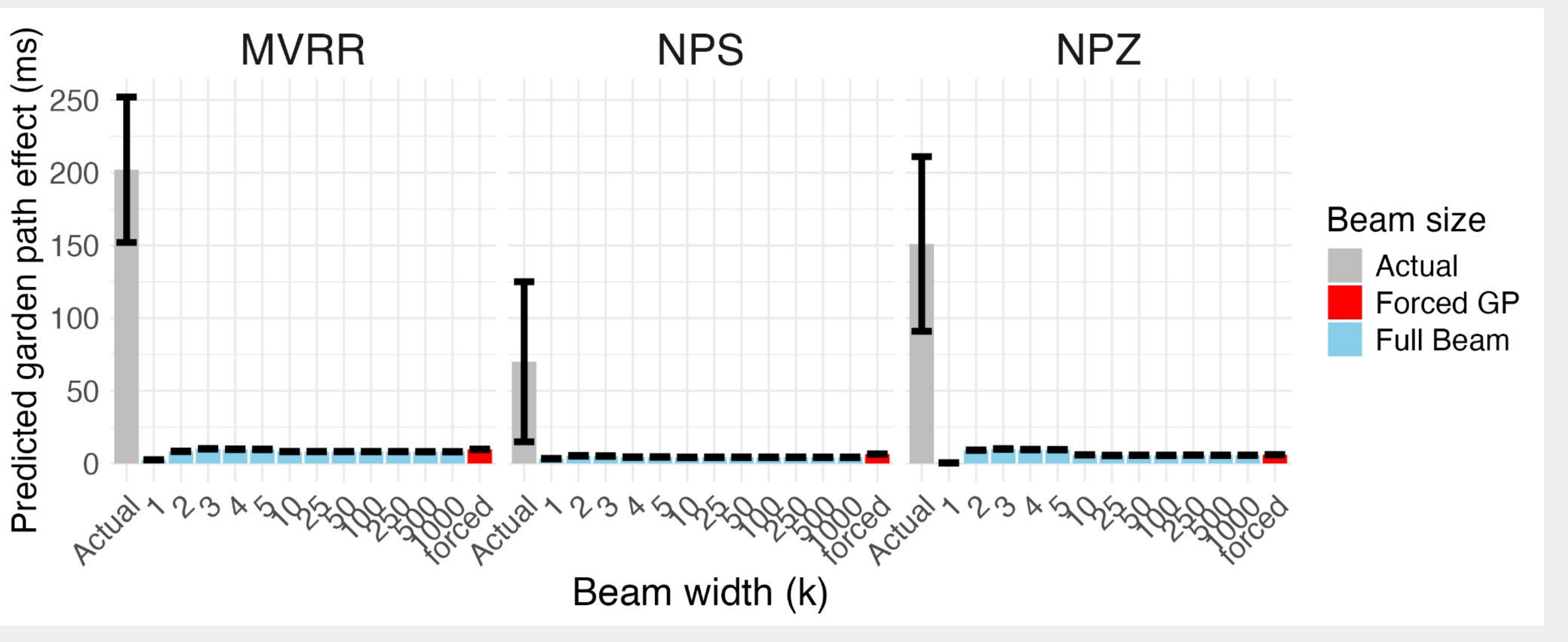
Unambiguous condition contains the green words in parentheses

Garden path effects:
RT("remained" | ambiguous) -
RT("remained" | unambiguous)

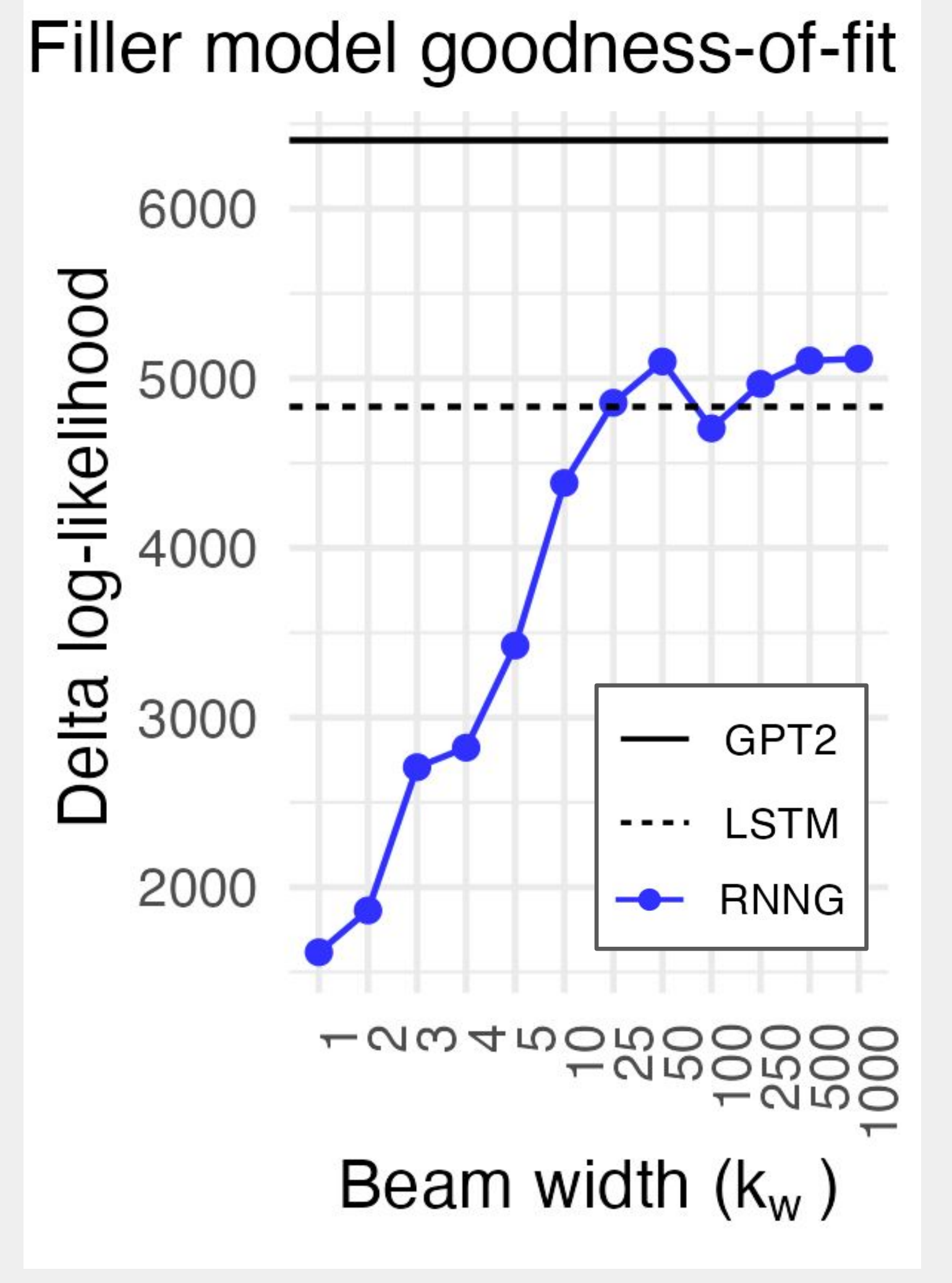
Less parallelism = larger garden path effects



…but effects are still drastically underpredicted



More parallelism = better fit to *filler* sentences



Conclusions

- Failure of LM surprisal to capture the magnitude of garden path effects in humans is not driven solely by differences in syntactic parallelism.
- Rather, LMs assign too much probability to the disambiguating word *even when forced to garden path.*
- Either LM probability estimates diverge from humans, or processing models also need a reanalysis component.