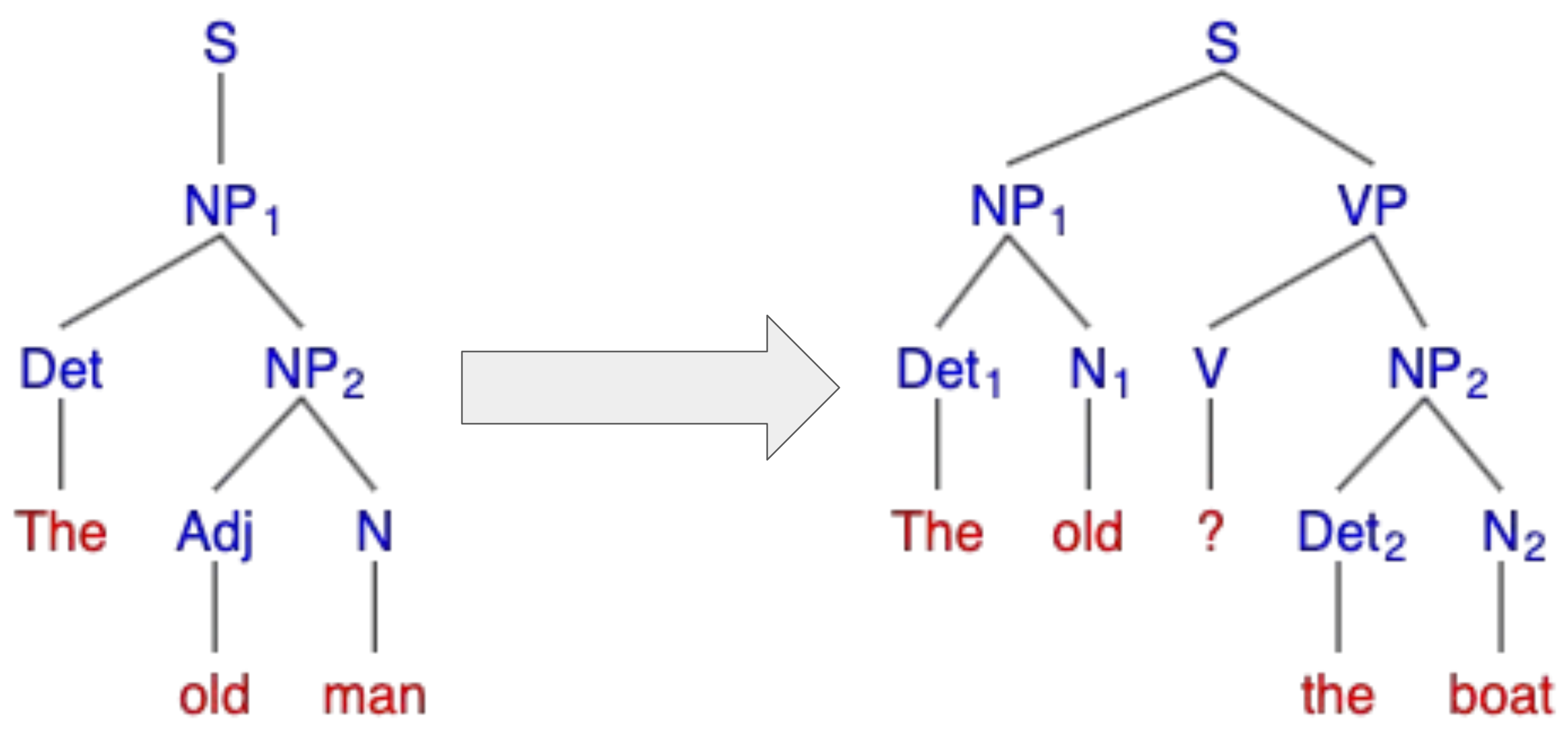
Does directionality influence a Language Model's ability to predict garden path effect?

Anzi Wang, Brian Kherlen and Sarah Cryan

Introduction

Example: The old man **the** boat



Insights from the example:

- It is common for humans to initially misanalyze the sentence
- "the": the **disambiguating word** of the garden path effect
- When **context** from both sides of the disambiguating word is provided, temporary ambiguity is lifted!

In human reading behavior, more context reduces the garden path effect and helps prevent misinterpretation (Fujita and Yoshida, 2021). Does more context (i.e., context from both sides of the disambiguating word) have a similar effect on how Language Models process sentences?

→ Is GPT-2, a causal model, better than BERT, a masked model, at predicting the garden path effects experienced by humans?

Background

Types of garden path sentences (Futrell et al., 2019; Huang et al., 2024):

- Main-verb/Reduced-relative (MV/RR)

  The woman brought the sandwich from the kitchen **fell** in the dining room
- NP/Z (Overt Object)

  As the gangster shot the woman **burst** into hysterics
- NP/Z (Verb Transitivity)

  When the dog scratched the vet **took** off the muzzle

Empirical findings

- Context plays a crucial role in reducing misanalysis
- More constraining context improves parsing accuracy (Fujita and Yoshida, 2021)

Findings on LMs

- LSTMs and neural networks demonstrate garden-path effects
- Evaluation metric: surprisal difference difference between ambiguous and unambiguous sentence pairs (Futrell et al. 2019)
- BERT performs similarly to humans in processing garden-path sentences (Irwin and Wilson, 2023)

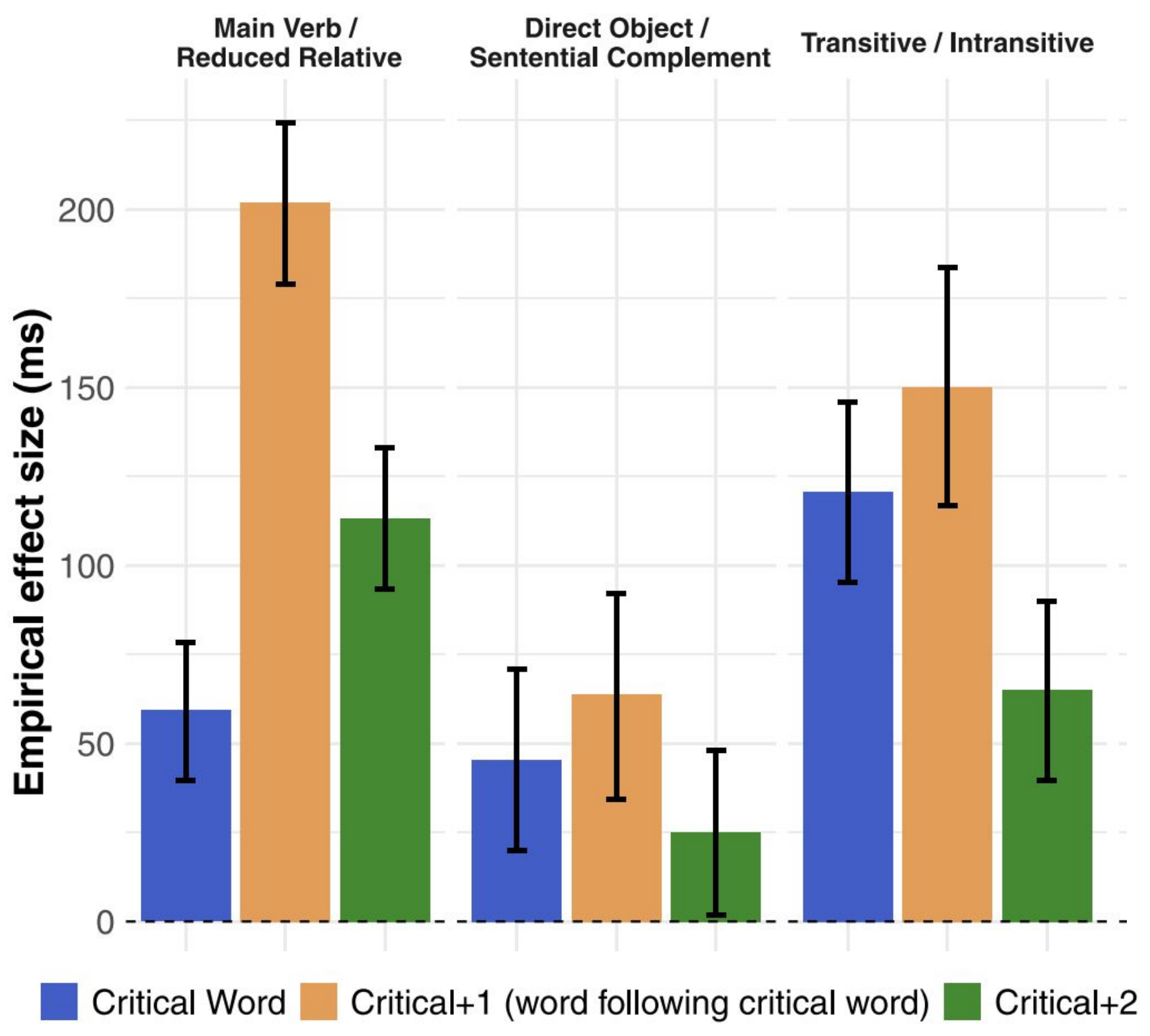


Fig 1. Empirical effect size at the disambiguating word (Huang et al., 2024)

Experimental setup

Datasets

We evaluate LMs under 3 conditions (Futrell et al., 2019):

- Main-verb/Reduced-relative (MV/RR)
- NP/Z (Overt Object)
- NP/Z (Verb Transitivity)

Each dataset: 50 garden path sentences and 50 unambiguous equivalents

- The woman brought the sandwich from the kitchen **fell** in the dining room (MV/RR)
- The woman who was brought the sandwich from the kitchen **fell** in the dining room (unambiguous equivalent)

Models

- 5 GPT-2 Small models (Bolton et al. 2021)
  - 124M parameters
  - Causal model (unidirectional)
  - Trained with same hyper-parameters, different random seeds
- 5 BERT-Base models (Sellam et al. 2022)
  - 110M parameters
  - Masked model (bidirectional)
  - Trained with same hyper-parameters, different random seeds

Evaluation metrics

Compare surprisal difference in ambig/unambig pairs across BERT/GPT-2

Results

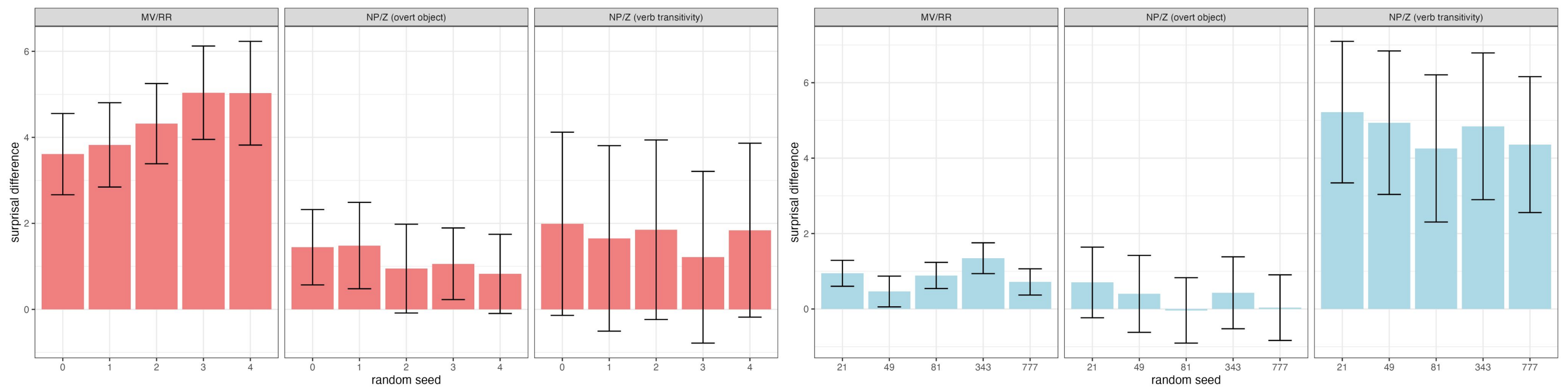


Fig 2. ROI surprisal difference across different conditions predicted by BERT (pink) and GPT-2 (blue) with different seeds

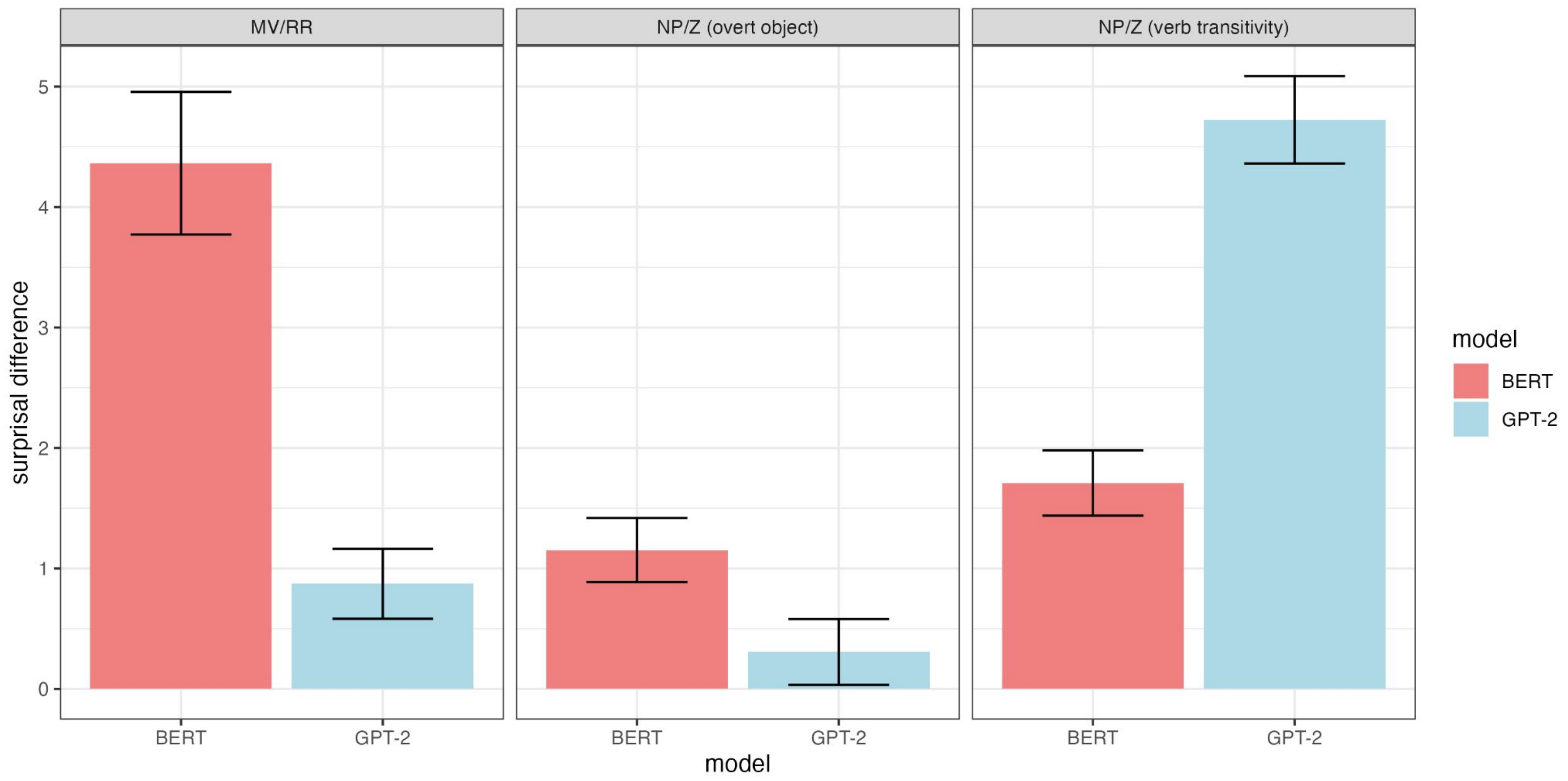


Fig 3. Overall ROI surprisal difference across different conditions predicted by BERT (pink) and GPT-2 (blue)

Summary

Comparing GPT-2 and BERT

- Both GPT-2 and BERT are able to predict garden path effects. Both models find garden path sentences more surprising than their unambiguous equivalents
  - (ambig - unambig) is almost always positive across different models/seeds
- However, there is a qualitative difference between how GPT-2 and BERT deal with processing difficulty
  - In MV/RR condition, BERT assigns a significantly higher surprisal to the disambiguating word than GPT-2 does
  - In verb transitivity condition, we observe the opposite

Comparing LMs and human behavior

- Comparing to empirical effect size at the disambiguating word (Huang et al., 2024), GPT-2 demonstrates a more human-like pattern in predicting processing difficulty
- GPT-2's performance aligns more closely with human reading behavior, so GPT-2 makes a better model of human behavior than BERT in processing garden path sentences
- This conclusion is in line with our initial hypothesis about context and directionality

Limitations + Future directions

- Only measured effect size of the disambiguating word

  → Wider region of interest: disambiguating word + two following words
- Large error bars and variation across different seeds

  Insignificant result in overt object condition for GPT 2

  → Scale up and include more random seeds for more general conclusions about each model

  → Need more garden path sentences! 50 pairs for each condition are unlikely to be enough for the argument we want to make
- Insufficient evidence to draw a general conclusion about model directionality (1 unidirectional and 1 bidirectional model)

  → Scale up with more models of each type in future experiments

References

- Hiroki Fujita, and Masaya Yoshida. (2024) Online reflexive resolution and interference. *Language, Cognition and Neuroscience*.
- 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 ACL*.
- Kuang-jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon and Tal Linzen. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*.
- *jsSyntaxTree*. (n.d.). Ironcreek.net. https://ironcreek.net/syntaxtree/.
- Lingling Zhou, Suzan Verberne, and Gijs Wijnholds. (2024) Tree Transformer's Disambiguation Ability of Prepositional Phrase Attachment and Garden Path Effects. *In Proceedings of ACL*.
- Tovah Irwin, Kyra Wilson, and Alec Marantz. (2023). BERT Shows Garden Path Effects. *EACL Proceedings*.