Scalable Predictive Metrics Beyond Surprisal

Michael Vrazitulis

Sentence Processing Workshop
Golm, May 26-28, 2026

Introduction

Surprisal Theory (Hale 2001; Levy 2008):

$$\text{Difficulty}(w) \propto -\log P(w \mid \text{context})$$

Surprisal

Introduction

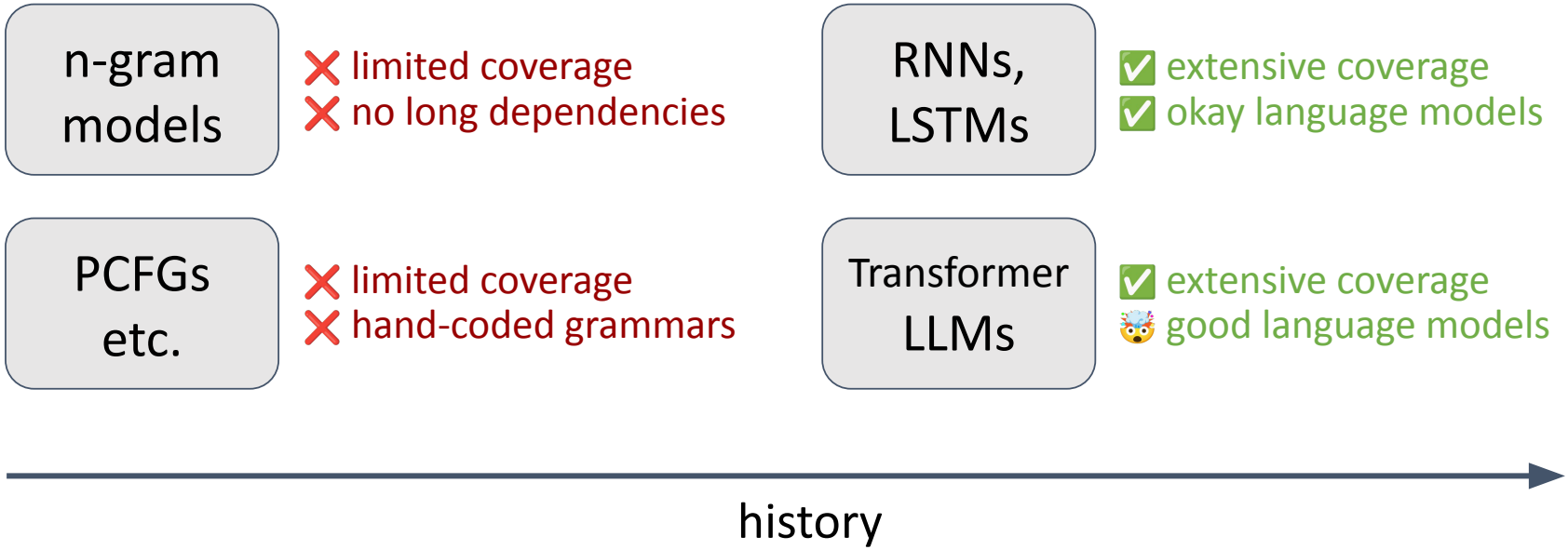
Probabilities derived from a **language model**

Surprisal Theory (Hale 2001; Levy 2008):

$$\text{Difficulty}(w) \propto -\log P(w \mid \text{context})$$

Introduction

Language modeling technology has **improved dramatically** in recent years.



Introduction

It's **easy to compute** next-word probabilities from LLMs for **arbitrary sentences.**

Transformer LLMs

✅ extensive coverage
🤯 good language models

Introduction

It's **easy to compute** next-word probabilities from LLMs for **arbitrary sentences.**

Surprisal Theory can make fine-grained **predictions** about **any construction** of interest.

Introduction

Surprisal has been shown to predict human **reading times** well (e.g., Demberg & Keller 2008; Smith & Levy 2013, Shain et al. 2024).

But it also has **limitations.**

Most notably, it **underpredicts syntactic disambiguation effects** (van Schijndel & Linzen 2021; **SAP dataset** by Huang et al. 2024).

Introduction

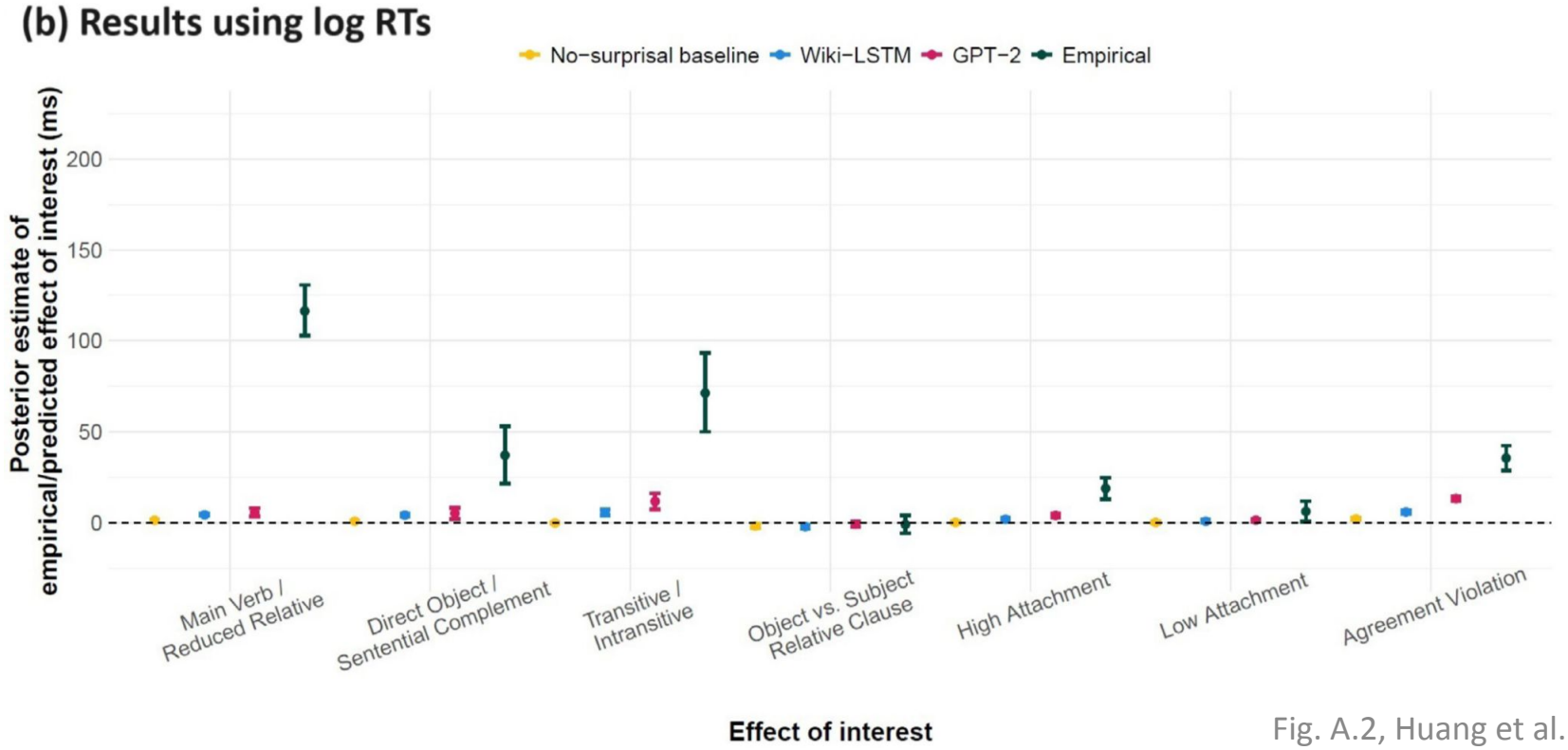


Fig. A.2, Huang et al. (2024)

Introduction

If **surprisal** is not the right non-trivial metric to **predict key sentence processing phenomena** like those in SAP, then **what is?**

```
RT ~ WordLength + WordFrequency + ???
```

Introduction

Can we compute **alternative,** theoretically motivated **metrics** that operate on the **same level as surprisal?**

Same level, i.e.:

- Word-by-word numerical predictions on arbitrary sentences
- Automatically computable (no hand-coding required)
- Linking hypothesis: simple scaling parameter to fit RTs

Introduction

Can we compute **alternative,** theoretically motivated **metrics** that operate on the **same level as surprisal?**

Well, **let's try,** and **evaluate** them on the **SAP** data!

Predictive Metrics

Here's the metrics we examined:

- (Plain) surprisal
- Resource-rational lossy context surprisal
- Attention entropy
- DLT integration cost (+ reanalysis cost)
- ACT-R interference (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics

Here's the metrics we examined:

- **(Plain) surprisal**
- Resource-rational lossy context surprisal
- Attention entropy
- DLT integration cost (+ reanalysis cost)
- ACT-R interference (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ (Plain) Surprisal

- Nothing new here
- Just **GPT-2**(-small) surprisal
- A **reference point** to compare the other, “more exciting” metrics against

Predictive Metrics

Here's the metrics we examined:

- (Plain) surprisal
- **Resource-rational lossy context surprisal**
- Attention entropy
- DLT integration cost (+ reanalysis cost)
- ACT-R interference (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ Resource-Rational Lossy Context Surprisal

- GPT-2 surprisal, but computed on a **distorted memory representation**
- **Resource-rational** distortion model (Hahn et al. 2022)
- Here: assess **20%, 50%, and 80% deletion rates** (McCurdy & Hahn 2024)

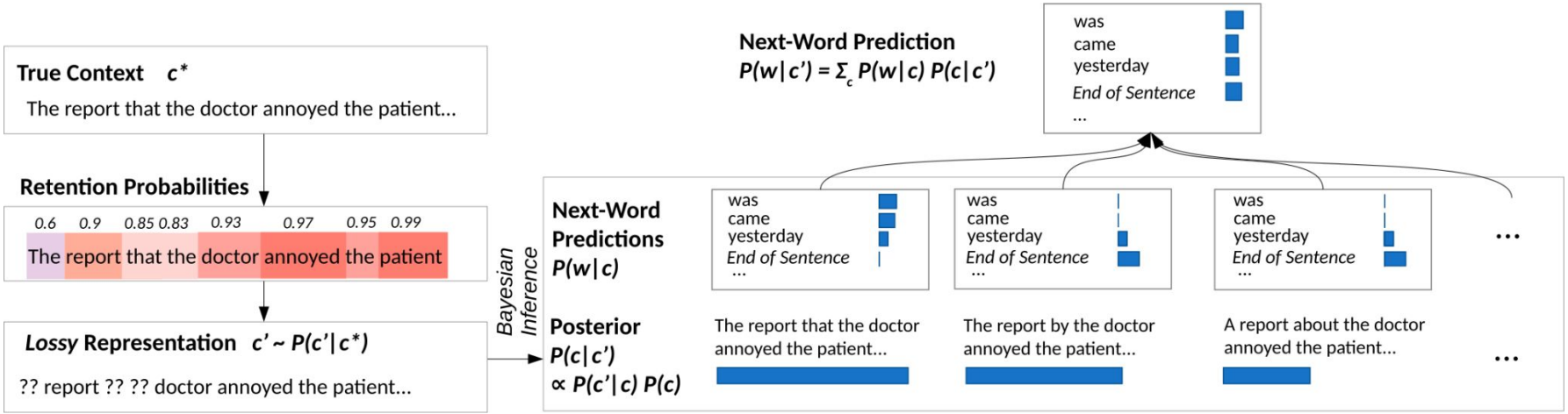


Fig. 2, Hahn et al. (2022)

Predictive Metrics

Here's the metrics we examined:

- (Plain) surprisal
- Resource-rational lossy context surprisal
- **Attention entropy**
- DLT integration cost (+ reanalysis cost)
- ACT-R interference (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ Attention Entropy

- Ryu & Lewis (2025): **Transformer attention entropy** as a proxy to **cue-based retrieval** (Lewis & Vasishth 2005)
- We use their GPT-2-based implementation, with the **same 20 attention heads** they had **selected** as relevant for **syntactic dependencies**

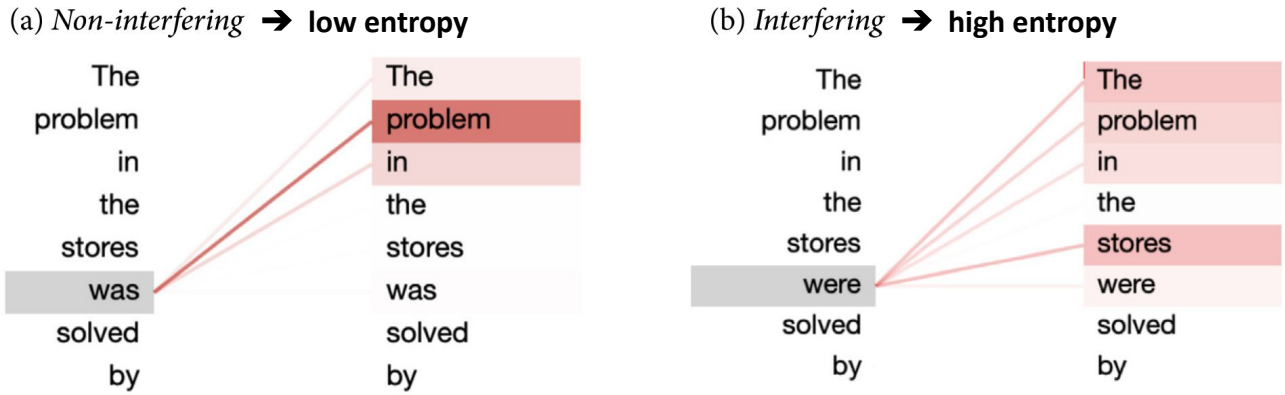
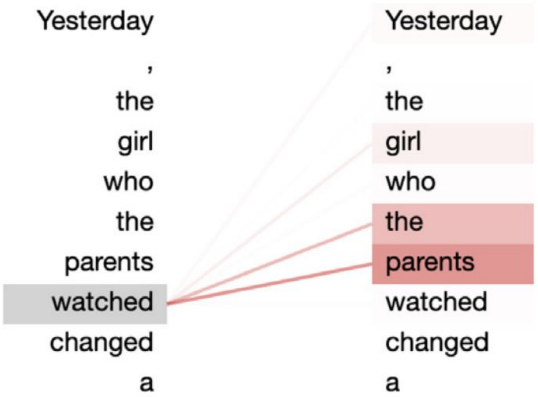


Fig. 3, Ryu & Lewis (2025)

Predictive Metrics ➔ Attention Entropy

- Ryu & Lewis (2025): **Transformer attention entropy** as a proxy to **cue-based retrieval** (Lewis & Vasishth 2005)
- We use their GPT-2-based implementation, with the **same 20 attention heads** they had **selected** as relevant for **syntactic dependencies**

(a) *ORC* ➔ **high entropy**



(b) *SRC* ➔ **low entropy**

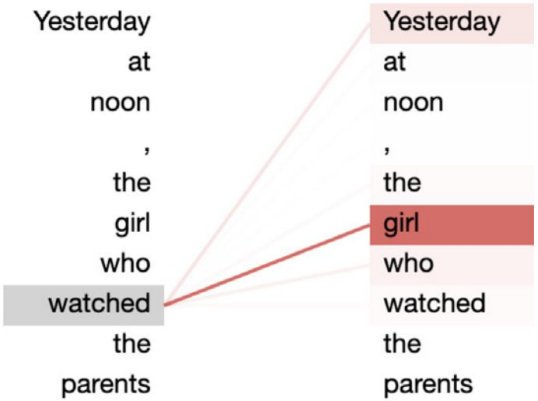


Fig. 6, Ryu & Lewis (2025)

Predictive Metrics ➔ Attention Entropy

- Ryu & Lewis (2025): **Transformer attention entropy** as a proxy to **cue-based retrieval** (Lewis & Vasishth 2005)
- We use their GPT-2-based implementation, with the **same 20 attention heads** they had **selected** as relevant for **syntactic dependencies**

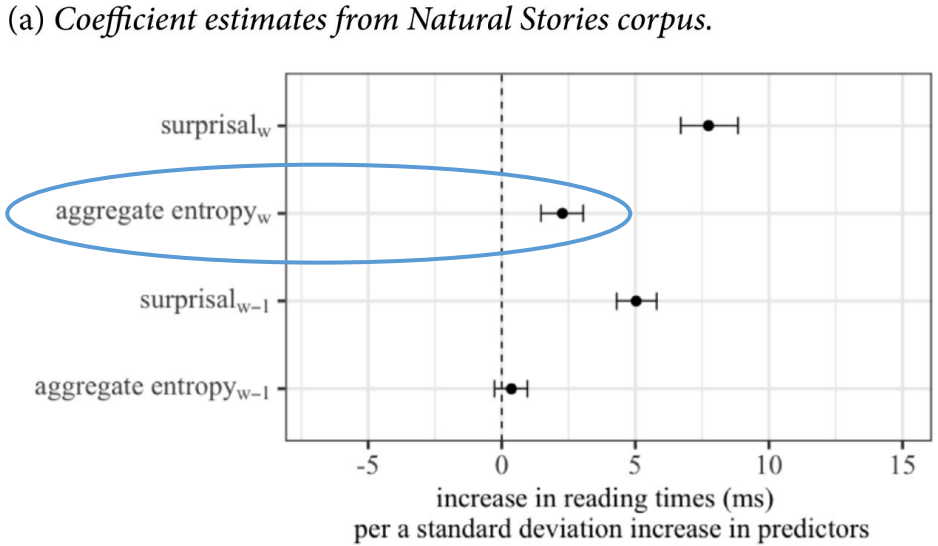


Fig. 10a, Ryu & Lewis (2025)

Predictive Metrics

Here's the metrics we examined:

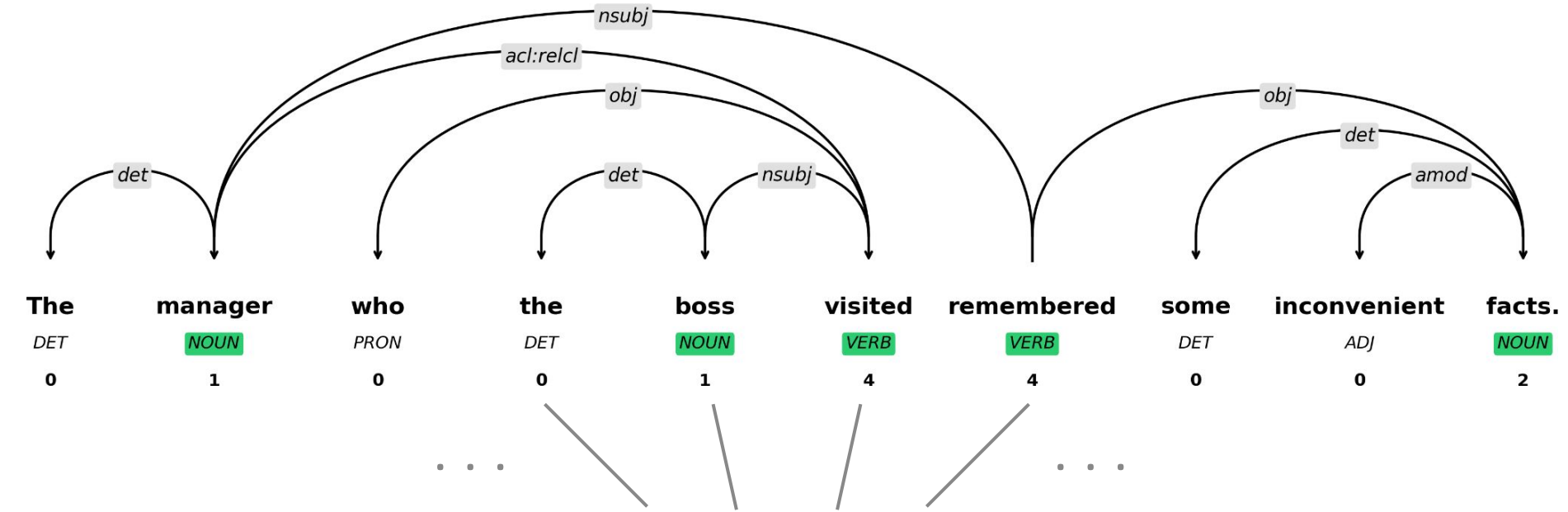
- (Plain) surprisal
- Resource-rational lossy context surprisal
- Attention entropy
- **DLT integration cost (+ reanalysis cost)**
- ACT-R interference (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ DLT Integration Cost

- Gibson (1998; 2000)'s Dependency Locality Theory (DLT):
  - Processing effort is determined by **integration cost** (and storage cost)
  - Derived by summing **dependency lengths** in terms of **# discourse referents**

Operationalize this as **automatically computable metric** with **Stanza dependency parser** (for dependency trees) and **spaCy's POS tagger** (to identify discourse referents).

Predictive Metrics ➔ DLT Integration Cost



Predictive Metrics ➔ DLT Integration Cost

One problem:

The **parse** so far **can change** as new words come in (ambiguity resolution).

So the **veridical global sentence** is **not** an **ideal basis** for computing DLT cost **at word w_n** (➔ uses future information).

Predictive Metrics ➔ DLT Integration Cost

Solution:

At each $w_1 \ldots w_n$ prefix, **pre-generate** a plausible **continuation.**

How? Use **Phi-2 LLM,** with prompt instructions to keep it short.

Then compute a **dependency parse** for **each prefix' continuation.**

Use those **"veridical-future-agnostic" parses** for **DLT cost computation** at each word.

Predictive Metrics ➔ DLT Integration Cost

During the summer I went to the beach.
During the summer I went to the beach.
During the budget meeting the CEO presented the financial report.
During the budget negotiation, the two parties reached an agreement.
During the budget negotiation, Janet and John discussed the terms.
During the budget negotiation, Janet charmed the committee members with her persuasive arguments.
During the budget negotiation, Janet charmed the committee members with her persuasive arguments.
During the budget negotiation, Janet charmed the assistant into giving her a discount.
During the budget negotiation, Janet charmed the assistant of the senator, and the senator offered her a job.
During the budget negotiation, Janet charmed the assistant of the senator, and the senator offered her a job.
During the budget negotiation, Janet charmed the assistant of the executive, and the executive was impressed by her.
During the budget negotiation, Janet charmed the assistant of the executive who was in charge of the budget.
During the budget negotiation, Janet charmed the assistant of the executive who decides the budget.
During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything.
During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything .
During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything in the company.
During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything in secret.

Predictive Metrics ➔ DLT Integration Cost

But another problem:

In pure DLT integration cost, there is still **no way to account for effort** due to a **parse change** (e.g., disambiguation).

Let's **complement** DLT cost **with a metric** tracking **parse changes…**

Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)

Simple idea:

- Go through **dependency parses** of auto-completed **continuations** at **each word**
- **How much** does the parse so far need to be **changed** between word w_{n-1} **and** w_n**?**

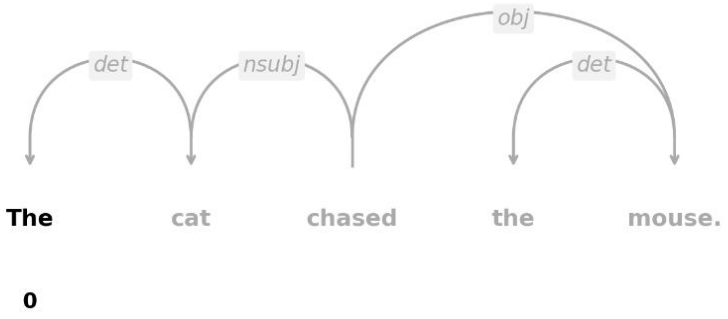
Once reaching w_n **,** did **any word** w_1 **…** w_{n-1} just receive a change to…

… the label of its incoming arrow (**upward** dependency)? Count +1 for each word

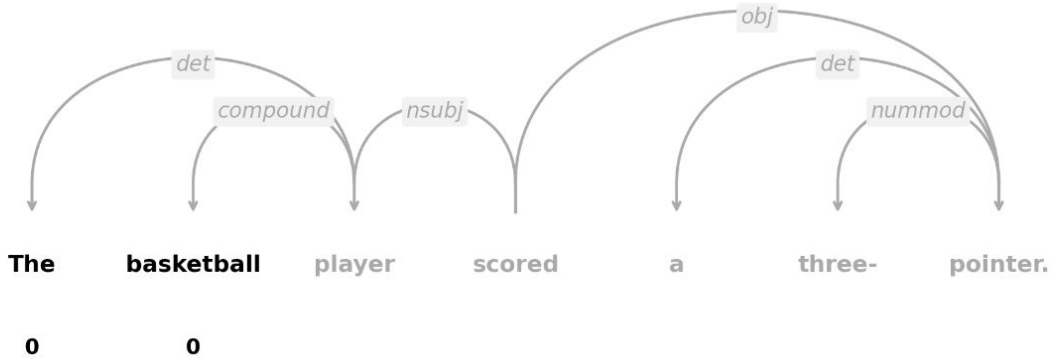
… some label of its outgoing arrows (**downward** dependencies)? Count +1 for each word

Similar ideas: Gorrell (1995), Sturt et al. (1999)

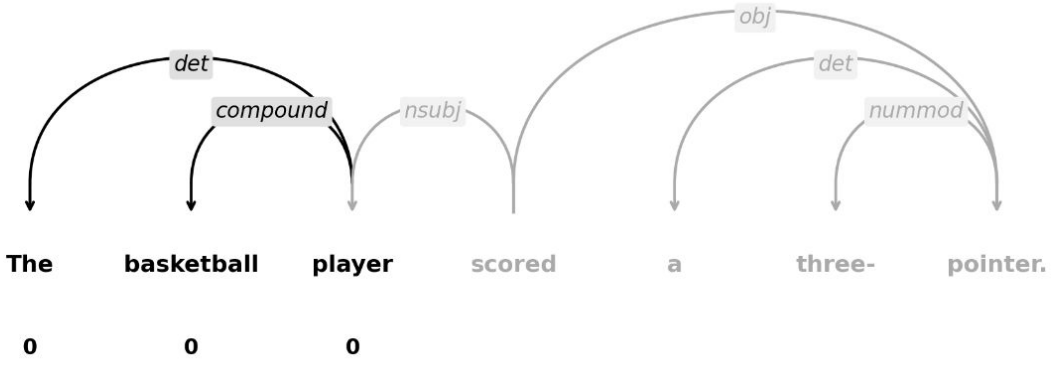
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



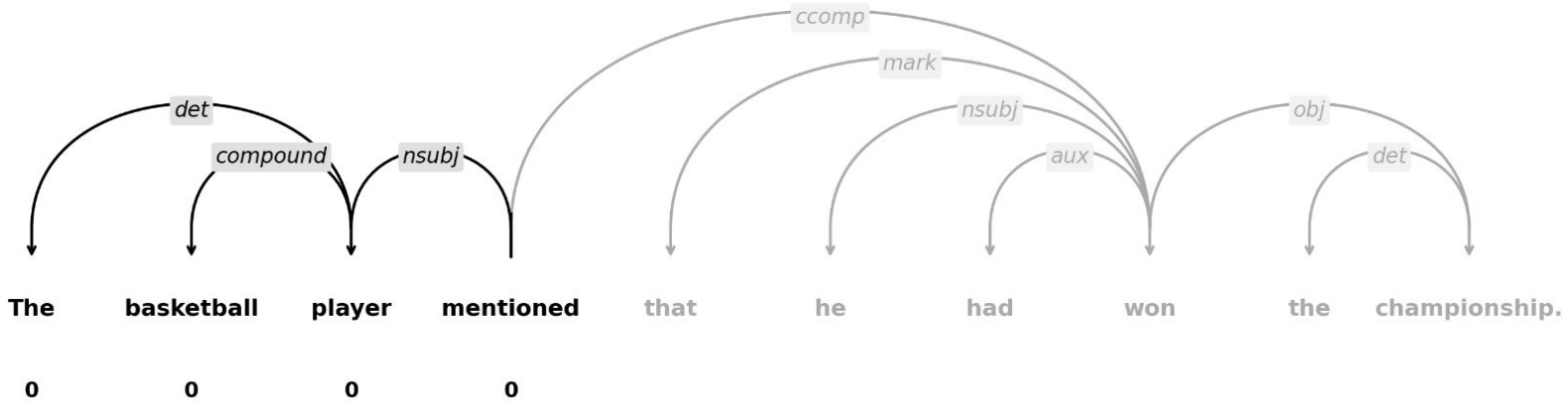
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



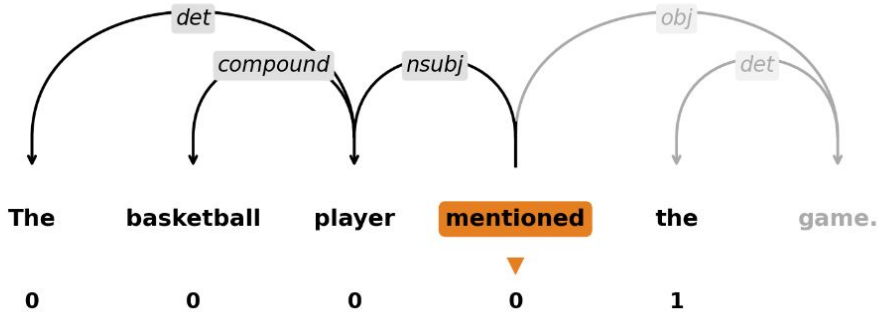
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



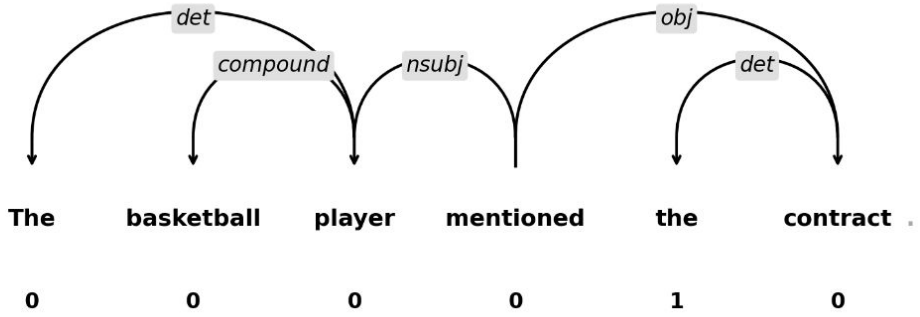
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



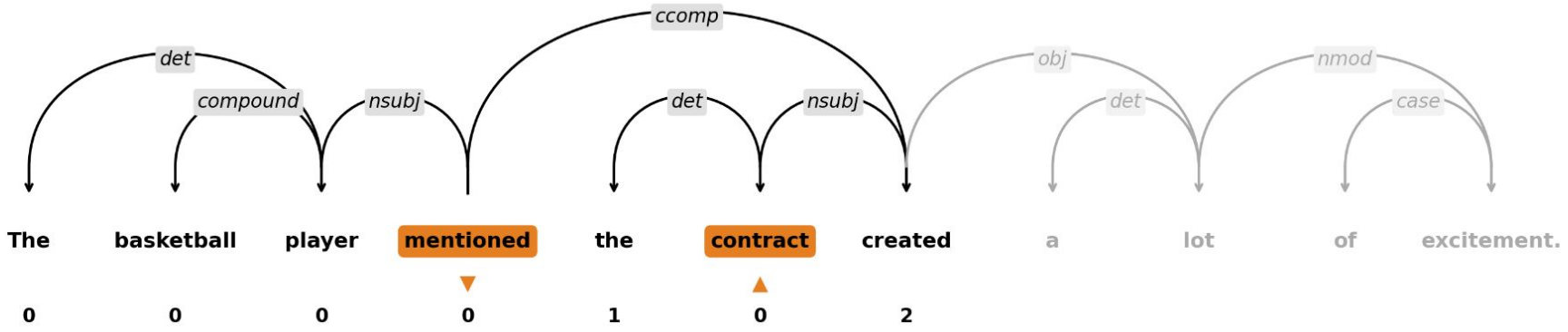
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



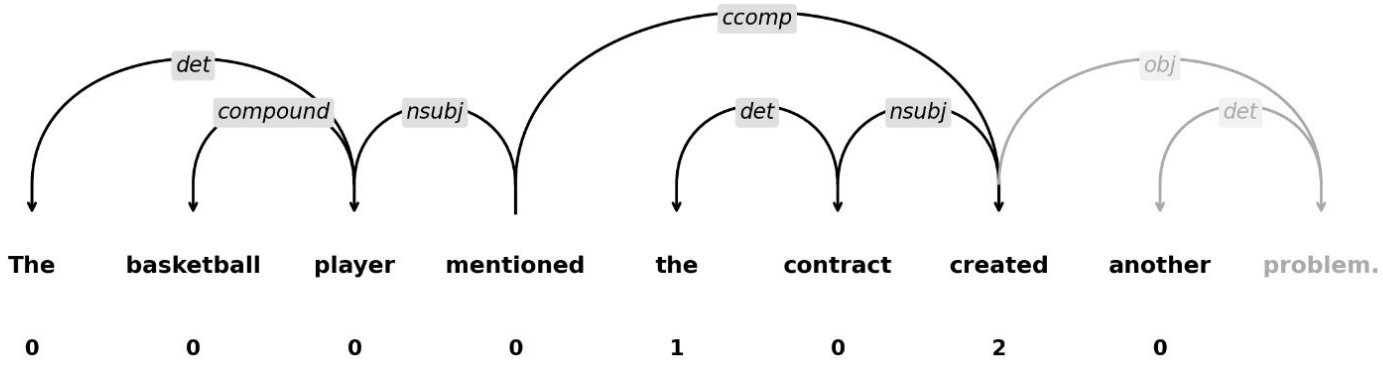
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



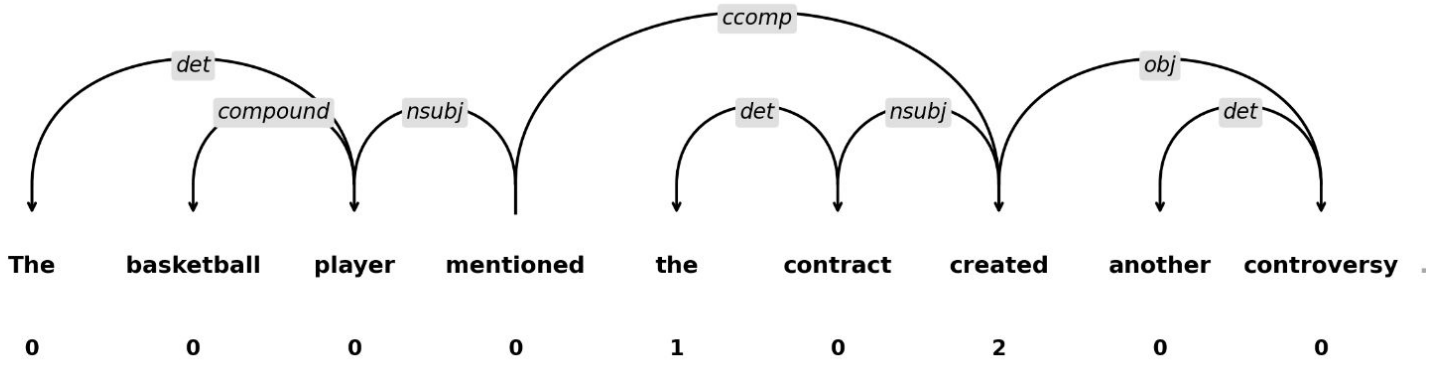
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



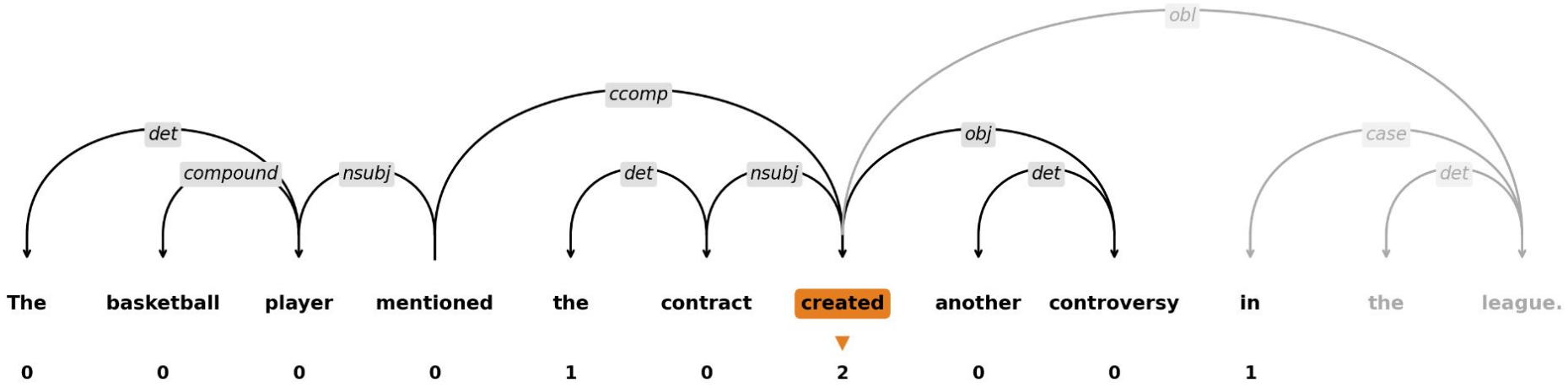
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



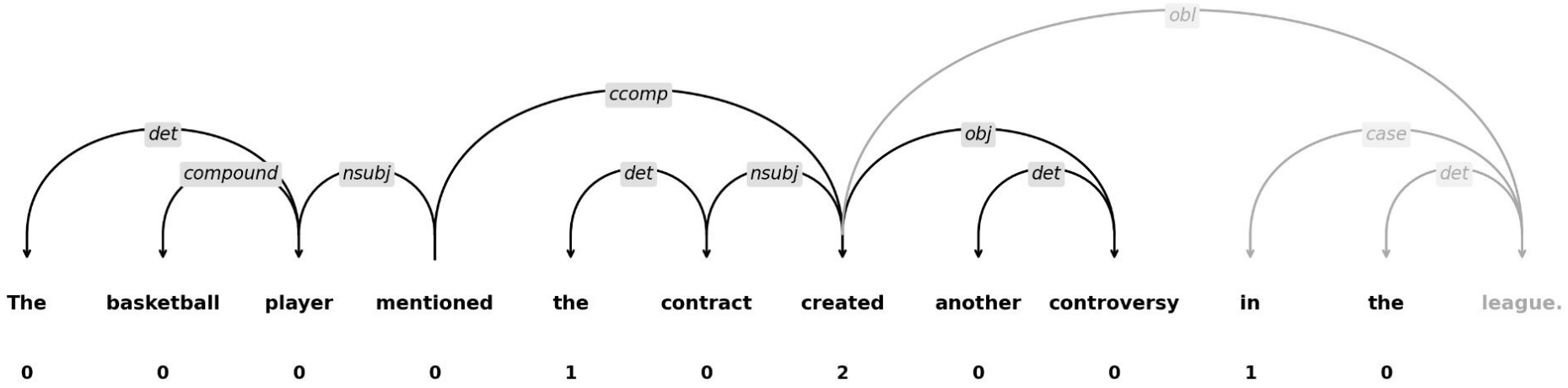
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



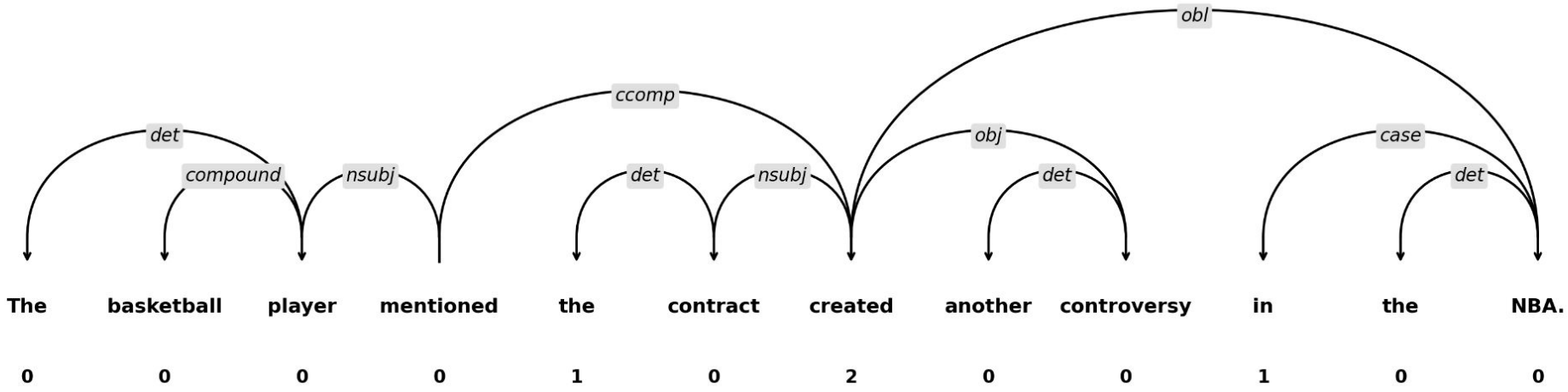
Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



Predictive Metrics ➔ Reanalysis Cost (Auxiliary Metric)



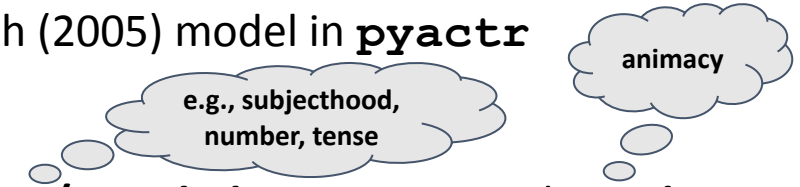
Predictive Metrics

Here's the metrics we examined:

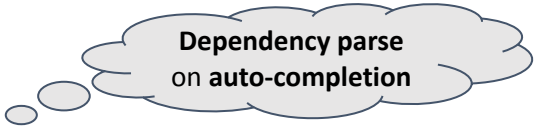
- (Plain) surprisal
- Resource-rational lossy context surprisal
- Attention entropy
- DLT integration cost (+ reanalysis cost)
- **ACT-R interference** (+ reanalysis cost)
- BERT interference (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ ACT-R Interference

- Simplified re-implementation of Lewis & Vasishth (2005) model in **`pyactr`**
- **Automatization:**
  Get categorical features from **spaCy's dependency/morphology parsers** and **WordNet**



- **Assumption:**
  If a word has a **codependent** to the left that is a **content word,** retrieval is needed!



- **Sum** of ACT-R **retrieval times** of all leftward codependents = **cost at word**
- Retrieval is susceptible to **interference** from feature-similar **intervening words**
- **Complement** with **reanalysis cost** metric (just as done for DLT)

Predictive Metrics ➔ ACT-R Interference

+subj +anim –subj –anim

The attorney who forgot about the important meeting **complains** a lot.

0 0 0 0 0.019 0 0 0 0.0435 0 0

+subj +anim –subj +anim

The attorney who forgot about the important visitor **complains** a lot.

0 0 0 0 0.019 0 0 0 0.0544 0 0

+subj +anim +subj –anim

The attorney who forgot that the meeting was important **complains** a lot.

0 0 0 0 0 0 0 0.0435 0.0672 0.0746 0 0

+subj +anim +subj +anim

The attorney who forgot about the visitor was important **complains** a lot.

0 0 0 0 0 0 0 0.0435 0.0672 0.0933 0 0

Predictive Metrics

Here's the metrics we examined:

- (Plain) surprisal
- Resource-rational lossy context surprisal
- Attention entropy
- DLT integration cost (+ reanalysis cost)
- ACT-R interference (+ reanalysis cost)
- **BERT interference** (+ dependency distance + reanalysis cost)

Predictive Metrics ➔ BERT Interference + Dependency Distance

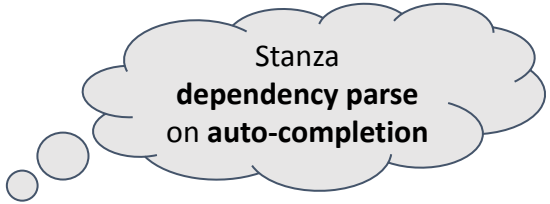
- Interference metric based on similarities between **BERT word embeddings**
- **Similarity**(Emb1, Emb2) = **max**(0, **CosineSim**(Emb1, Emb2))
- **Interference** = $\Sigma_{i \in interferents}$ **Similarity**(Emb_i , $Emb_{retrievalTarget}$)
- **Difficulty** = **Interference** + β **log(Position** $_{currentWord}$ **– Position** $_{retrievalTarget}$ **)**
- **SumDifficulty** = $\Sigma_{r \in retrievalTargets}$ **Difficulty** $_r$

Relative weighting interference vs. distance ⇒ equivalent to **two** separate **predictors** in **linear model**

BERT: Devlin et al. (2019)

Predictive Metrics ➔ BERT Interference + Dependency Distance

Some caveats:



- Only select **left codependents** as **retrieval targets** if they are **content words**
- Only words with **same spaCy POS tag** as retrieval target are **eligible** as **interferents**
- Further **complement** with **reanalysis cost**

The SAP Self-Paced Reading Dataset (Huang et al. 2024)

- **N = 2,000 participants** (post-exclusion)
- **Included experimental designs:**
  - MV/RR garden paths `(ambiguous | unambiguous)`
  - NP/S garden paths `(ambiguous | unambiguous)`
  - NP/Z garden paths `(ambiguous | unambiguous)`
  - Relative clause asymmetry `(subject | object)`
  - Attachment ambiguity `(high | low | ambiguous)`
  - Agreement violation `(violation | no-violation)`
- Some trials were followed by **comprehension questions**
- They also included **naturalistic fillers,** but we ignore these here

Statistical Models

Compare these **log-normal** models on **spillover-region reading times** (frequentist):

Baseline

RT ~ trialOrder + lengthCritical + lengthSpillover + logFreqCritical + logFreqSpillover **+ (1|subject) + (1|item) + (1|design)**

Surprisal

Baseline **+ surprisalCritical + surprisalSpillover** + (**surprisalCritical + surprisalSpillover** || subject)

RR-LCS with 20% deletion rate

Baseline **+ RRLCS20pCritical + RRLCS20pSpillover** + (**RRLCS20pCritical + RRLCS20pSpillover** || subject)

RR-LCS with 50% deletion rate

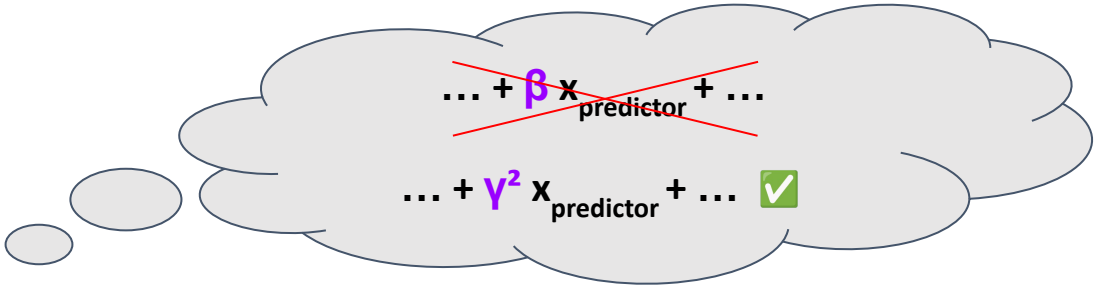
Baseline **+ RRLCS50pCritical + RRLCS50pSpillover** + (**RRLCS50pCritical + RRLCS50pSpillover** || subject)

RR-LCS with 80% deletion rate

Baseline **+ RRLCS80pCritical + RRLCS80pSpillover** + (**RRLCS80pCritical + RRLCS80pSpillover** || subject)

• • •

Statistical Models



predictor := **coefficient** forced to be positive

Surprisal

Baseline + surprisalCritical + surprisalSpillover + (surprisalCritical + surprisalSpillover || subject)

RR-LCS with 20% deletion rate

Baseline + RRLCS20pCritical + RRLCS20pSpillover + (RRLCS20pCritical + RRLCS20pSpillover || subject)

RR-LCS with 50% deletion rate

Baseline + RRLCS50pCritical + RRLCS50pSpillover + (RRLCS50pCritical + RRLCS50pSpillover || subject)

RR-LCS with 80% deletion rate

Baseline + RRLCS80pCritical + RRLCS80pSpillover + (RRLCS80pCritical + RRLCS80pSpillover || subject)

• • •

Statistical Models

Compare these **log-normal** models on **spillover-region reading times** (frequentist):

• • •

Attention entropy

Baseline + attEntropyCritical + attEntropySpillover + (attEntropyCritical + attEntropySpillover || subject)

DLT integration cost + reanalysis cost

Baseline + DLTIntgrCritical + DLTIntgrSpillover + ReanalysisCritical + ReanalysisSpillover
+ (DLTIntgrCritical + DLTIntgrSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

ACT-R interference + reanalysis cost

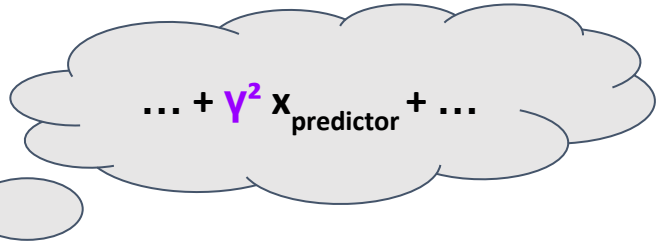
Baseline + ACTRIntfCritical + ACTRIntfSpillover + ReanalysisCritical + ReanalysisSpillover
+ (ACTRIntfCritical + ACTRIntfSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

BERT interference + dependency distance + reanalysis cost

Baseline + BERTIntfCritical + BERTIntfSpillover + DepDistCritical + DepDistSpillover
+ ReanalysisCritical + ReanalysisSpillover + (BERTIntfCritical + BERTIntfSpillover
+ DepDistCritical + DepDistSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

Statistical Models

predictor := **coefficient** forced to be positive



Attention entropy

Baseline + attEntropyCritical + attEntropySpillover + (attEntropyCritical + attEntropySpillover || subject)

DLT integration cost + reanalysis cost

Baseline + DLTIntgrCritical + DLTIntgrSpillover + ReanalysisCritical + ReanalysisSpillover
+ (DLTIntgrCritical + DLTIntgrSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

ACT-R interference + reanalysis cost

Baseline + ACTRIntfCritical + ACTRIntfSpillover + ReanalysisCritical + ReanalysisSpillover
+ (ACTRIntfCritical + ACTRIntfSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

BERT interference + dependency distance + reanalysis cost

Baseline + BERTIntfCritical + BERTIntfSpillover + DepDistCritical + DepDistSpillover
+ ReanalysisCritical + ReanalysisSpillover + (BERTIntfCritical + BERTIntfSpillover
+ DepDistCritical + DepDistSpillover + ReanalysisCritical + ReanalysisSpillover || subject)

Statistical Models

Our approach, for each predictive metric…

Spillover regions from…

- MV/RR garden paths
- NP/Z garden paths
- NP/S garden paths
- Relative clause asymmetry
- Attachment ambiguity
- Agreement violation

Fit **one unified model** across designs, with **by-design random intercepts.**

Then **predict differences between conditions** of each design.

- MV/RR garden paths
- NP/Z garden paths
- NP/S garden paths
- Relative clause asymmetry
- Attachment ambiguity
- Agreement violation

Statistical Models

…is different from Huang et al. (2024)'s:

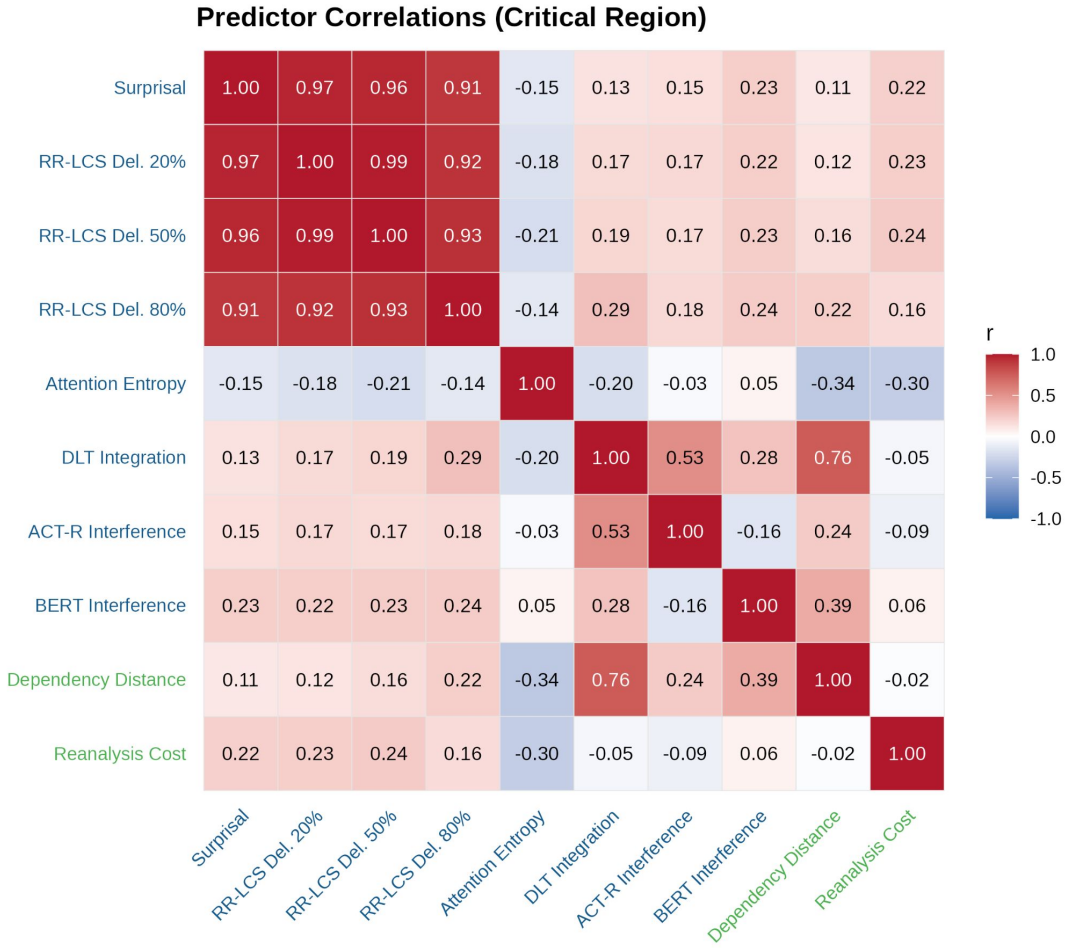
naturalistic filler sentences
(all words except first 3 and last 1)

Fit **a model on fillers.**

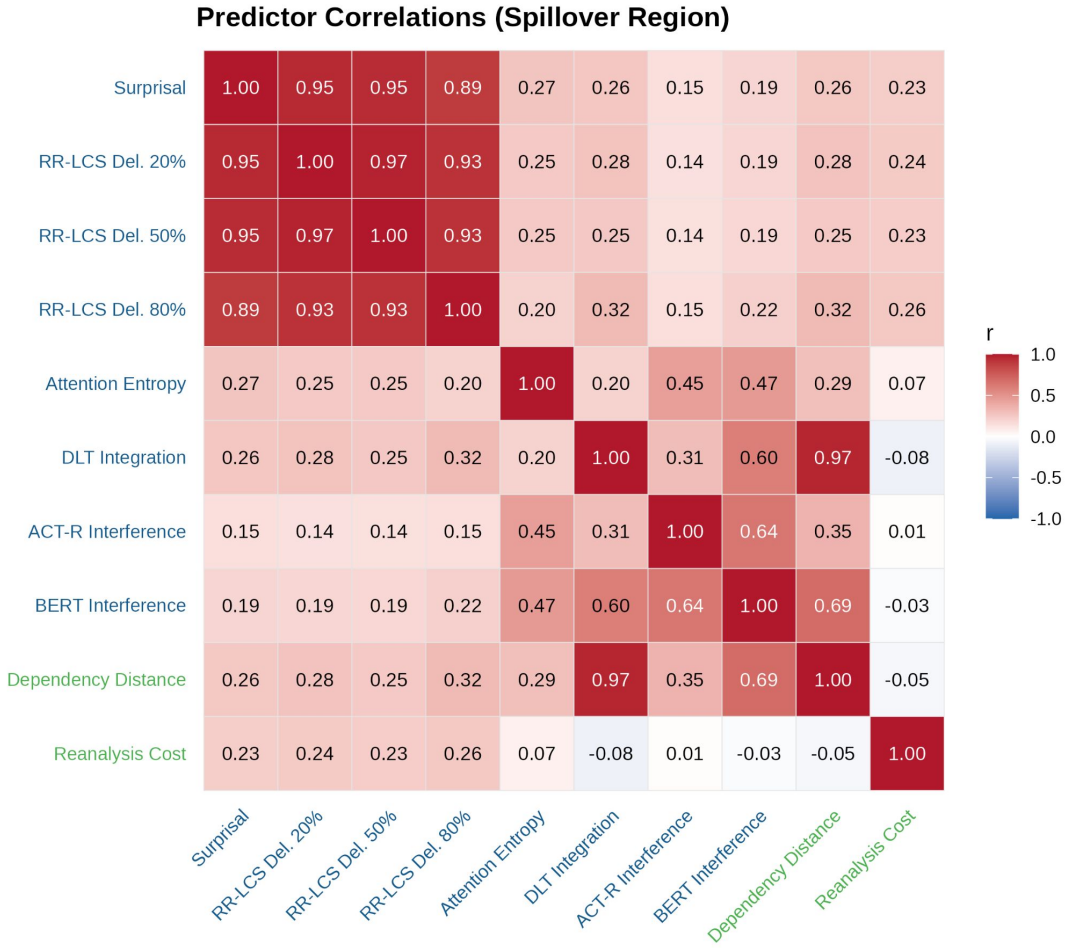
Then **predict differences between conditions** of each design.

- MV/RR garden paths
- NP/Z garden paths
- NP/S garden paths
- Relative clause asymmetry
- Attachment ambiguity
- Agreement violation

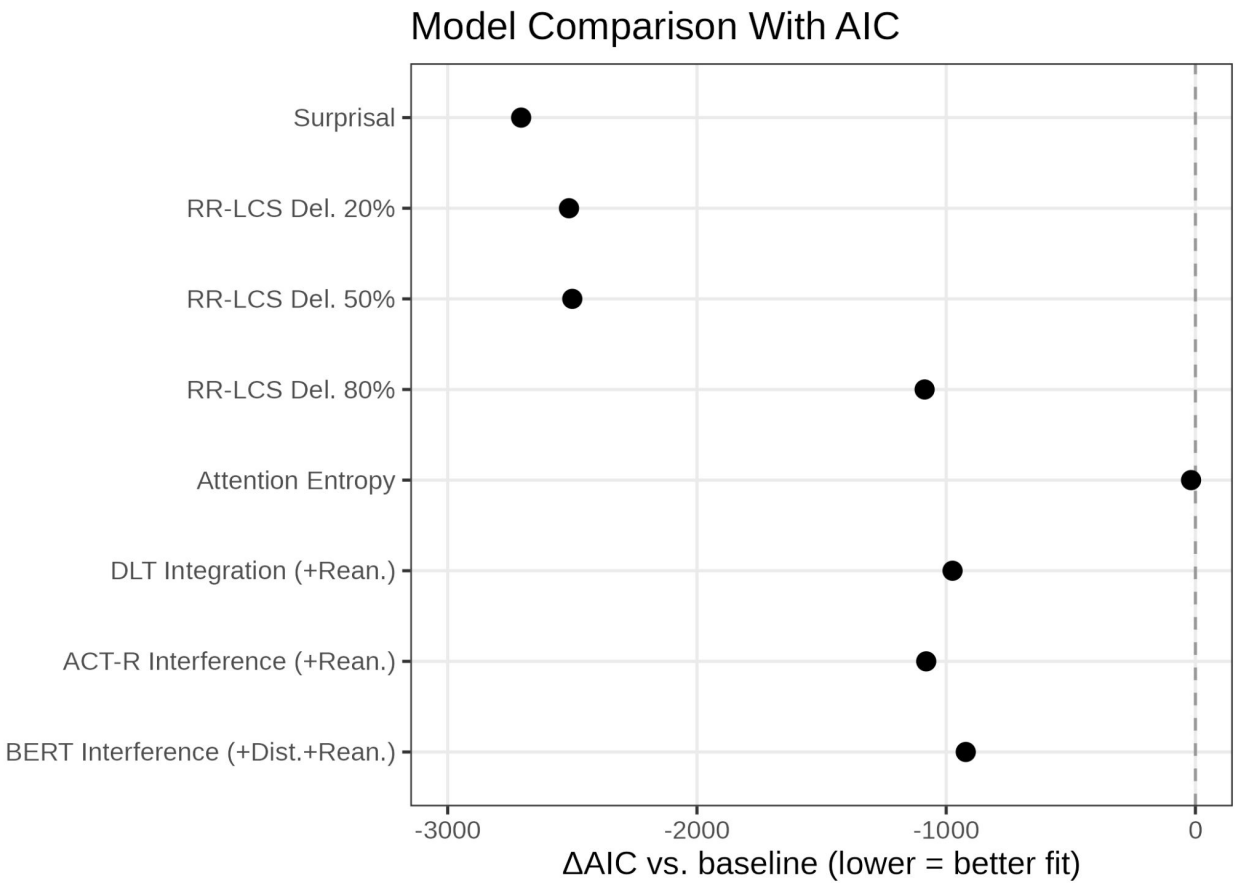
Results


Predictor Correlations (Critical Region)
Surprisal 1.00 0.97 0.96 0.91 -0.15 0.13 0.15 0.23 0.11 0.22
RR-LCS Del. 20% 0.97 1.00 0.99 0.92 -0.18 0.17 0.17 0.22 0.12 0.23
RR-LCS Del. 50% 0.96 0.99 1.00 0.93 -0.21 0.19 0.17 0.23 0.16 0.24
RR-LCS Del. 80% 0.91 0.92 0.93 1.00 -0.14 0.29 0.18 0.24 0.22 0.16
Attention Entropy -0.15 -0.18 -0.21 -0.14 1.00 -0.20 -0.03 0.05 -0.34 -0.30
DLT Integration 0.13 0.17 0.19 0.29 -0.20 1.00 0.53 0.28 0.76 -0.05
ACT-R Interference 0.15 0.17 0.17 0.18 -0.03 0.53 1.00 -0.16 0.24 -0.09
BERT Interference 0.23 0.22 0.23 0.24 0.05 0.28 -0.16 1.00 0.39 0.06
Dependency Distance 0.11 0.12 0.16 0.22 -0.34 0.76 0.24 0.39 1.00 -0.02
Reanalysis Cost 0.22 0.23 0.24 0.16 -0.30 -0.05 -0.09 0.06 -0.02 1.00
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration
ACT-R Interference
BERT Interference
Dependency Distance
Reanalysis Cost
r
1.0
0.5
0.0
-0.5
-1.0

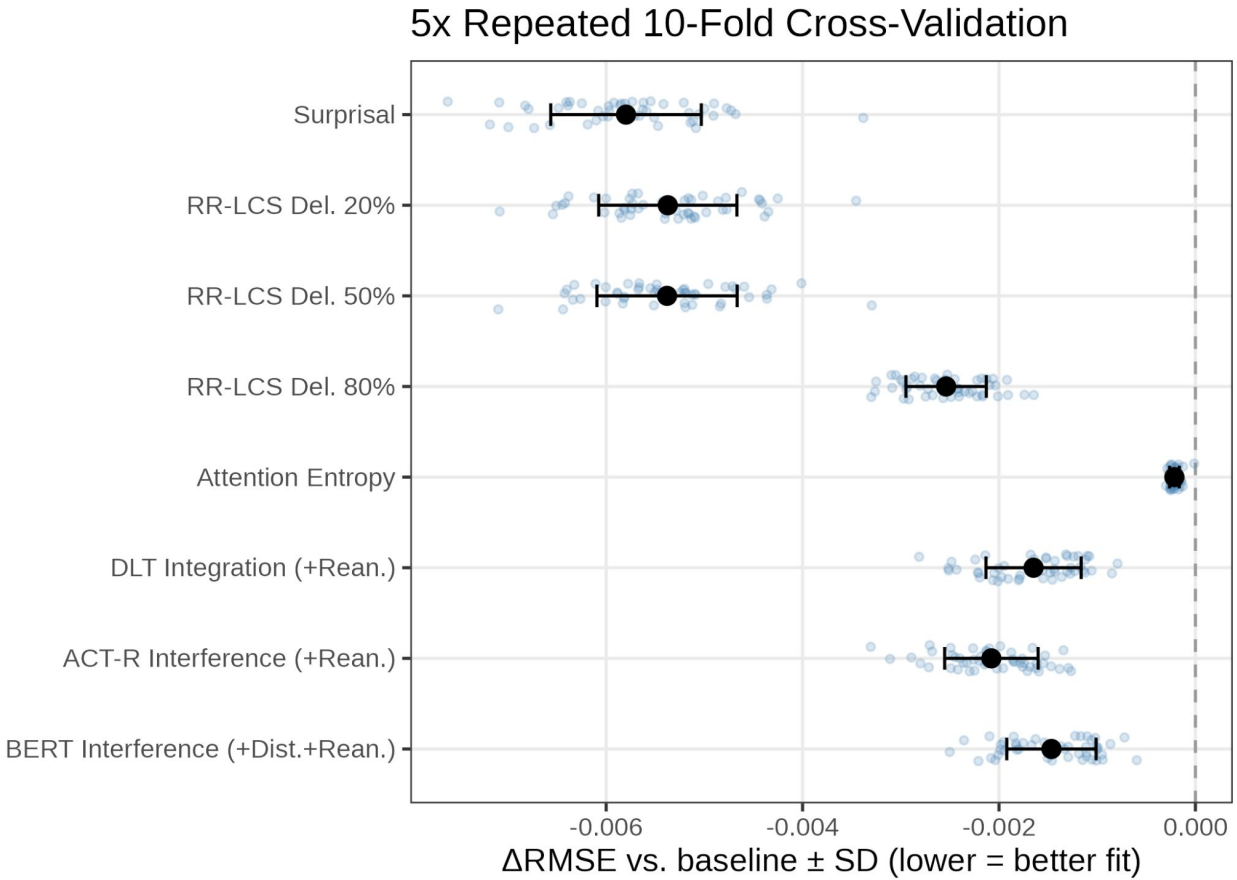
Results


Predictor Correlations (Spillover Region)
Surprisal 1.00 0.95 0.95 0.89 0.27 0.26 0.15 0.19 0.26 0.23
RR-LCS Del. 20% 0.95 1.00 0.97 0.93 0.25 0.28 0.14 0.19 0.28 0.24
RR-LCS Del. 50% 0.95 0.97 1.00 0.93 0.25 0.25 0.14 0.19 0.25 0.23
RR-LCS Del. 80% 0.89 0.93 0.93 1.00 0.20 0.32 0.15 0.22 0.32 0.26
Attention Entropy 0.27 0.25 0.25 0.20 1.00 0.20 0.45 0.47 0.29 0.07
DLT Integration 0.26 0.28 0.25 0.32 0.20 1.00 0.31 0.60 0.97 -0.08
ACT-R Interference 0.15 0.14 0.14 0.15 0.45 0.31 1.00 0.64 0.35 0.01
BERT Interference 0.19 0.19 0.19 0.22 0.47 0.60 0.64 1.00 0.69 -0.03
Dependency Distance 0.26 0.28 0.25 0.32 0.29 0.97 0.35 0.69 1.00 -0.05
Reanalysis Cost 0.23 0.24 0.23 0.26 0.07 -0.08 0.01 -0.03 -0.05 1.00
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration
ACT-R Interference
BERT Interference
Dependency Distance
Reanalysis Cost
r
1.0
0.5
0.0
-0.5
-1.0

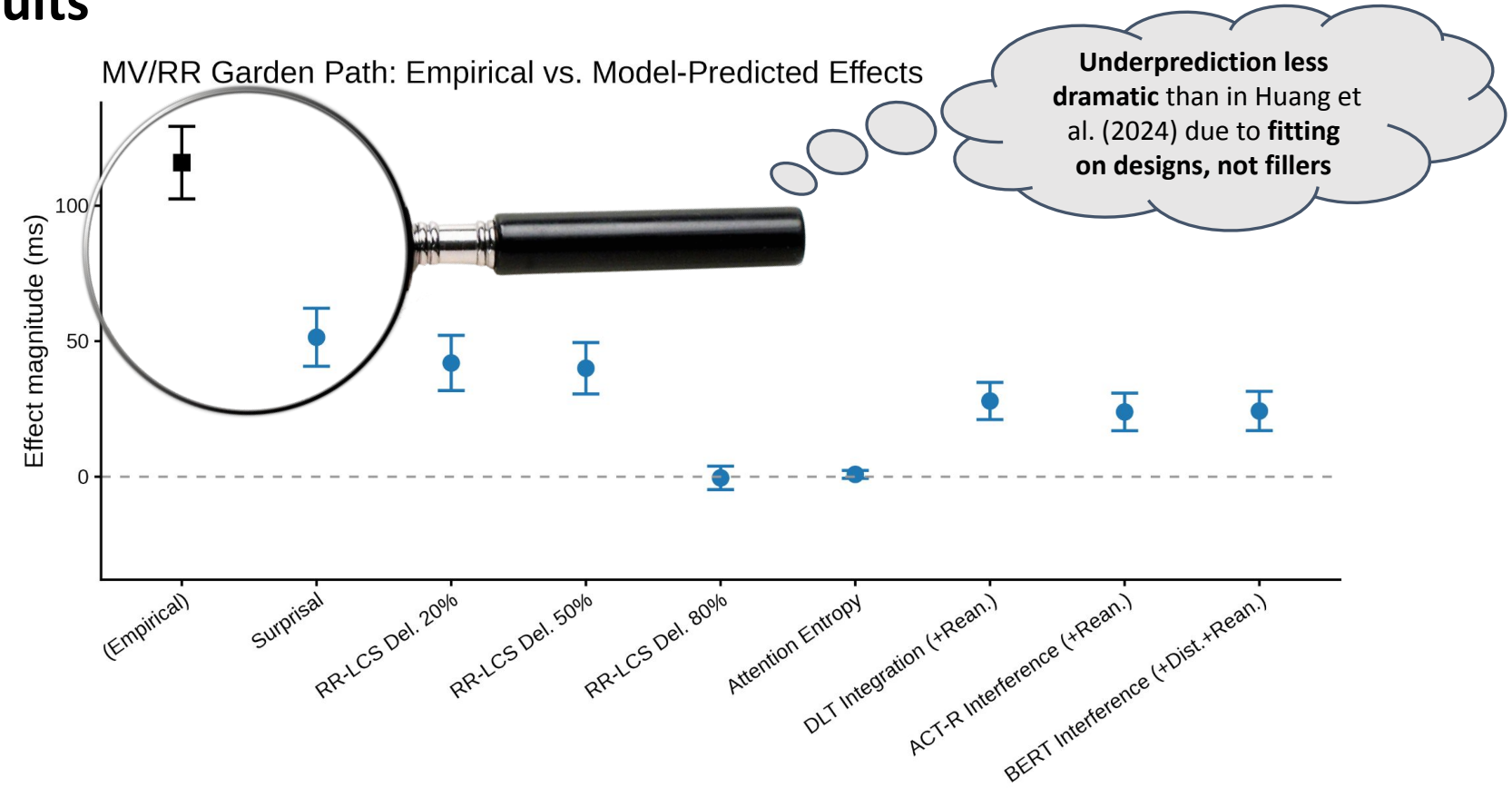
Results


Model Comparison With AIC
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)
-3000
-2000
-1000
0
ΔAIC vs. baseline (lower = better fit)

Results

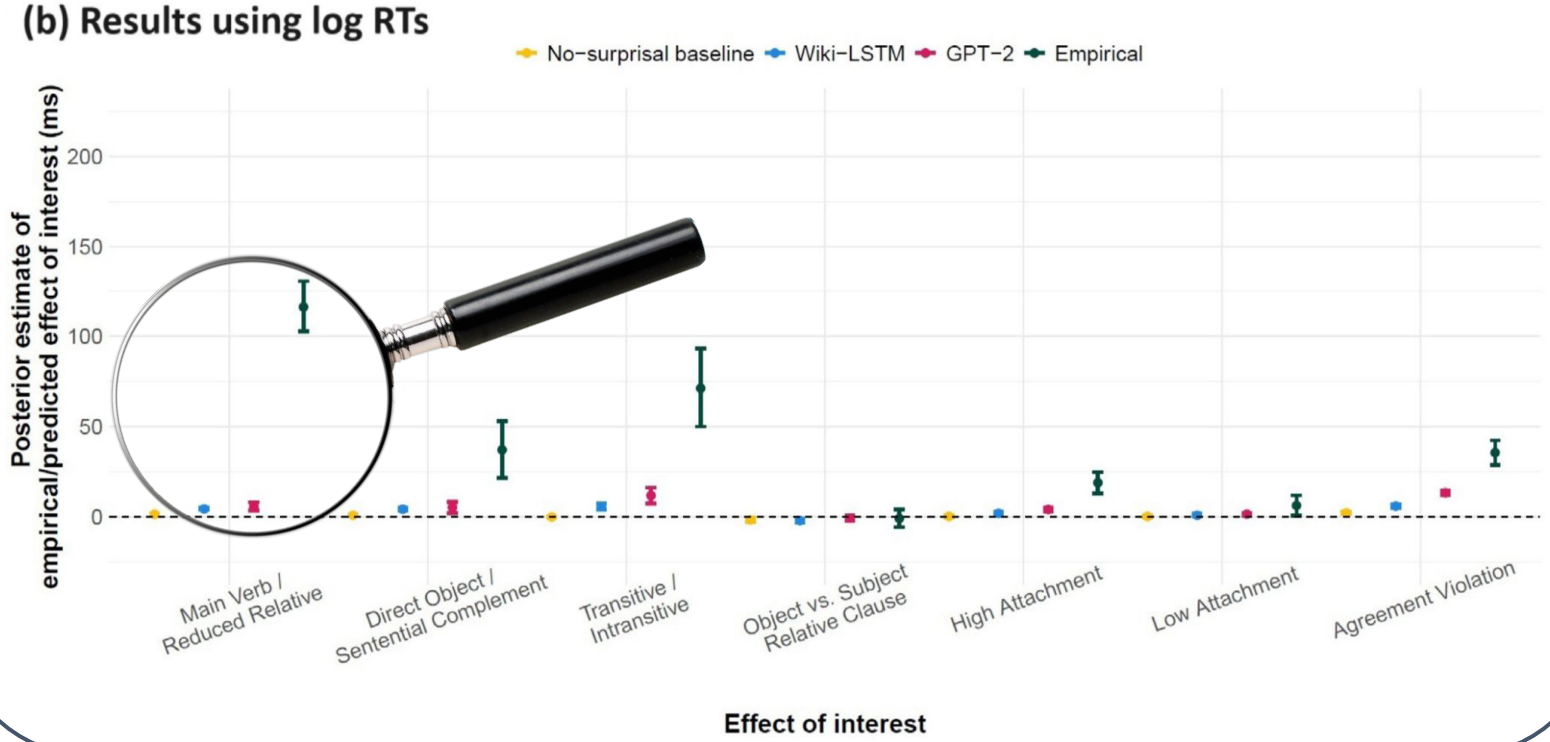

5x Repeated 10-Fold Cross-Validation
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)
-0.006
-0.004
-0.002
0.000
ΔRMSE vs. baseline ± SD (lower = better fit)

Results

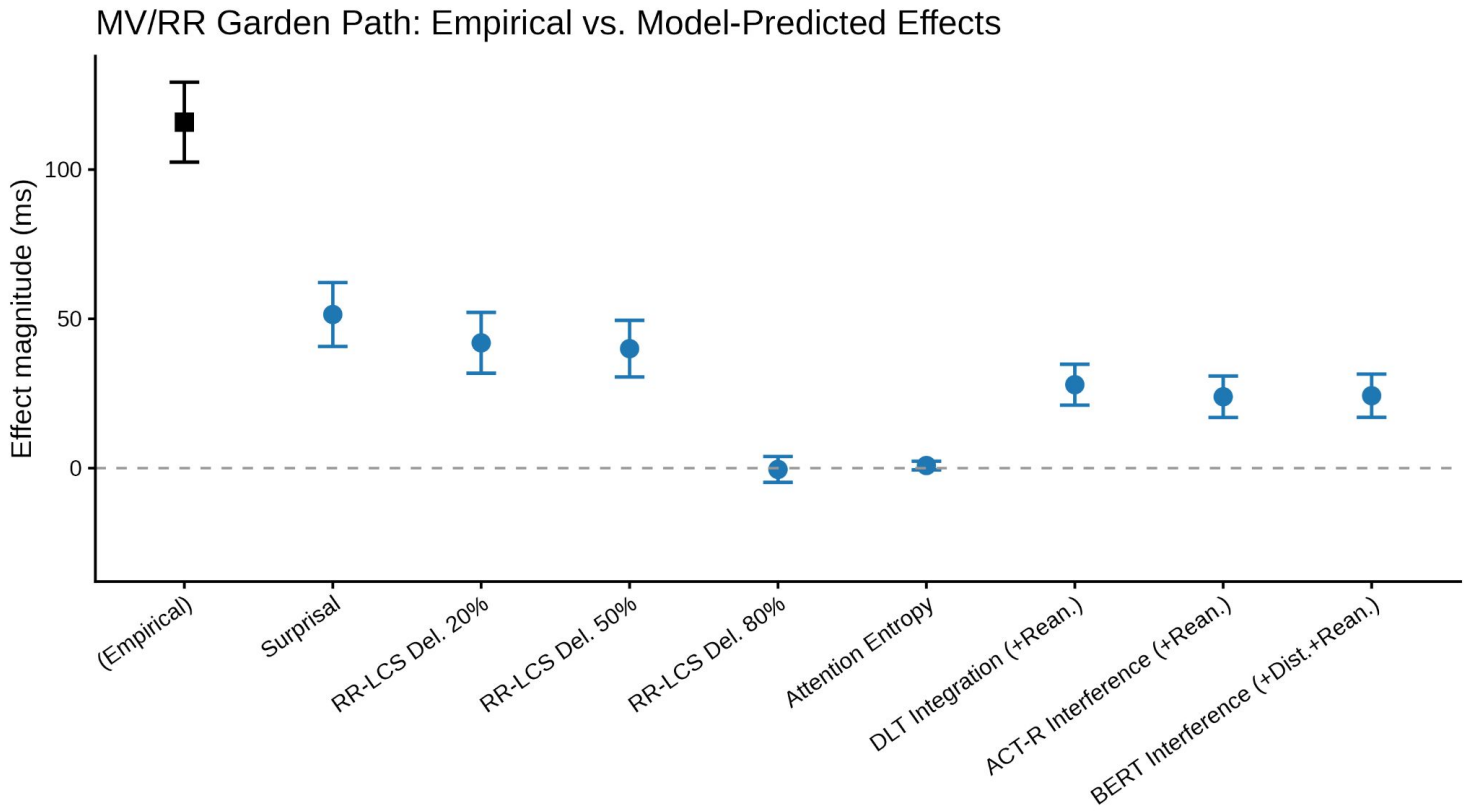

MV/RR Garden Path: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)
Underprediction less dramatic than in Huang et al. (2024) due to fitting on designs, not fillers

Results

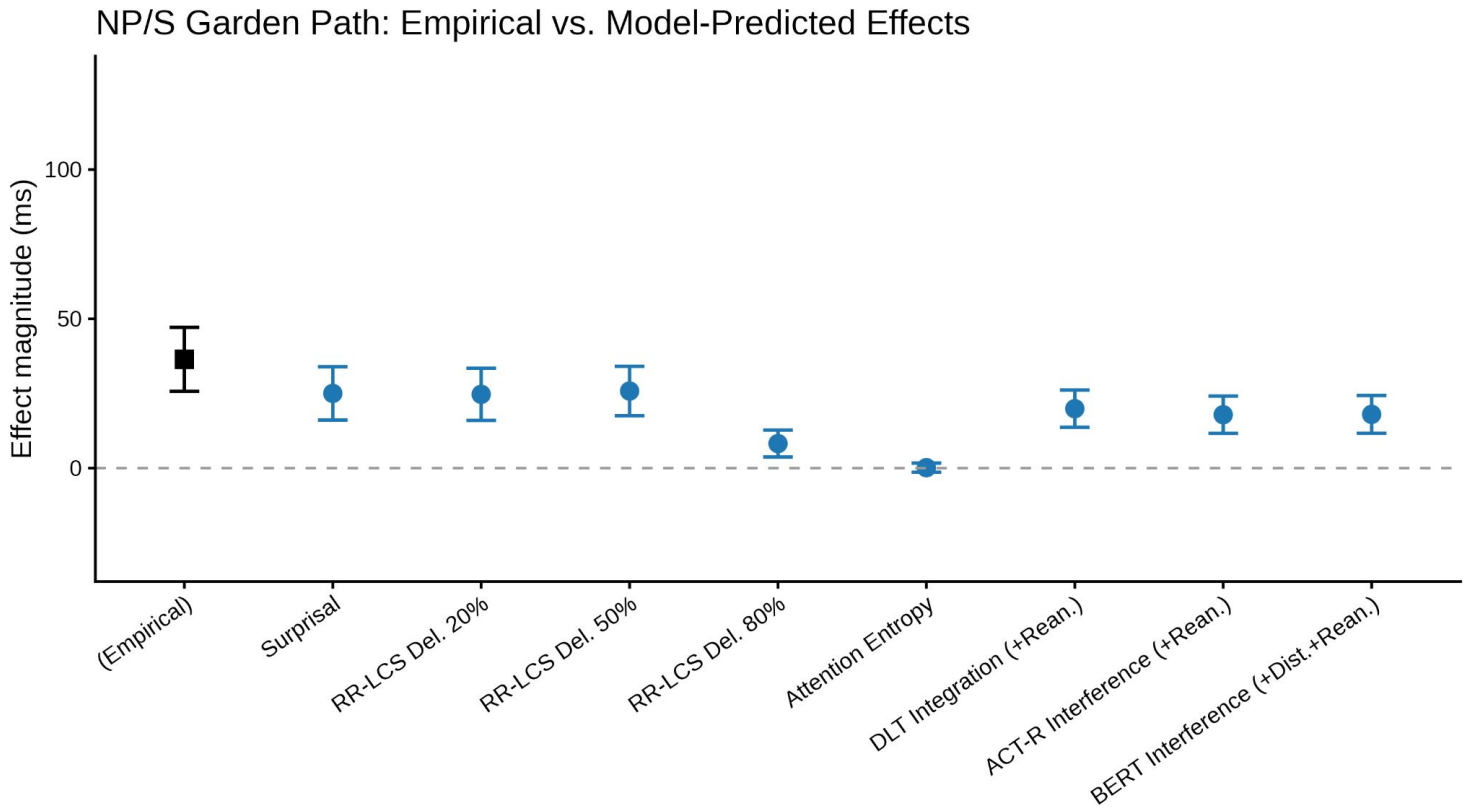
Fig. A.2, Huang et al. (2024)



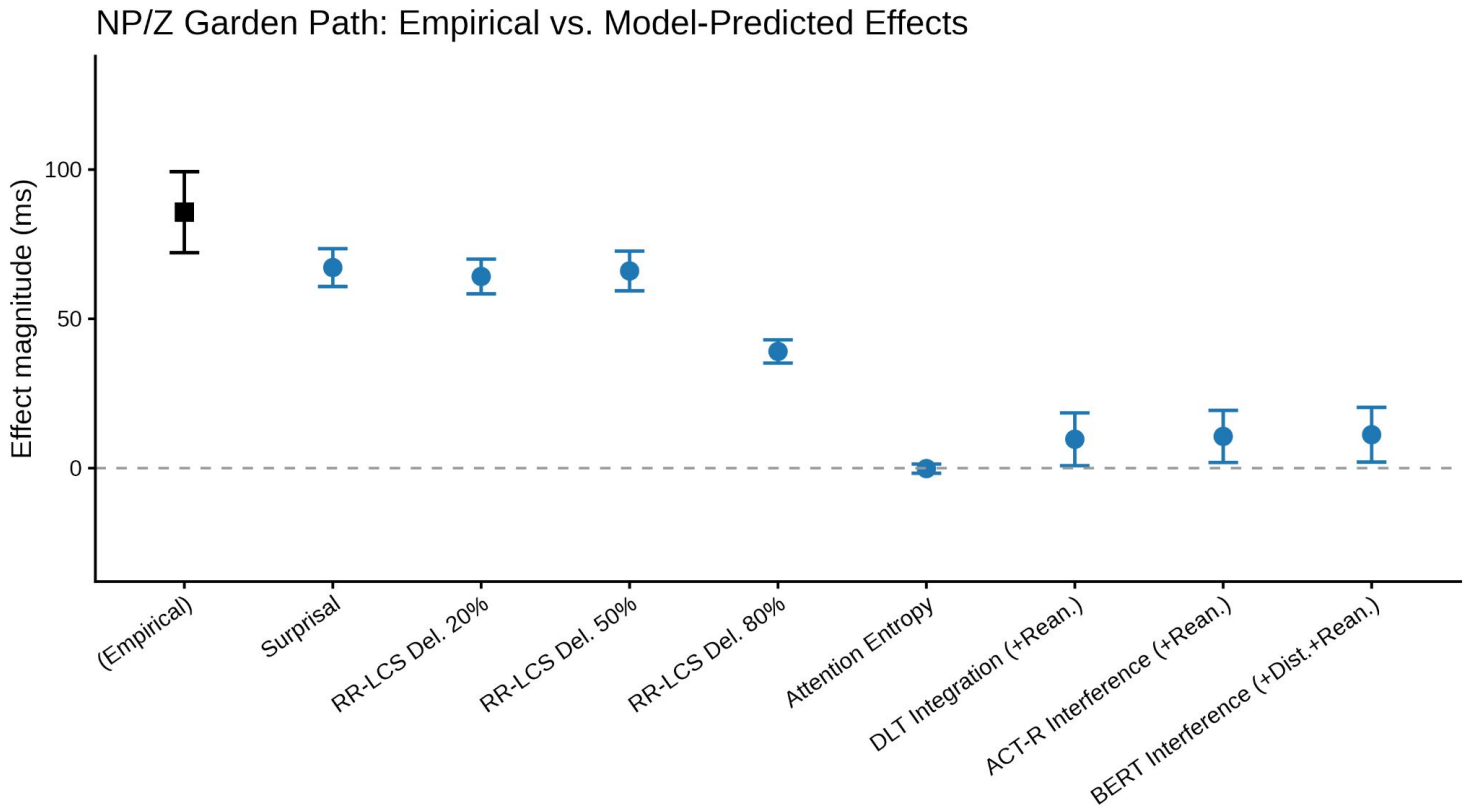
Results


MV/RR Garden Path: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

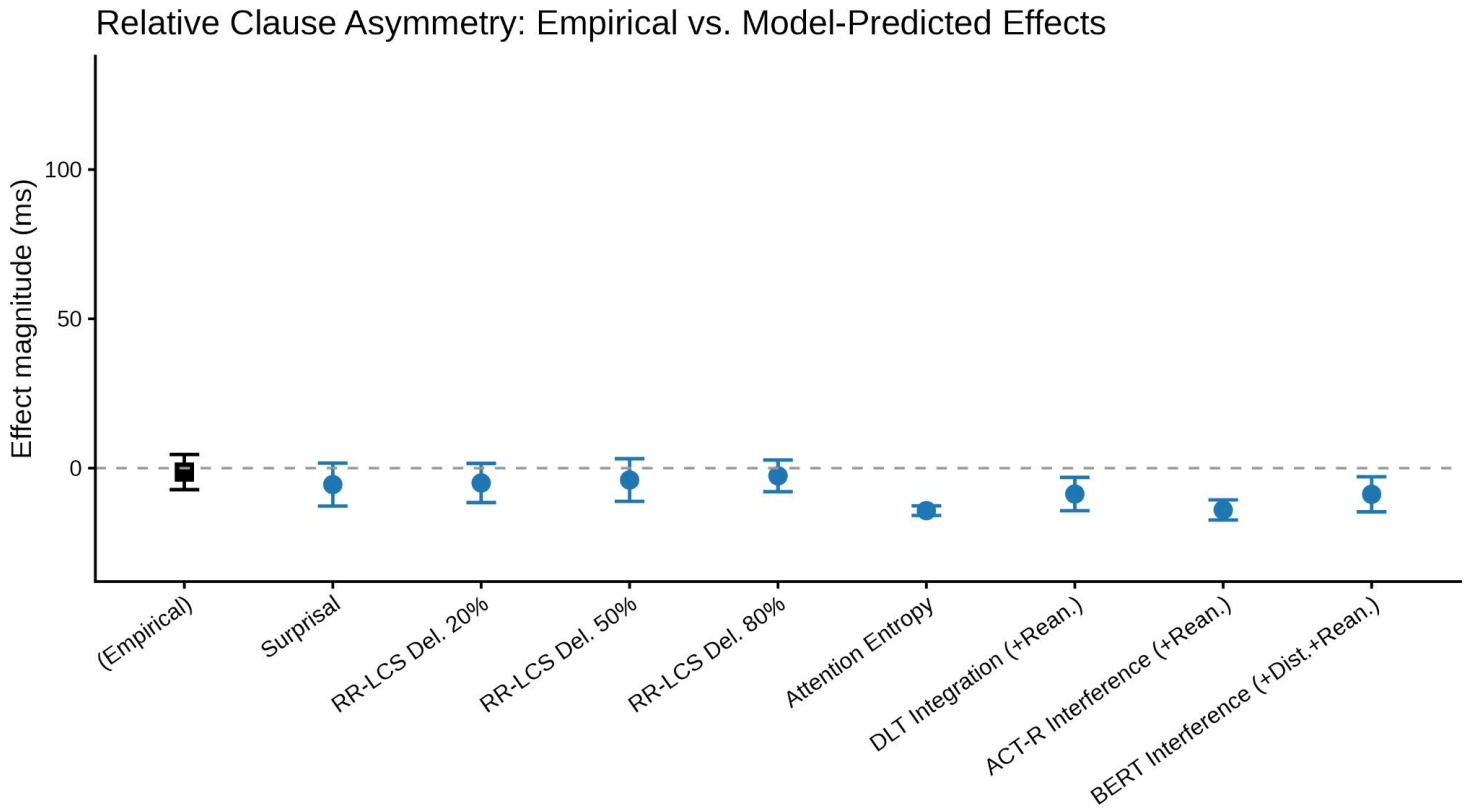
Results


NP/S Garden Path: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

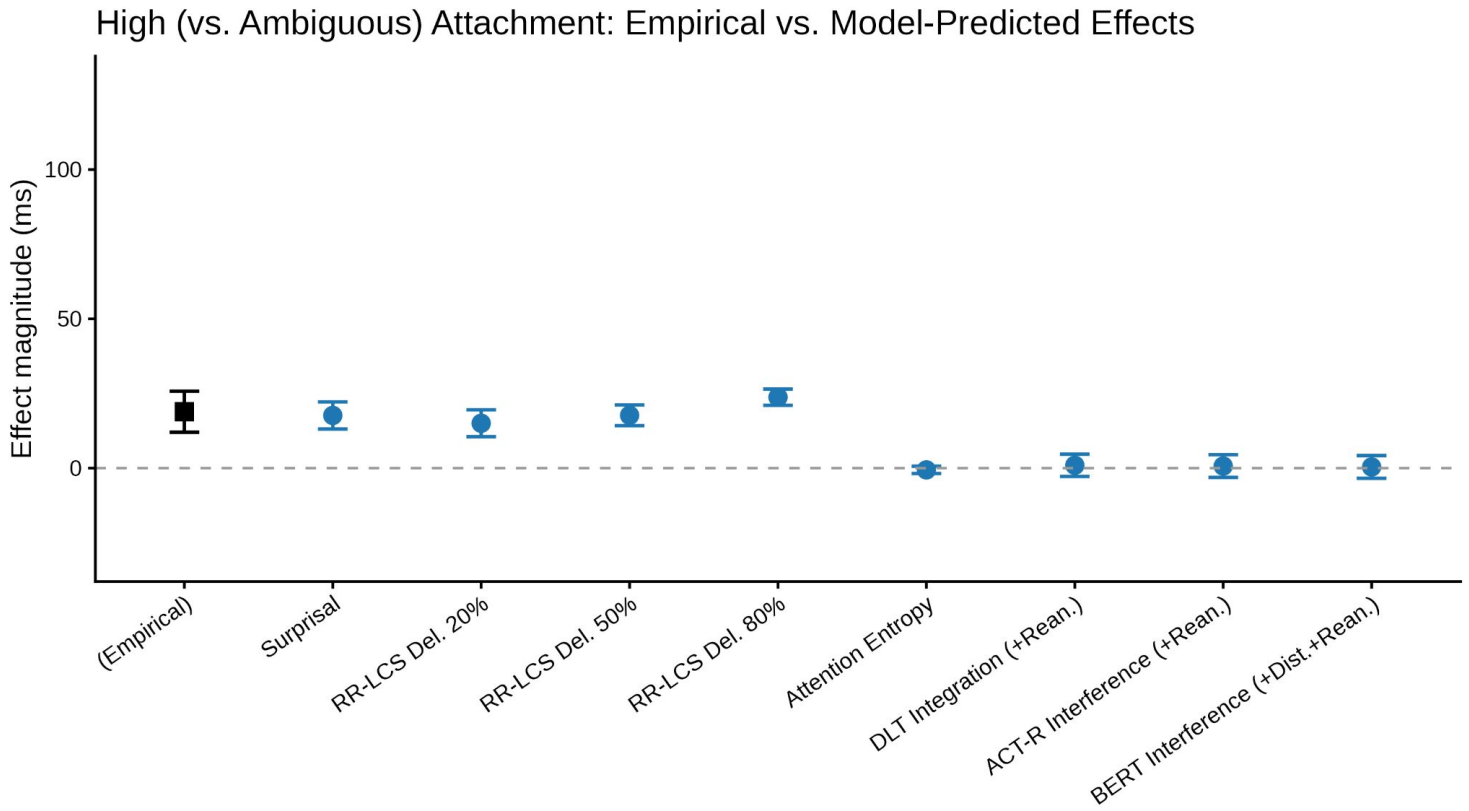
Results


NP/Z Garden Path: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

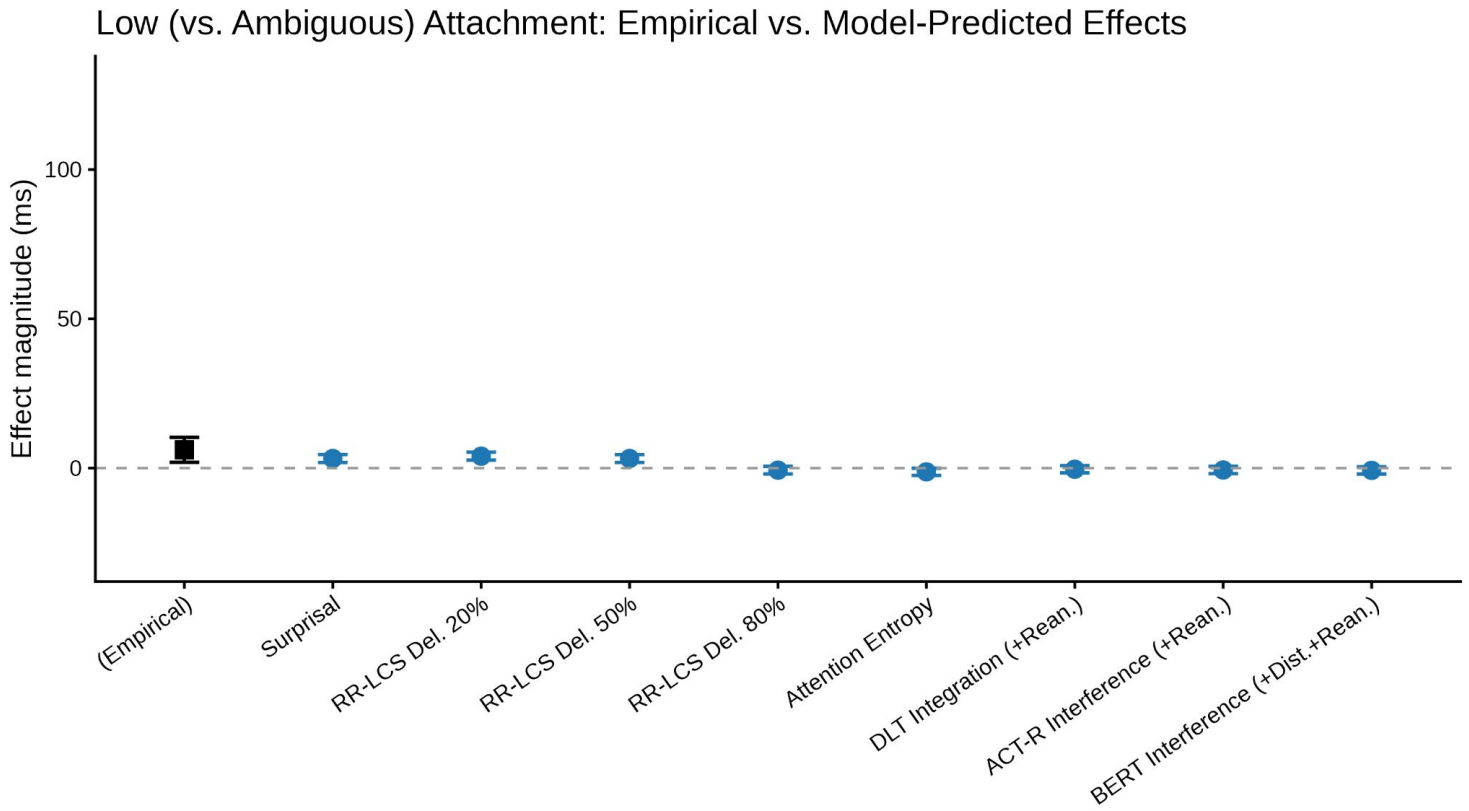
Results


Relative Clause Asymmetry: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

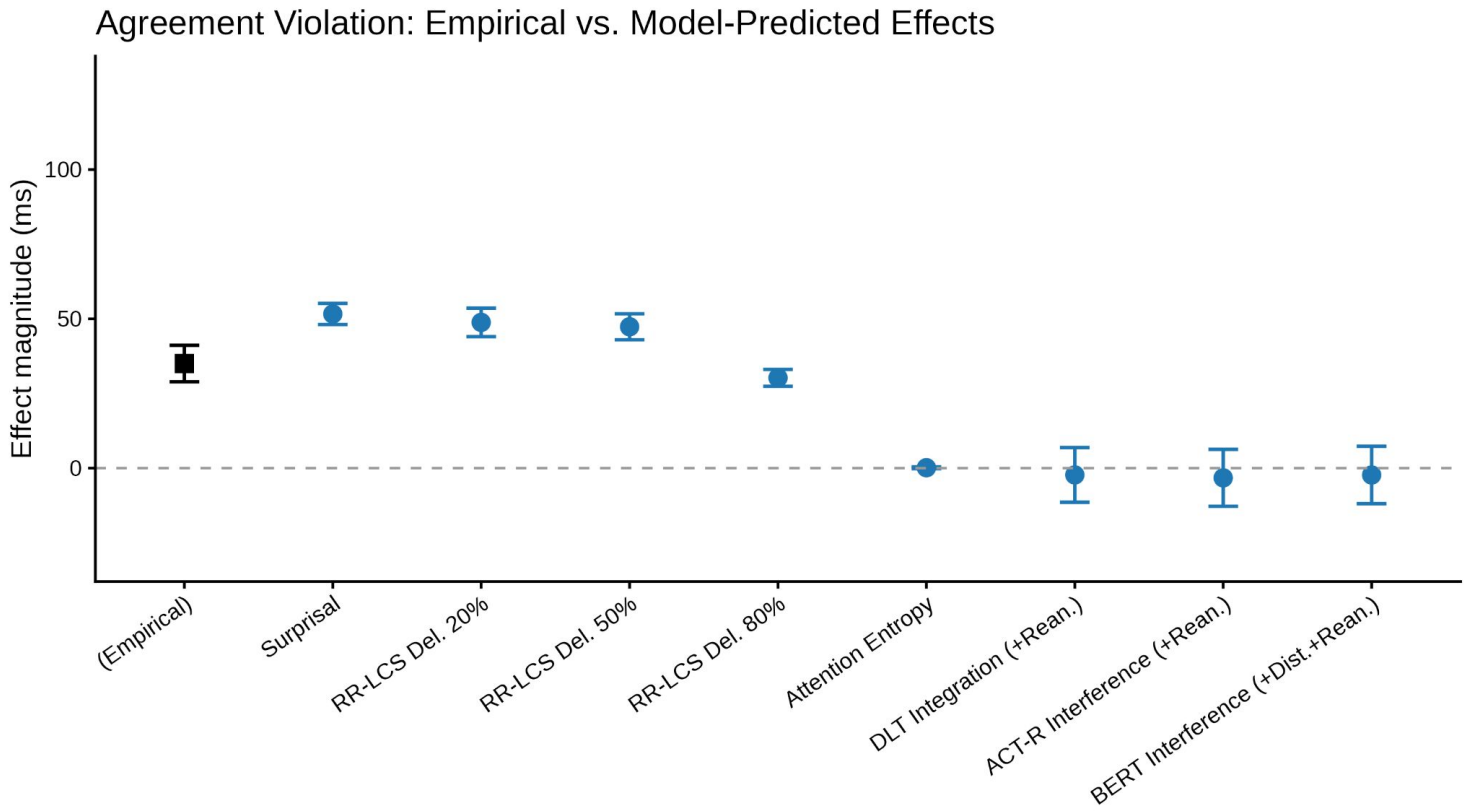
Results


High (vs. Ambiguous) Attachment: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

Results


Low (vs. Ambiguous) Attachment: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

Results

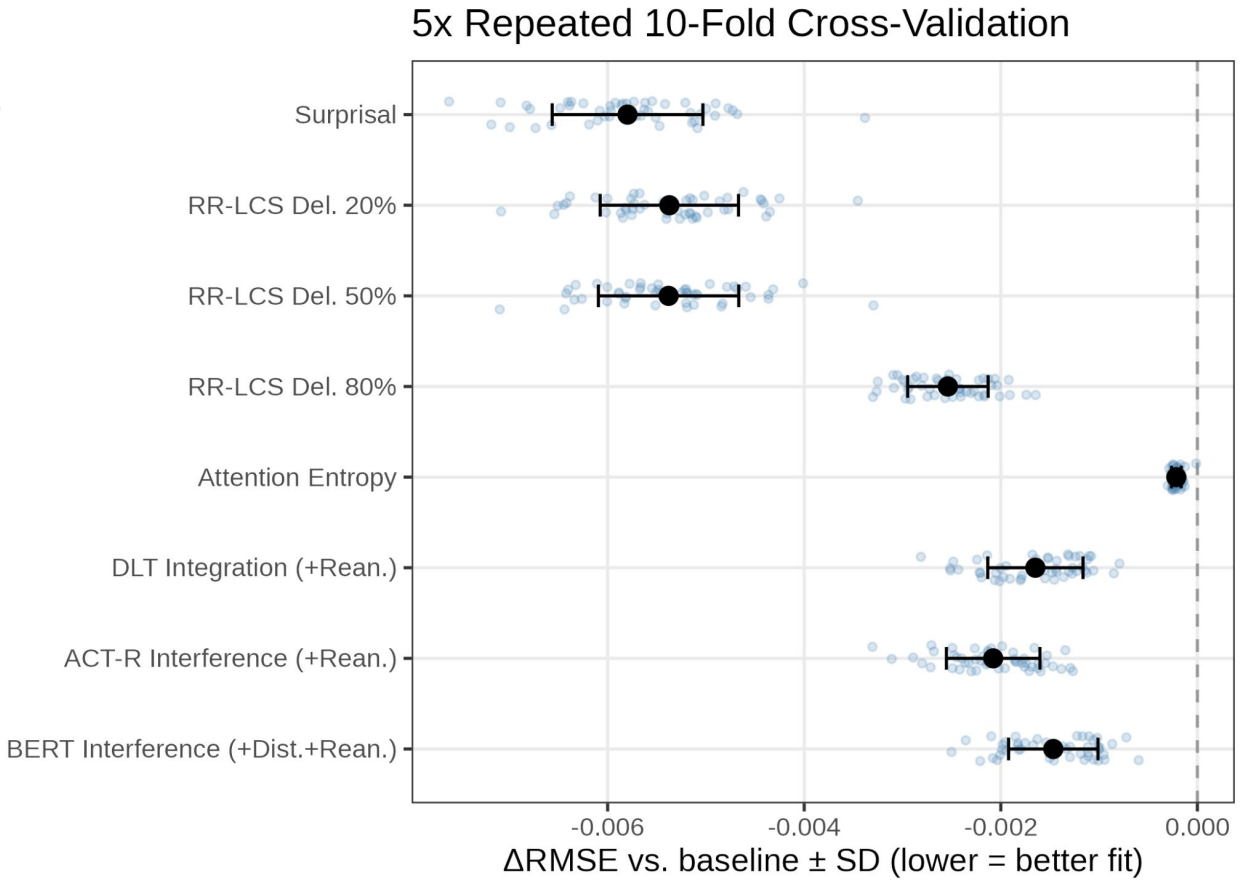

Agreement Violation: Empirical vs. Model-Predicted Effects
Effect magnitude (ms)
100
50
0
(Empirical)
Surprisal
RR-LCS Del. 20%
RR-LCS Del. 50%
RR-LCS Del. 80%
Attention Entropy
DLT Integration (+Rean.)
ACT-R Interference (+Rean.)
BERT Interference (+Dist.+Rean.)

Discussion and Conclusion

- **NLP tools** can help to **operationalize predictive metrics** of human processing effort, beyond surprisal, as **automatically computable**
- We computed predictions from **a range of metrics** motivated by **memory-based, reanalysis-based,** or hybrid **expectation–memory** accounts
- However, **no assessed metric** so far **outperformed surprisal** on the SAP dataset
- Among **models given no other information than text itself,** surprisal still reigns supreme, despite limitations in absolute terms
- **More work** is **needed,** perhaps better scalable metrics can be implemented (I call for a **“beat-surprisal-on-SAP” challenge**…)

Thanks for your attention!

- MV/RR garden paths
- NP/Z garden paths
- NP/S garden paths
- Relative clause asymmetry
- Attachment ambiguity
- Agreement violation



References

- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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)* (pp. 4171-4186).
- Gorrell, P. (1995). *Syntax and parsing* (Vol. 76). Cambridge University Press.
- Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences, 119*(43), e2122602119.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics*.
- Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language, 137*, 1026104510.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375-419.
- Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. P*roceedings of the National Academy of Sciences, 121*(10), e2307876121.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319.
- Ryu, S. H., & Lewis, R. L. (2025). Memory for prediction: A Transformer-based theory of sentence processing. *Journal of Memory and Language, 145*, 104670.
- Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language, 40*(1), 136-150.
- van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science, 45*(6), e12988.