Language model surprisal does not underpredict garden path effects in early eye-tracking measures

Byung-Doh Oh

October 31, 2025
CLiMB Lab Meeting
Material adapted from Brian Dillon

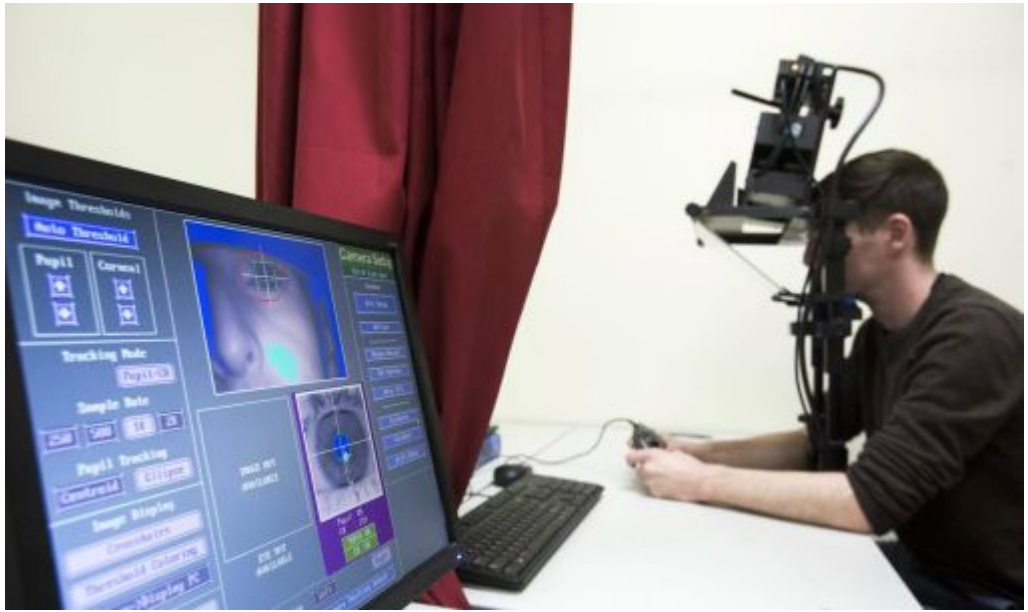
Computation and language in minds and brains (!)

The girl found the lamb …

You (and your brain) were probably able to:

- Build some mental representation without seeing the end of sentence
- Do so incrementally without much conscious effort

Eye movements in reading



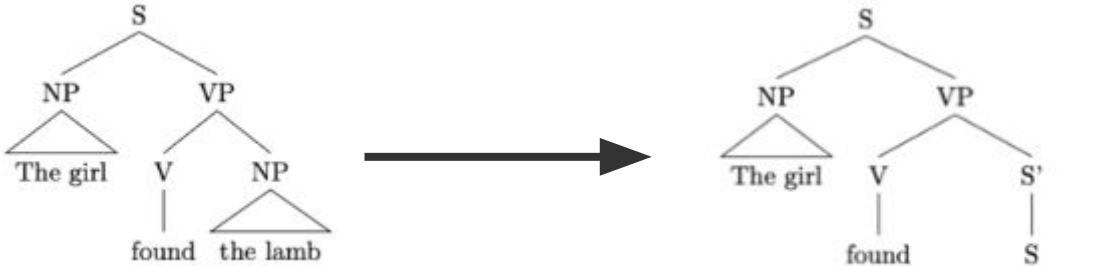
Schotter and Dillon (2025). A beginner's guide to eye tracking for psycholinguistic studies of reading.

The girl found the lamb

Schotter and Dillon (2025). A beginner’s guide to eye tracking for psycholinguistic studies of reading.

The girl found the lamb was extremely tasty.

Frazier and Rayner (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences.



???

The girl found the lamb <u>was</u> extremely tasty.

What does this tell us about processing?

- **An idea**: Language processing can be construed as a process of **inference under uncertainty**
- Comprehenders maintain a mental probability model over possible interpretations of the input
- As each word arrives, we use it to update our probability model

- **Predictable words carry less information and cause a smaller update**
- **Unpredictable words carry more information and cause a larger update**

Hale (2001), Levy (2008)

Surprisal: Effects of context

That is a very nice gin and tonic …

$$RT \propto -\log_2 P(w_t \mid w_{1..t-1})$$

Hale (2001), Levy (2008)

Surprisal: Effects of context

- Readers are sensitive to contextual probability contrasts over six orders of magnitude (Smith & Levy, 2013)
- RT effects are linear in surprisal across this whole range (Shain et al. 2024)
- Surprisal effects are not modulated by the presence of a highly predicted alternative (Frisson et al. 2017; Wong et al. 2024; cf. Cevoli et al. 2022)

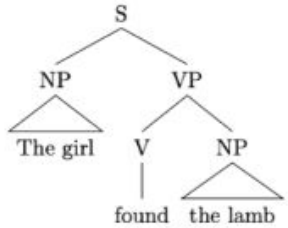
- These are non-obvious predictions of a theory that casts language processing as incremental belief update over sentence structure

Hale (2001), Levy (2008)

What does surprisal say about garden path effects?

The girl found the lamb <u>was</u> extremely tasty.

$-\log_2 P(\text{was} \mid \text{The girl found the lamb})$ is high as was is incompatible with



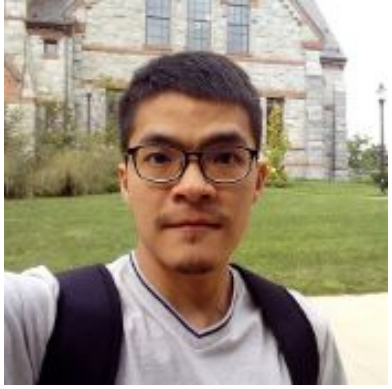
See Levy (2008) for how surprisal is derived from distributions over incremental parses;
This idea is embodied by incremental, generative parsers, e.g. van Schijndel et al. (2013), Hale et al. (2018)

The SAP Benchmark

- Diverse set of psycholinguistic contrasts, targeting various aspects of syntactic processing
- Collected large reading time datasets in multiple reading paradigms
  (SPR, Huang et al. 2024; eye-tracking, Timkey et al. in prep)

- **Can surprisal explain all processing difficulty associated with these diverse constructions?**

Will Timkey,
NYU



Kuan-Jung Huang,
MIT/UMD

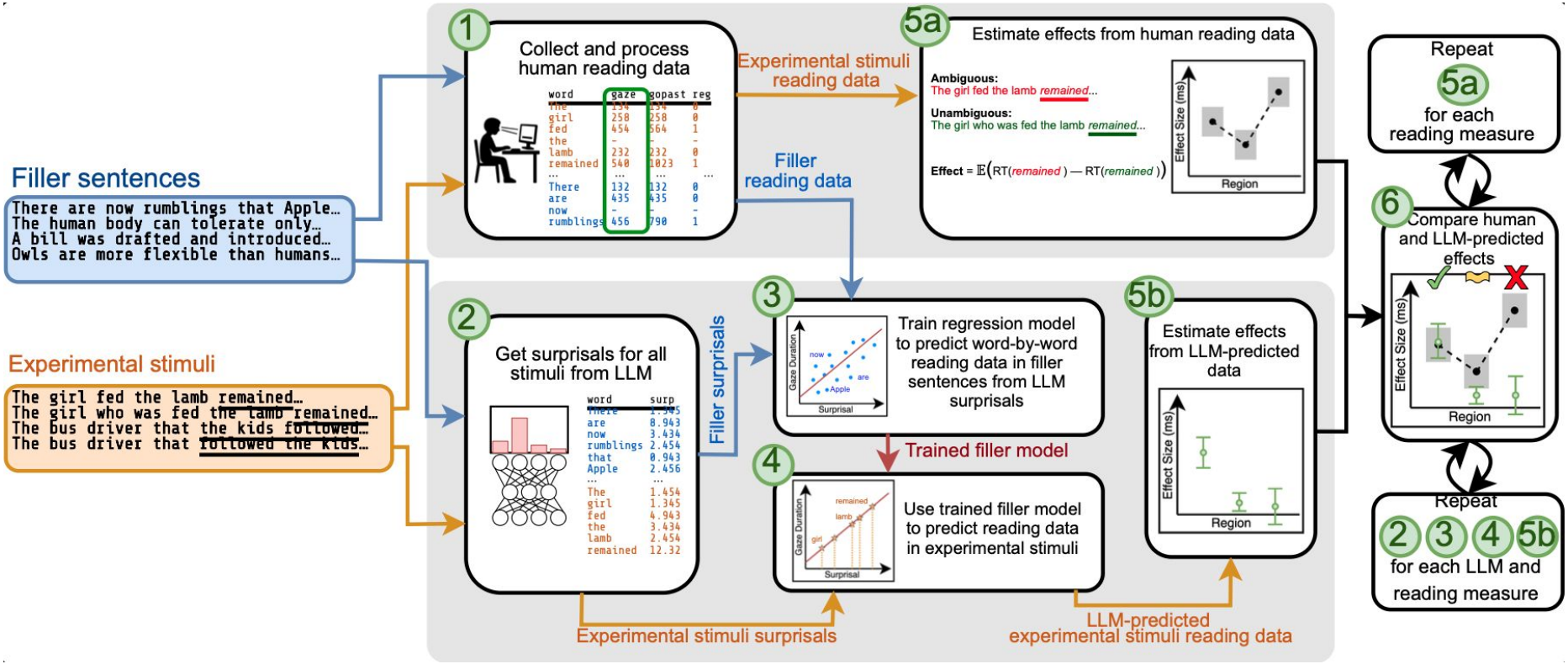
Suhas Arehalli,
Macalester

Grusha Prasad,
Colgate

Brian Dillon,
UMass

Tal Linzen,
NYU

Overview of methodology



1

Data collection

Main verb / reduced relative clause garden path

The little girl fed the lamb **remained** relatively calm …
The little girl who was fed the lamb **remained** relatively calm …

Direct object / sentential complement garden path

The little girl found the lamb **remained** relatively calm …
The little girl found that the lamb **remained** relatively calm …

Transitive / intransitive garden path

When little girl attacked the lamb **remained** relatively calm …
When little girl attacked, the lamb **remained** relatively calm …

Subject / object relative clauses

The bus driver that **the children** followed …
The bus driver that followed **the children** …

Relative clause modifies recent noun (low attachment)

Janet charmed the executives of the assistant who **decides** almost everything …
Janet charmed the executive of the assistant who **decides** almost everything …

Relative clause modifies distant noun (high attachment)

Janet charmed the executive of the assistants who **decides** almost everything …
Janet charmed the executive of the assistant who **decides** almost everything …

Subject-verb agreement mismatch

Whenever the nurse calls, the doctors **stops** working immediately …
Whenever the nurse calls, the doctor **stops** working immediately …

+ 40 Filler sentences from the Provo Corpus (Luke & Christianson, 2018)

Large eye-tracking dataset:
n=368!

Leveraging LMs to test theory

$-\log_2 P(\text{was} \mid \text{The girl fed the lamb})$

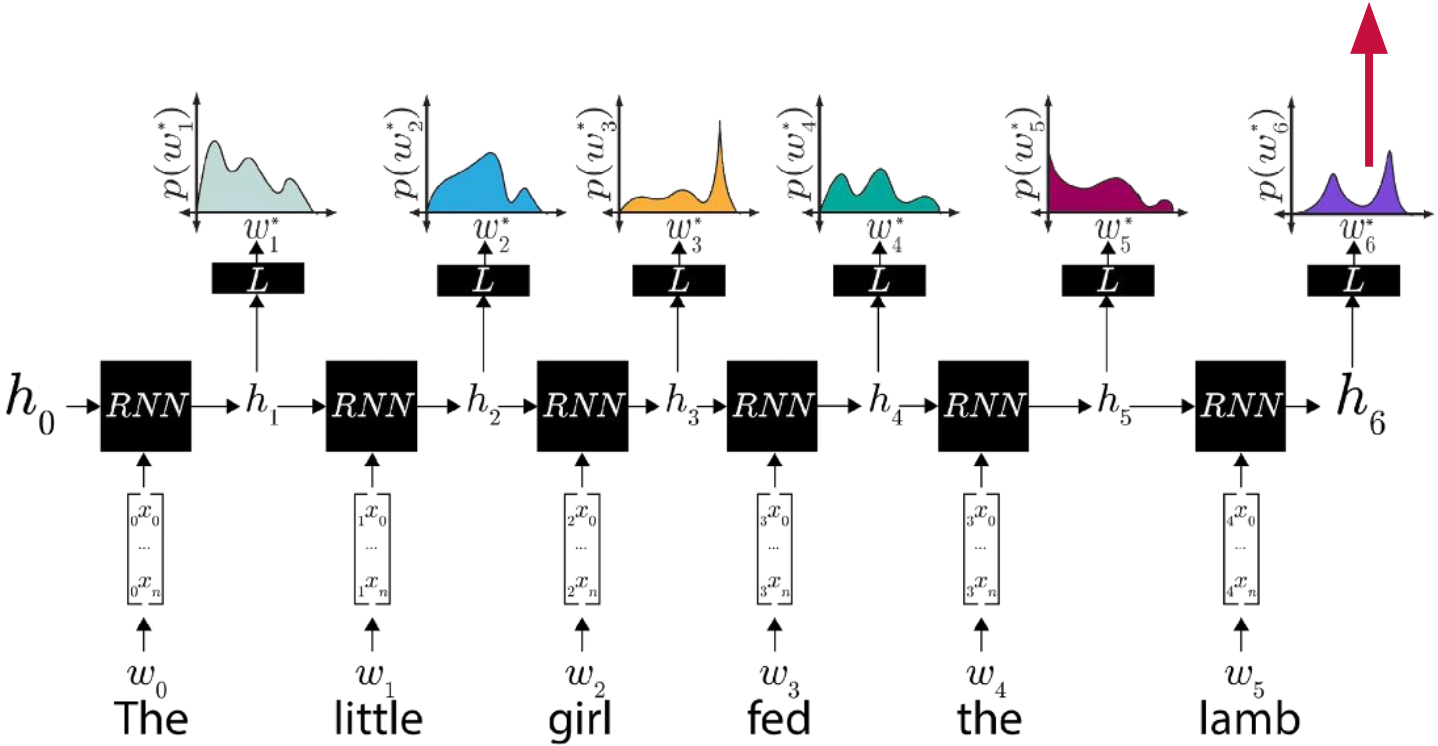


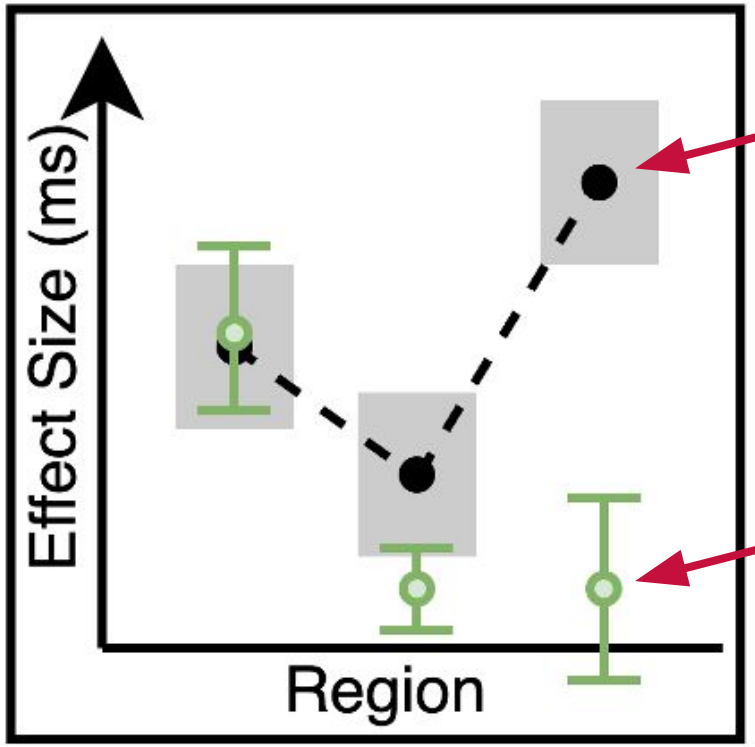
Figure from Suhas Arehalli

Leveraging LMs to test theory

- Pythia (Biderman et al. 2023): 8 sizes x 19 training steps
- GPT-2 (Radford et al. 2019): 4 sizes
- vSLSTM (van Schijndel et al. 2019): 5 sizes x 5 training data amounts x 5 corpora
- WikiLSTM (Gulordava et al. 2018)
- RNNG (Dyer et al. 2016): 5 seeds x {top-down, left-corner} x 12 beam sizes
- Mamba (Gu & Dao, 2023): 5 sizes

- Testing the predictions of surprisal from 407 LMs!
- Results from one model from each ‘family’ that best fits filler RTs

Comparing effects of interest



Actual garden path effect (humans): Critical sentence RT minus control sentence RT; estimated by effect model

Predicted garden path effect (LMs): Given the surprisal difference, how big of an effect do we predict?

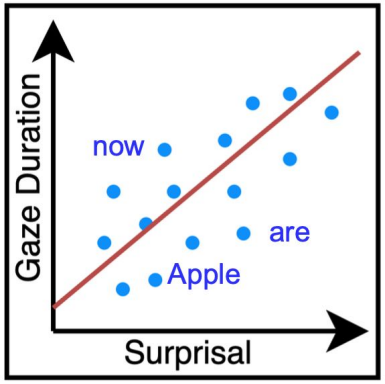
Estimating effects of interest

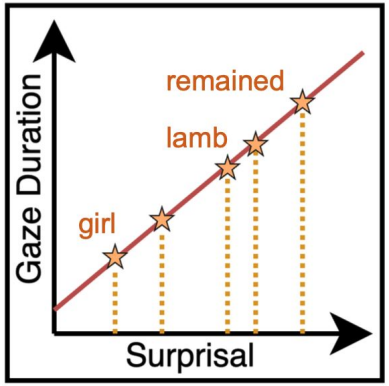
Filler sentences

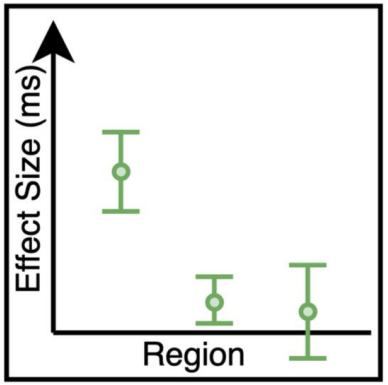
There are now rumblings that Apple…
The human body can tolerate only…
A bill was drafted and introduced…
Owls are more flexible than humans…

Experimental stimuli

The girl fed the lamb remained…
The girl who was fed the lamb remained…
The bus driver that the kids followed…
The bus driver that followed the kids…







Fit a filler model to RTs of filler sentences

Use filler model to predict RTs of stimuli sentences

Fit effect model to the predicted RTs

Assumption: Surprisal is linearly related to RTs

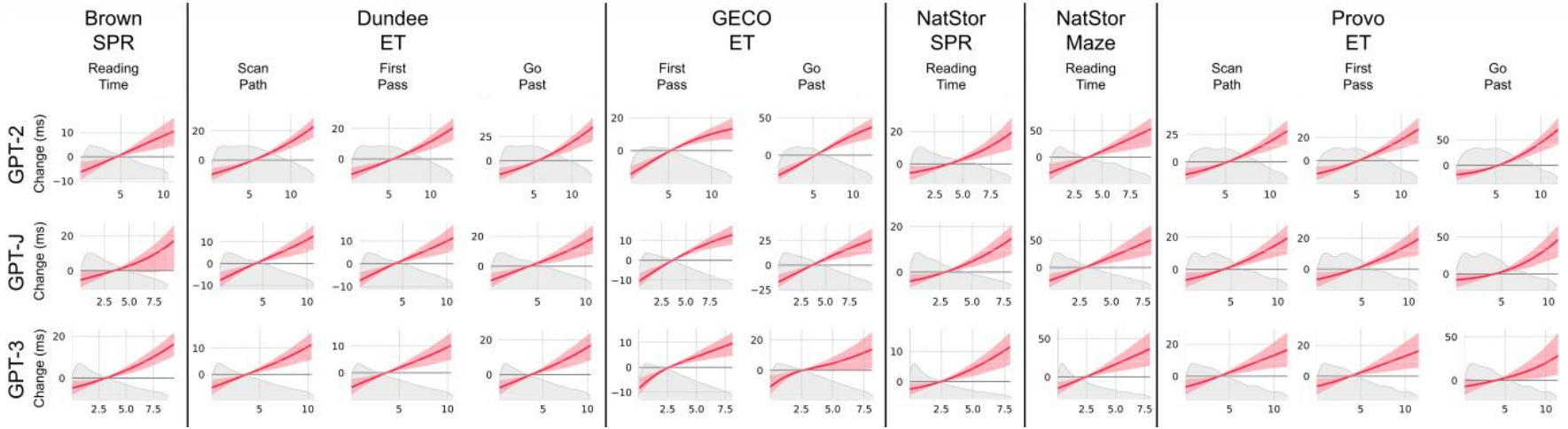
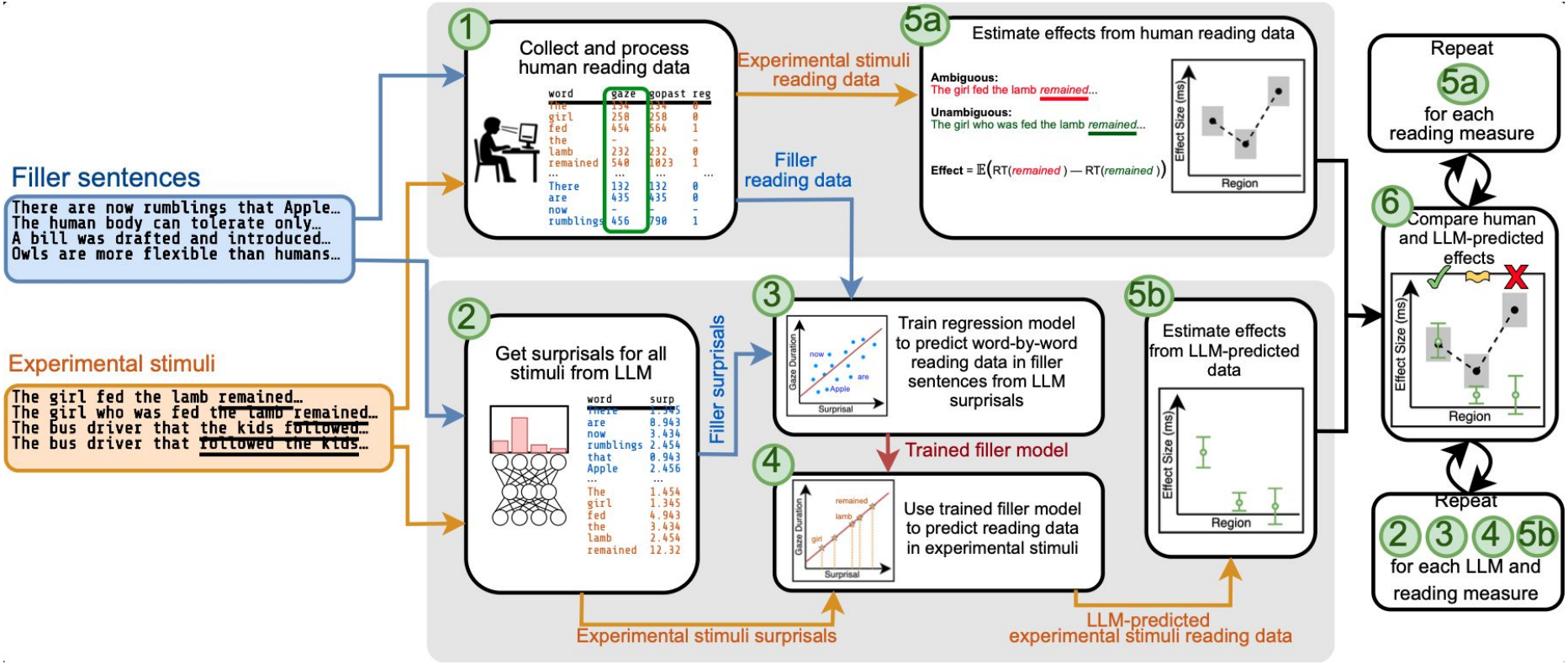
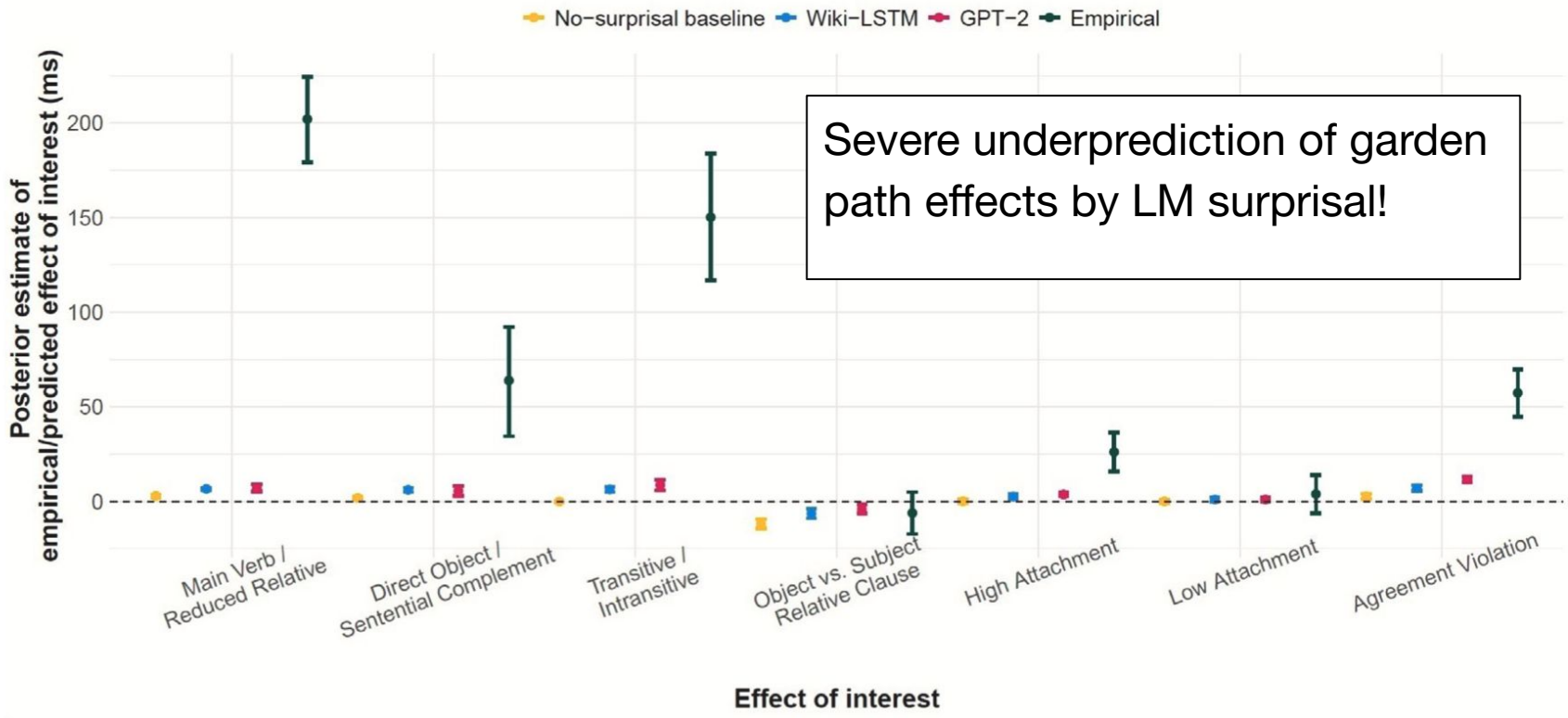


Figure from Shain et al. (2024)

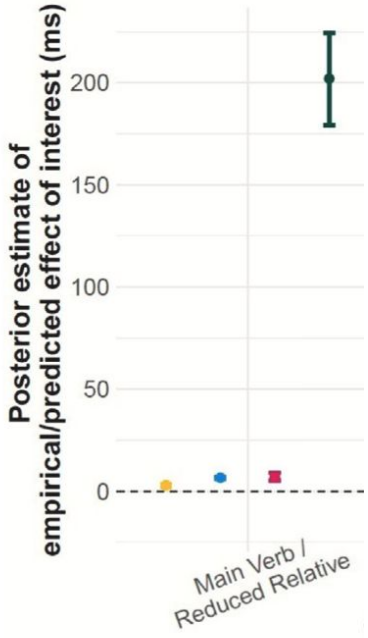

Filler sentences
There are now rumblings that Apple…
The human body can tolerate only…
A bill was drafted and introduced…
Owls are more flexible than humans…
Experimental stimuli
The girl fed the lamb remained…
The girl who was fed the lamb remained…
The bus driver that the kids followed…
The bus driver that followed the kids…
1
Collect and process human reading data
word gaze gopast reg
The 134 134 0
girl 258 258 0
fed 454 564 1
the - - -
lamb 232 232 0
remained 540 1023 1
... ...
There 132 132 0
are 435 435 0
now - - -
rumblings 456 790 1
Experimental stimuli reading data
Filler reading data
5a
Estimate effects from human reading data
Ambiguous:
The girl fed the lamb remained...
Unambiguous:
The girl who was fed the lamb remained...
Effect = 𝔼(RT(remained) — RT(remained))
Effect Size (ms)
Region
2
Get surprisals for all stimuli from LLM
word surp
There 1.345
are 8.943
now 3.434
rumblings 2.454
that 0.943
Apple 2.456
... ...
The 1.454
girl 1.345
fed 4.943
the 3.434
lamb 2.454
remained 12.32
Filler surprisals
3
Gaze Duration
Surprisal
now
are
Apple
Train regression model to predict word-by-word reading data in filler sentences from LLM surprisals
Trained filler model
4
remained
lamb
girl
Use trained filler model to predict reading data in experimental stimuli
5b
Estimate effects from LLM-predicted data
Experimental stimuli surprisals
LLM-predicted experimental stimuli reading data
Repeat
5a
for each reading measure
6
Compare human and LLM-predicted effects
Repeat
2 3 4 5b
for each LLM and reading measure

Previous results with SPR data (Huang et al. 2024)



Severe underprediction of garden path effects by LM surprisal!

Previous results with SPR data (Huang et al. 2024)

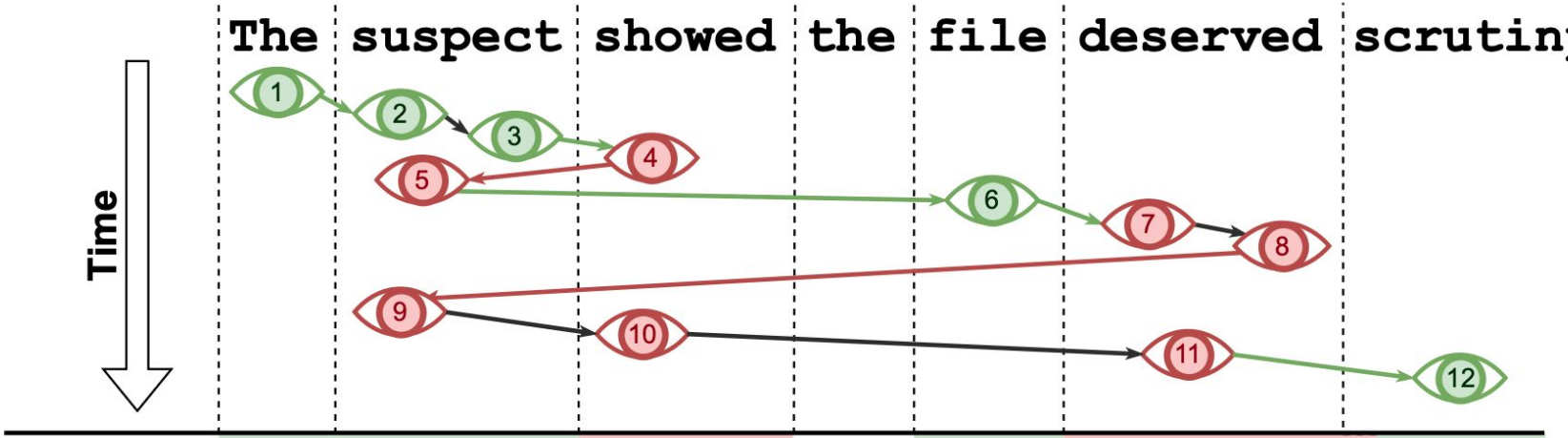


LMs are too ‘superhuman’ as surprisal estimators

Surprisal and structural processing make distinct contributions to reading difficulty

Eye movement measures

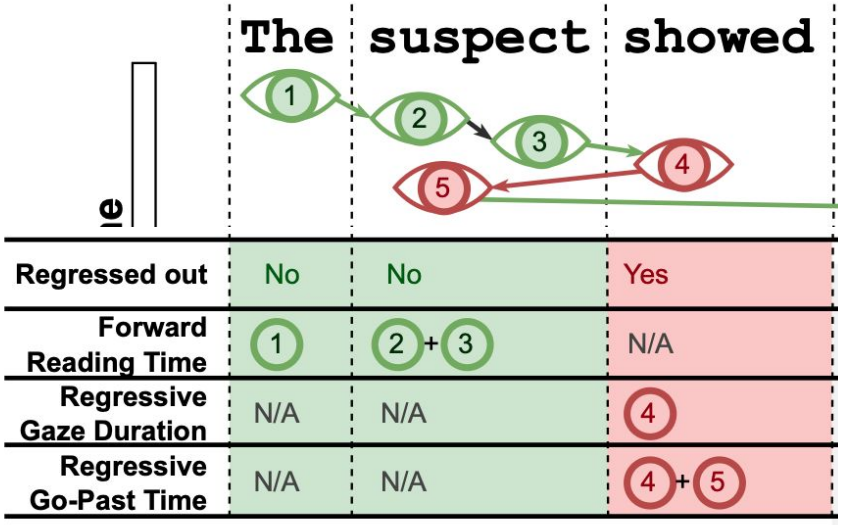
Regression-contingent measures; Does a reader exit to the left or right after?



	The	suspect	showed	the	file	deserved	scrutiny
Regressed out	No	No	Yes	N/A	No	Yes	No
Forward Reading Time	1	2 + 3	N/A	N/A	6	N/A	12
Regressive Gaze Duration	N/A	N/A	4	N/A	N/A	7 + 8	N/A
Regressive Go-Past Time	N/A	N/A	4 + 5	N/A	N/A	7 + 8 + 9 + 10 + 11	N/A

Eye movement measures

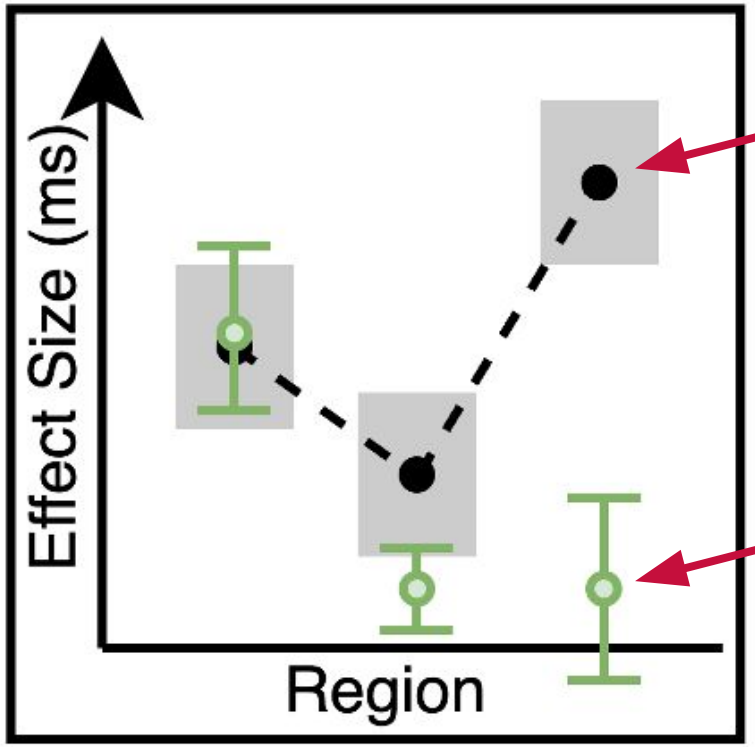
Regression-contingent measures; Does a reader exit to the left or right after?



Gaze Duration	1	2 + 3	4
Go-Past Time	1	2 + 3	4 + 5

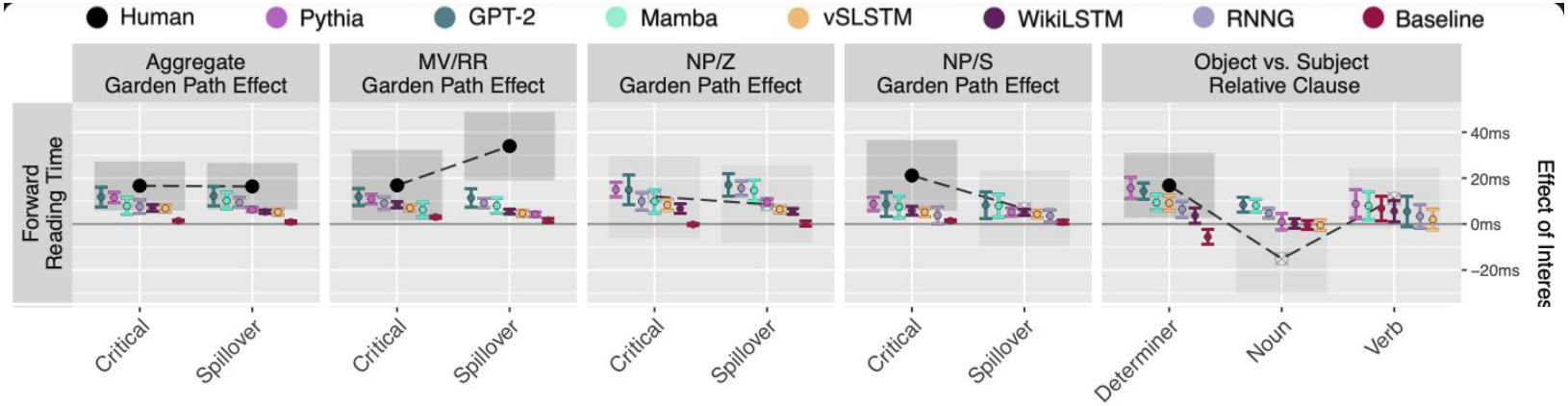
‘Traditional’ measures can be defined for every word, but confounds forward reading and backward reading

One last priming session before the actual results…



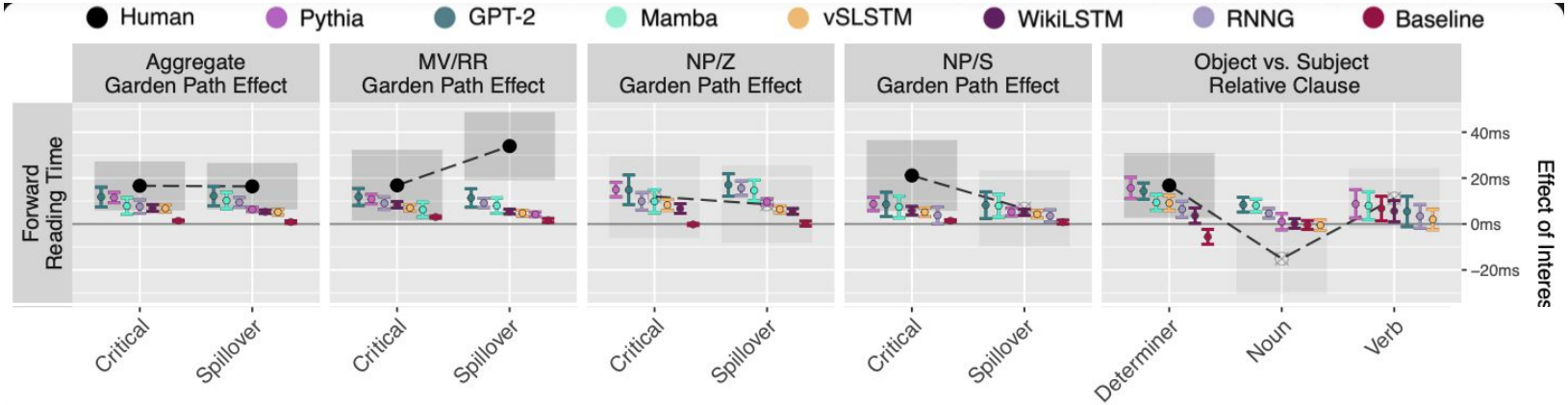
Actual garden path effect (humans): Critical sentence RT minus control sentence RT; estimated by effect model

Predicted garden path effect (LMs): Given the surprisal difference, how big of an effect do we predict?

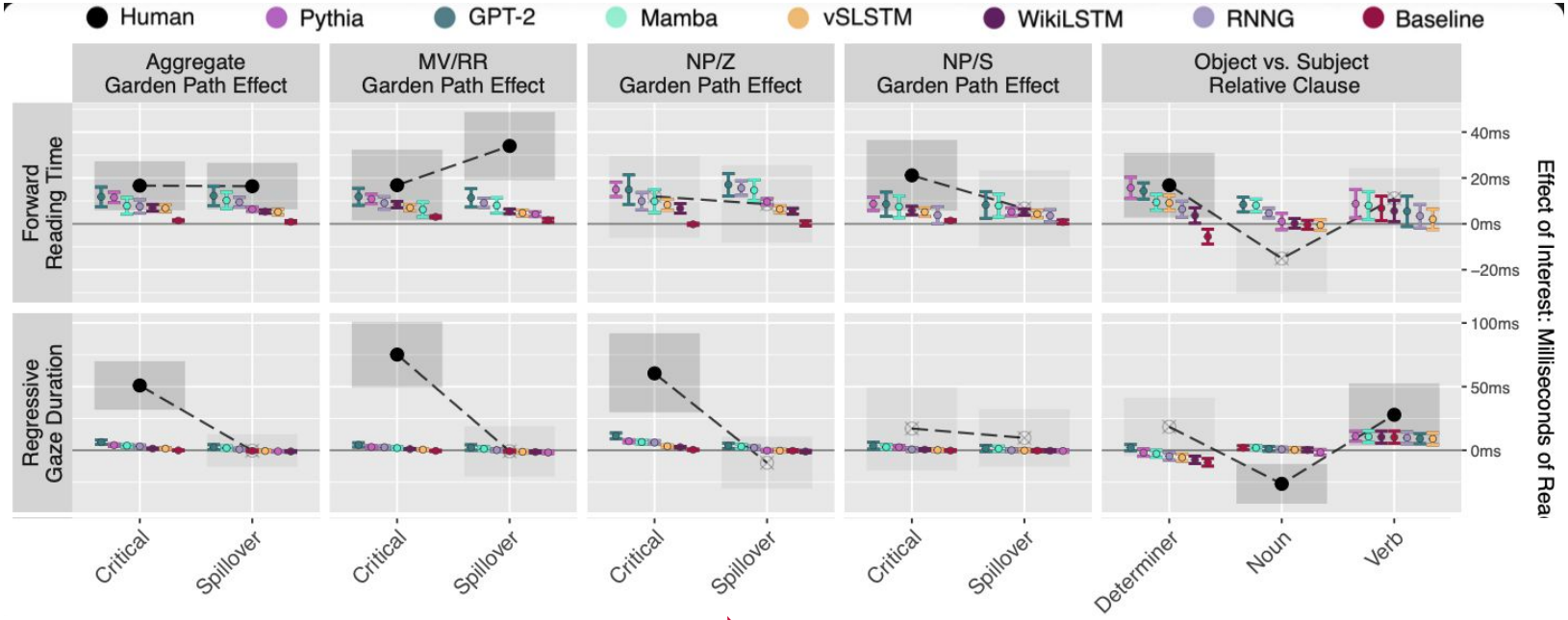


The girl found the lamb <u>was</u> <u>extremely</u> tasty.

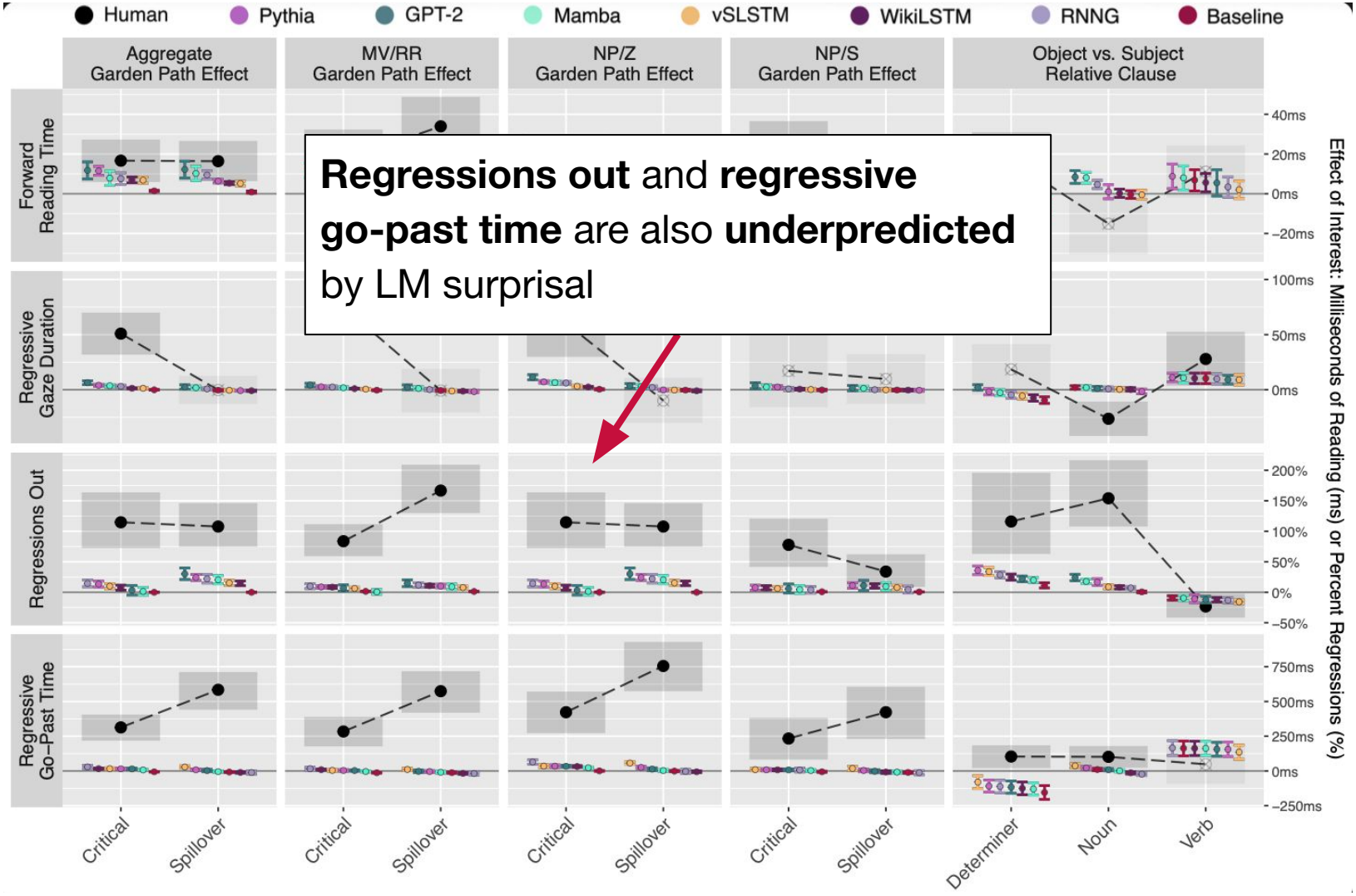
Critical Spillover



Forward reading time (gaze duration w/o regression) is **well predicted** by LM surprisal!



Regressive gaze duration (gaze duration followed by regression) is often **underpredicted** by LM surprisal


Human
Pythia
GPT-2
Mamba
vSLSTM
WikiLSTM
RNNG
Baseline
Aggregate Garden Path Effect
MV/RR Garden Path Effect
NP/Z Garden Path Effect
NP/S Garden Path Effect
Object vs. Subject Relative Clause
Forward Reading Time
Regressive Gaze Duration
Regressions Out
Regressive Go–Past Time
40ms
20ms
0ms
–20ms
100ms
50ms
0ms
200%
150%
100%
50%
0%
–50%
750ms
500ms
250ms
0ms
–250ms
Effect of Interest: Milliseconds of Reading (ms) or Percent Regressions (%)
Critical
Spillover
Determiner
Noun
Verb
Regressions out and **regressive go-past time** are also **underpredicted** by LM surprisal


Human
Pythia
GPT-2
Mamba
vSLSTM
WikiLSTM
RNNG
Baseline
Aggregate Garden Path Effect
MV/RR Garden Path Effect
NP/Z Garden Path Effect
NP/S Garden Path Effect
Object vs. Subject Relative Clause
Forward Reading Time
Regressive Gaze Duration
Regressions Out
Regressive Go–Past Time
Trials where readers did not regress
Trials where readers did regress
40ms
20ms
0ms
−20ms
100ms
50ms
0ms
200%
150%
100%
50%
0%
−50%
750ms
500ms
250ms
0ms
−250ms
Effect of Interest: Milliseconds of Reading (ms) or Percent Regressions (%)
Critical
Spillover
Critical
Spillover
Critical
Spillover
Critical
Spillover
Determiner
Noun
Verb

Interim summary

- LM surprisal **can capture** garden path effects in ‘early’ measures
- LM surprisal **can’t capture** garden path effects in ‘late’ measures
- The gap between the two is due to rereading of earlier words in the sentence
- Which words do people reread when they get garden-pathed?

Hypothesis 1: Selective Reanalysis Hypothesis

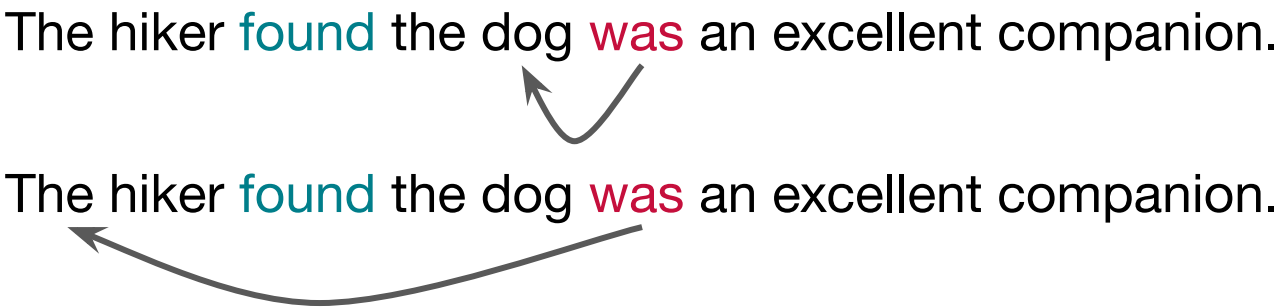
- Rereading is guided by the structural representation of the sentence
  (Frazier & Rayner, 1982; von der Malsburg & Vasishth, 2011)

- Readers keep track of uncertain parts of the sentence
  (Levy et al., 2009; Bicknell & Levy, 2010)

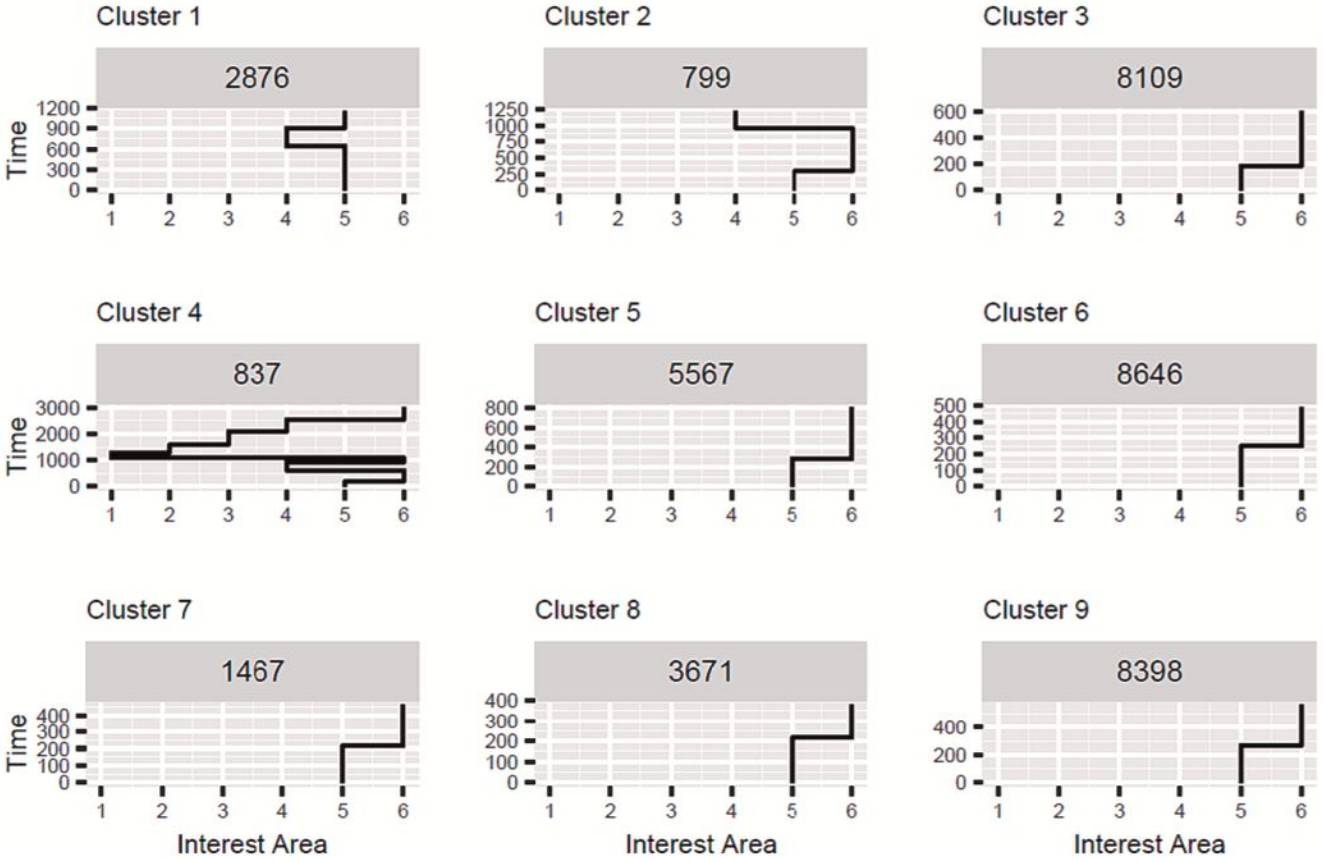
The hiker found the dog was an excellent companion.

Hypothesis 2: Time Out Hypothesis

- Rereading prevents the intake of new information and buys readers time
  (Inhoff & Weger, 2005; Mitchell et al., 2008)



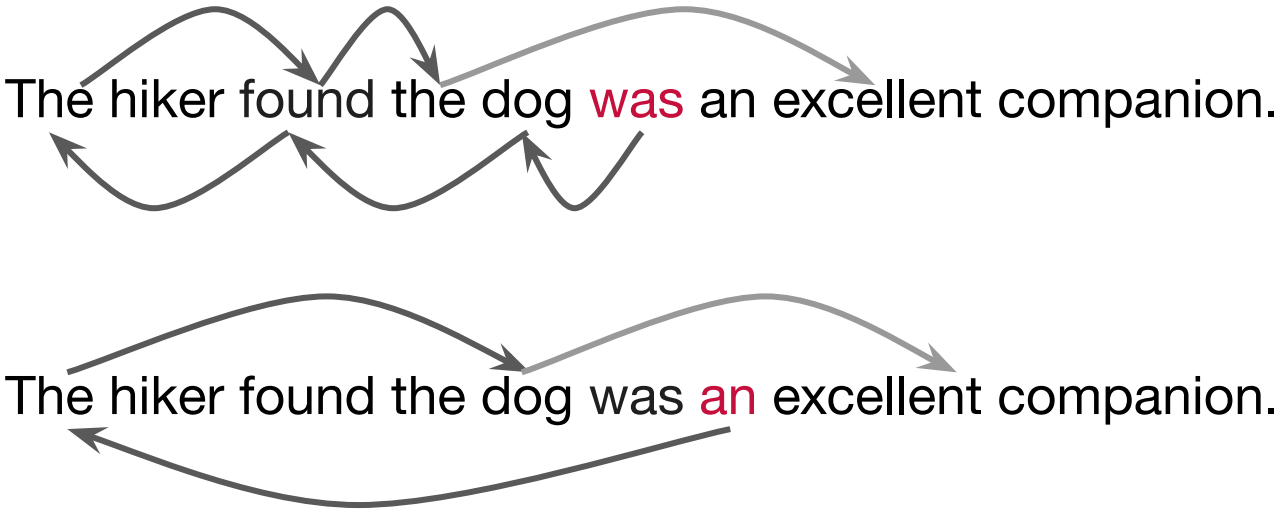
But reading behavior is highly variable



Christianson et al. (2024)

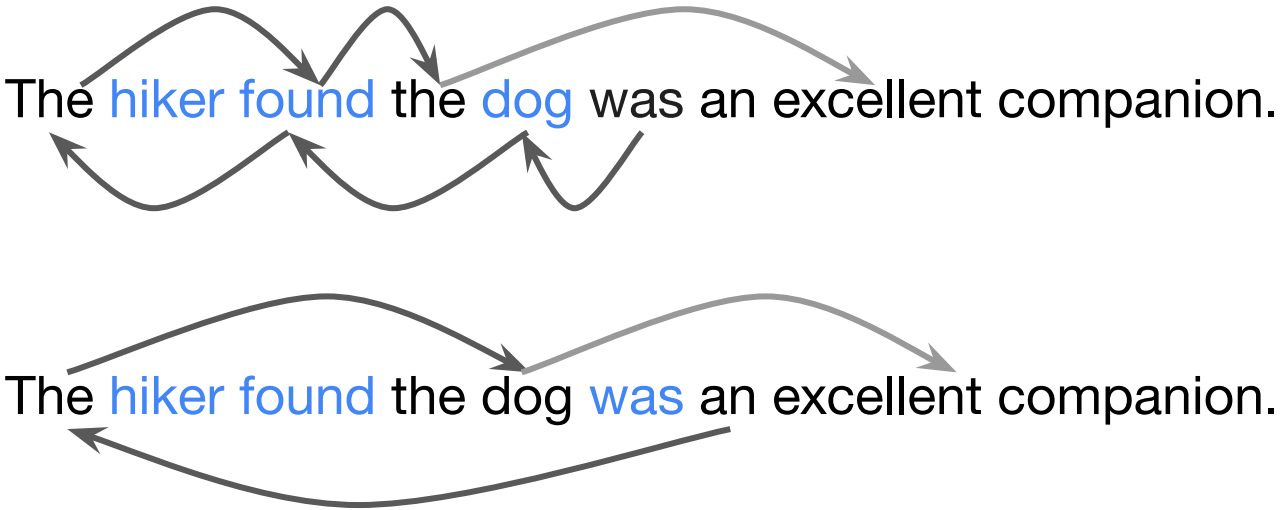
Regression modeling setup

- Dependent variable: Number of fixations on earlier word region
- Counted from two source regions: critical, spillover
- Similar assumption about where processing difficulty will manifest



Regression modeling setup

- Dependent variable: Number of fixations on earlier word region
- Counted at three target regions: previous, verb, noun
- Verb is the region predicted by SRH; previous and noun by TOH



Regression modeling setup

- Dependent variable: Number of fixations on earlier word region
- Counted at three target regions: previous, verb, noun
- Verb is the region predicted by SRH; previous and noun by TOH

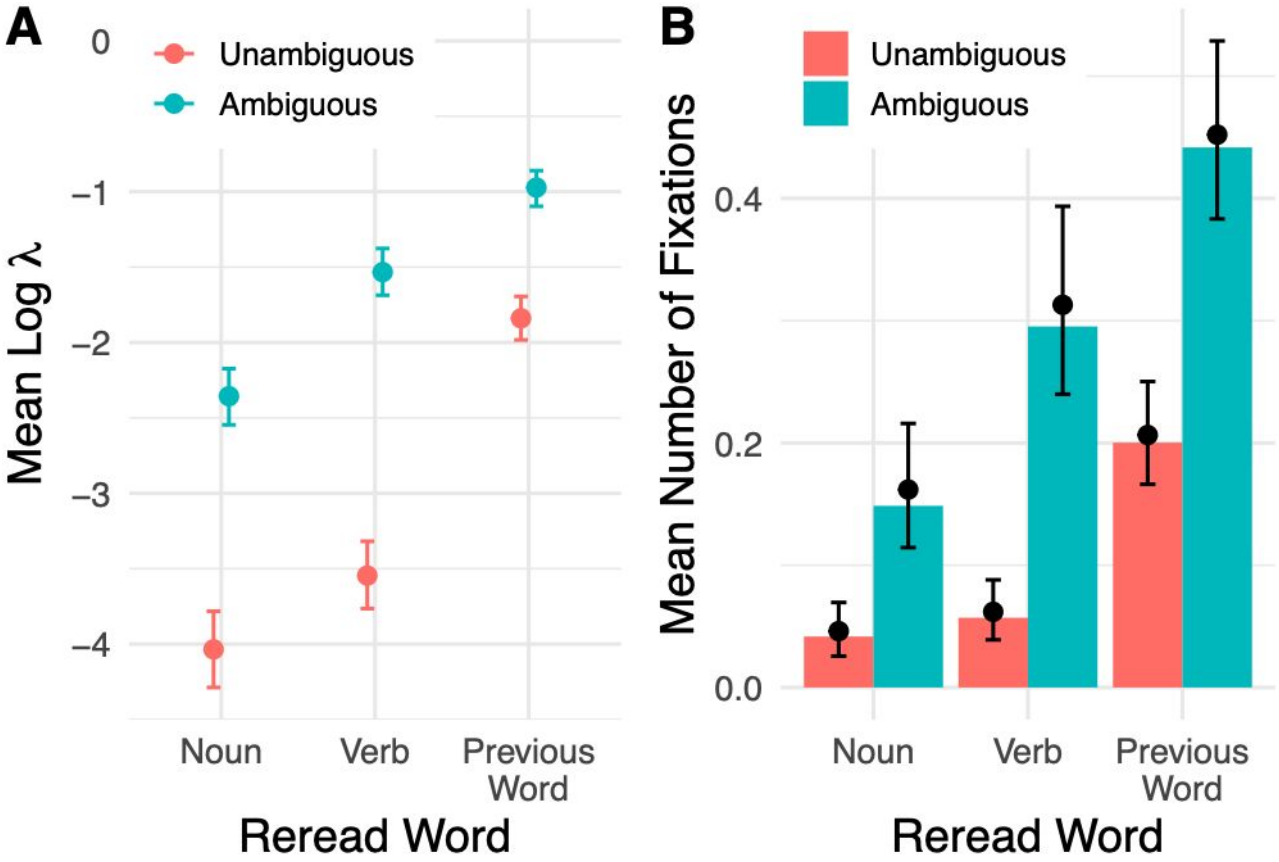
(C, P)	(C, V)	(C, N)	(S, P)	(S, V)	(S, N)
1	2	0	0	0	0

Regression modeling setup

Poisson regression. A Poisson distribution captures a counting process and is parameterized by λ, which is its expected value and variance. Its probability mass function is defined as follows:

$$\mathrm{P}(Y = y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \quad [1]$$

Is there an interaction effect between ambiguity and target region?



Ambiguity drives a larger proportional increase in fixations on the verb (**consistent with SRH**)

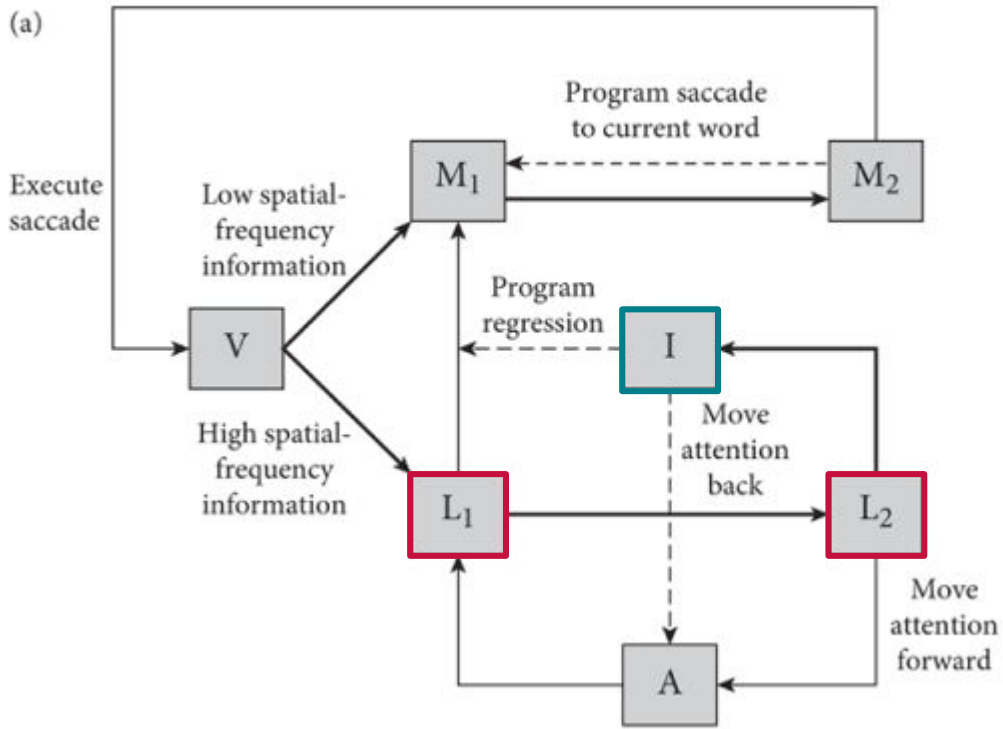
Structure dissociates from surprisal

- Surprisal underpredicts structural processing difficulty in different reading contexts (Huang et al. 2024; Kobzeva & Kush, 2024; Wilcox et al. 2021; van Schijndel & Linzen, 2019)

- Eye-tracking data show dissociable effects: Surprisal captures early measures, structural difficulty is indexed by later measures

- Readers reread words that are most helpful for amending their structural representation of the sentence

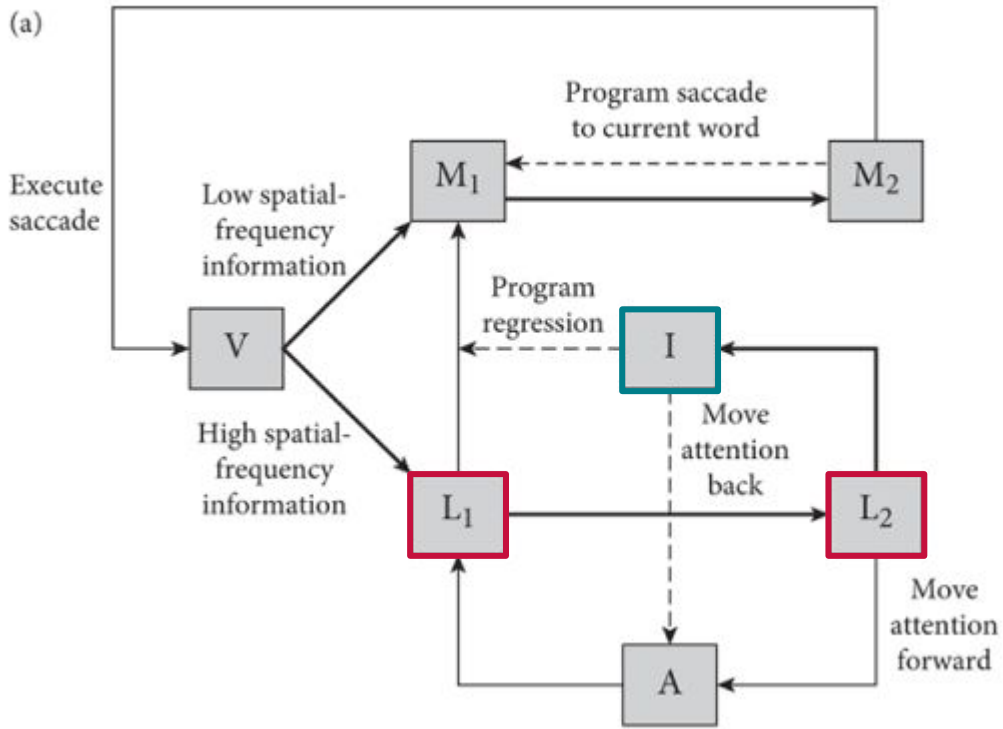
EZ Reader



- Theory of oculomotor control in reading
- Posits two stages of lexical access, determined by frequency and predictability
- Posits probability of integration failure as a construct of post-lexical difficulty

Reichle and Sheridan (2015)

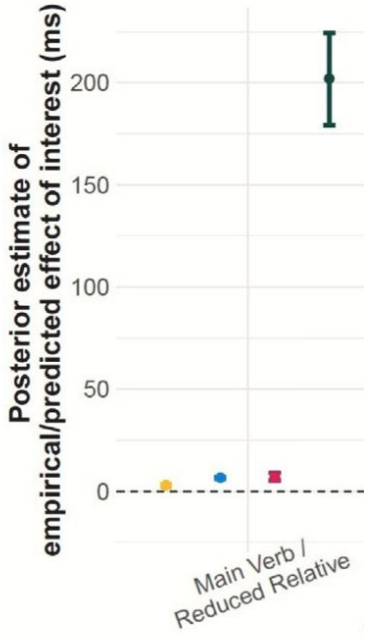
Hypothesis



- **Surprisal** impacts lexical access components
- **Structural difficulty** impacts post-lexical difficulty: Probability of integration failure

Reichle and Sheridan (2015)

New nuance to earlier findings (Huang et al. 2024)

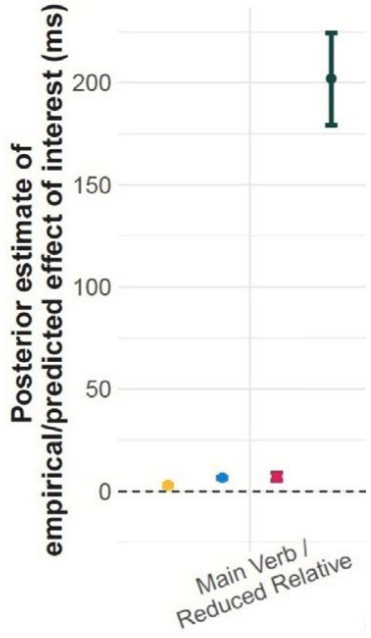


LMs are too ‘superhuman’ as surprisal estimators

Surprisal and structural processing make distinct contributions to reading difficulty

- What surprisal actually underpredicts is late eye-tracking measures
- Confirmed with a lot of data and LMs

New nuance to earlier findings (Huang et al. 2024)



LMs are too ‘superhuman’ as surprisal estimators

Surprisal and structural processing make distinct contributions to reading difficulty

But LM surprisal isn't perfect

LMs have much stronger long-term memory of training data

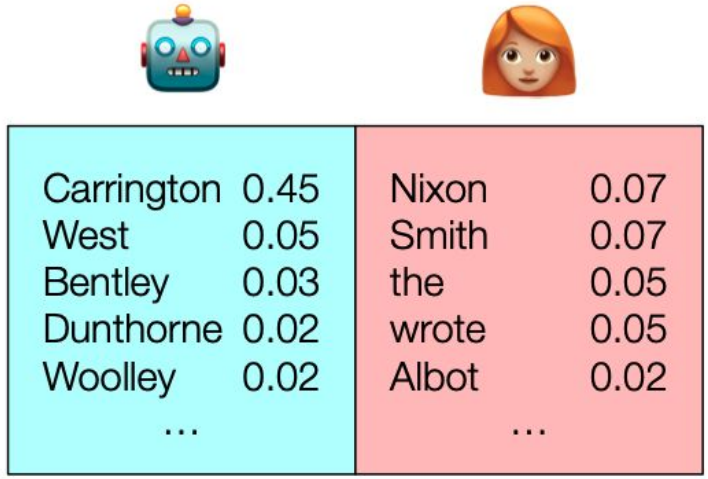


Figure 2: **Mainstream LLMs have much more factual knowledge than average humans.** Top

But LM surprisal isn't perfect

LMs have much stronger short-term memory of input text

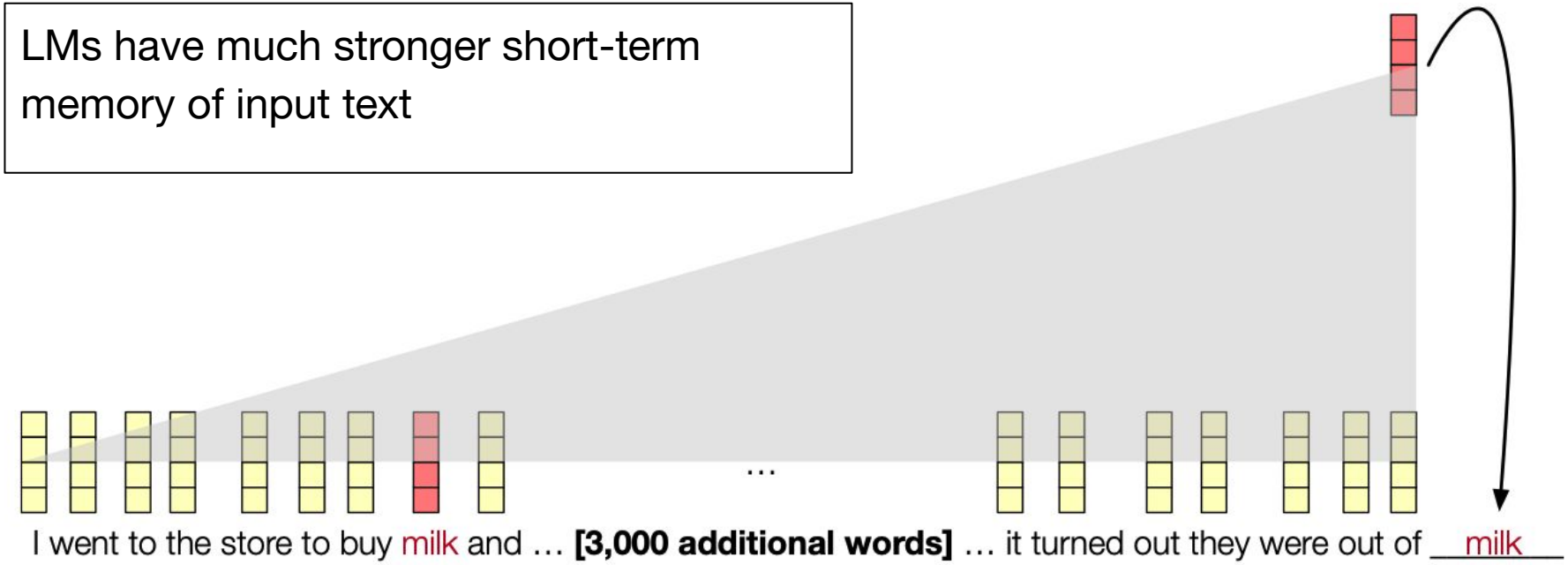


Figure 3: **Transformers have perfect access to the input sequence.** Even with thousands of

Shameless plug (https://arxiv.org/pdf/2510.05141)

To model human linguistic prediction, make LLMs less superhuman

Byung-Doh Oh and Tal Linzen
New York University
{oh.b, linzen}@nyu.edu

Interpretability agenda: LMs could entertain too many parses in parallel!

Abstract

When people listen to or read a sentence, they actively make predictions about upcoming words: words that are less predictable are generally read more slowly than predictable ones. The success of large language models (LLMs), which, like humans, make predictions about upcoming words, has motivated exploring the use of these models as cognitive models of hu-

Conclusion

Preprint coming very soon!

- Surprisal isn’t the only determinant of processing difficulty: Eye-tracking data shows structural processing difficulty dissociates from surprisal
- It may be helpful to think of two constructs: surprisal and integration failure
- LM surprisal isn’t perfect, and there is exciting work to be done for modeling human-like prediction

Thanks for listening!

oh.b@nyu.edu

NYU