EMNLP2026 BUDAPEST

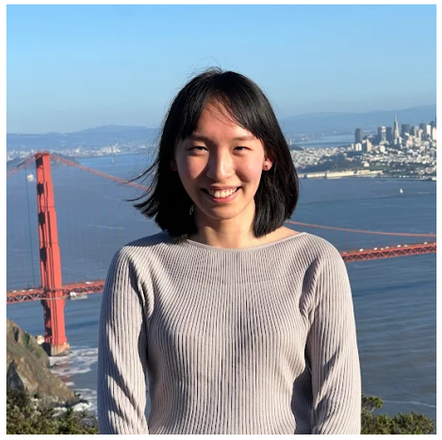
GEORGETOWN UNIVERSITY

Mohamed bin Zayed University of Artificial Intelligence

TOHOKU UNIVERSITY

Using LMs to Model the Effects of Context and Coreference during Sentence Comprehension









Kohei Kajikawa[1] **Lin Ai**[1] **Tatsuki Kuribayashi**[2,3] **Ethan Wilcox**[1]

[1]Georgetown University [2]MBZUAI [3]Tohoku University

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing
  - *Surprisal* predicts reading times [e.g., Wilcox et al., 2020; Oh et al., 2022]

$$-\log p(w \mid \text{context})$$

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing
  - *Surprisal* predicts reading times [e.g., Wilcox et al., 2020; Oh et al., 2022]

    $-\log p(w \mid \text{context})$

  - LM representations align with brain activity [e.g., Schrimpf et al., 2021; Goldstein et al., 2022]

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing
  - *Surprisal* predicts reading times [e.g., Wilcox et al., 2020; Oh et al., 2022]

    $-\log p(w \mid \text{context})$

  - LM representations align with brain activity [e.g., Schrimpf et al., 2021; Goldstein et al., 2022]
- But better LMs are not always better models of human processing [Kuribayashi et al., 2021; Oh & Schuler, 2023]

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing
  - *Surprisal* predicts reading times [e.g., Wilcox et al., 2020; Oh et al., 2022]

    $-\log p(w \mid \text{context})$
  - LM representations align with brain activity [e.g., Schrimpf et al., 2021; Goldstein et al., 2022]
- But better LMs are not always better models of human processing [Kuribayashi et al., 2021; Oh & Schuler, 2023]
- One explanation: modern LMs are "***superhuman***" [Oh & Linzen, 2026]
  - They have "**near-perfect short-term memory**"

Can a Transformer be a Model of Human Reading?

- Transformer LMs predict human processing
  - *Surprisal* predicts reading times [e.g., Wilcox et al., 2020; Oh et al., 2022]

    $-\log p(w \mid \text{context})$
  - LM representations align with brain activity [e.g., Schrimpf et al., 2021; Goldstein et al., 2022]
- But better LMs are not always better models of human processing [Kuribayashi et al., 2021; Oh & Schuler, 2023]
- One explanation: modern LMs are "***superhuman***" [Oh & Linzen, 2026]
  - They have "**near-perfect short-term memory**"
- What makes a Transformer more "human-like"?

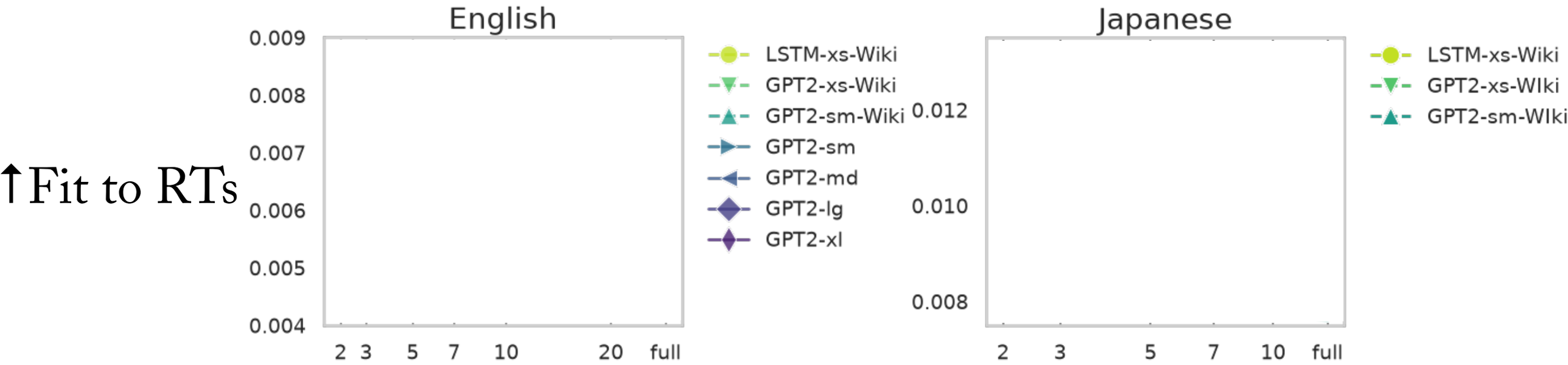
Limiting Context Makes LMs More "Human-like"

- Restricting models' access to prior context simulates human working memory limits [e.g., Hahn et al., 2022; Kuribayashi et al., 2022; Timkey and Linzen, 2023; De Varda and Marelli, 2024; Clark et al., 2025; Xu et al., 2026]

Limiting Context Makes LMs More "Human-like"

- Restricting models' access to prior context simulates human working memory limits [e.g., Hahn et al., 2022; Kuribayashi et al., 2022; Timkey and Linzen, 2023; De Varda and Marelli, 2024; Clark et al., 2025; Xu et al., 2026]

- Kuribayashi et al. [2022; EMNLP]:



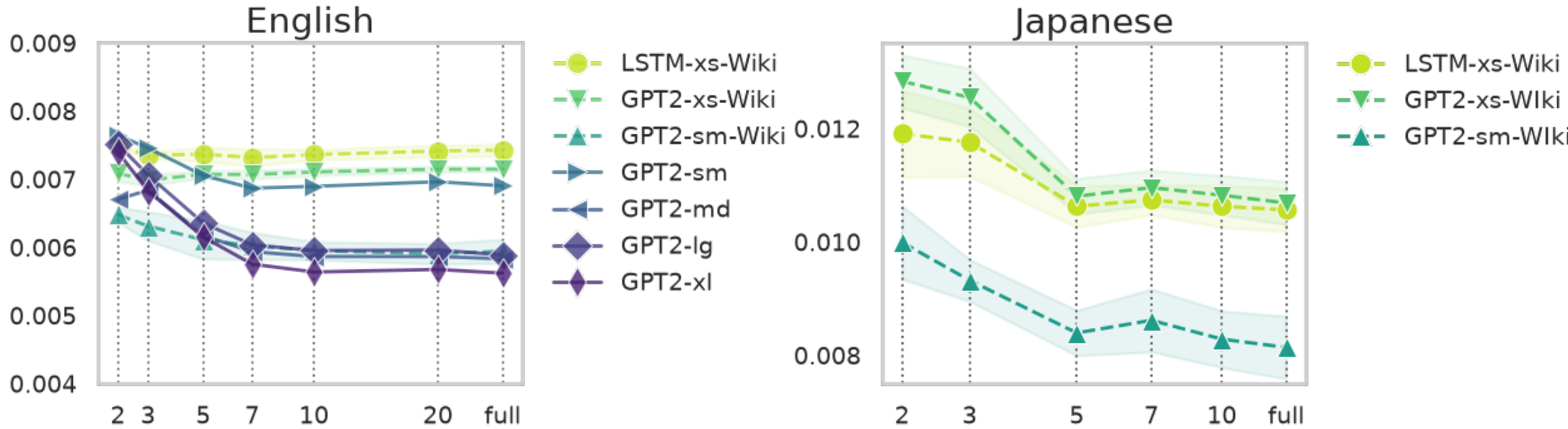
(Adapted from [Kuribayashi et al., 2022])

Limiting Context Makes LMs More "Human-like"

- Restricting models' access to prior context simulates human working memory limits [e.g., Hahn et al., 2022; Kuribayashi et al., 2022; Timkey and Linzen, 2023; De Varda and Marelli, 2024; Clark et al., 2025; Xu et al., 2026]

- Kuribayashi et al. [2022; EMNLP]:

↑Fit to RTs



←Shorter Context ←Shorter Context

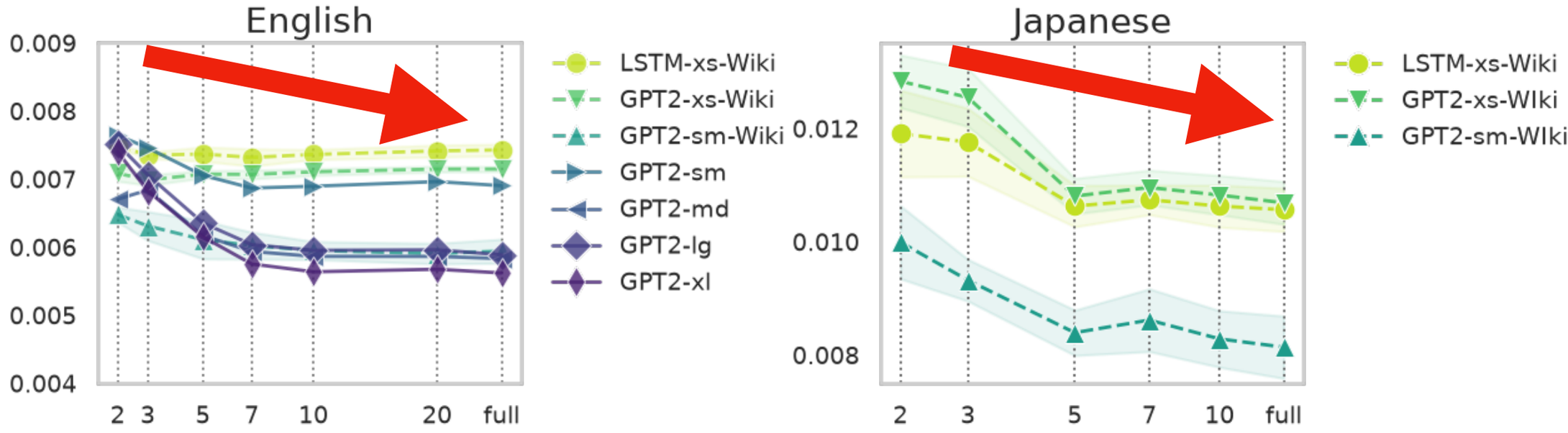
(Adapted from [Kuribayashi et al., 2022])

Limiting Context Makes LMs More "Human-like"

- Restricting models' access to prior context simulates human working memory limits [e.g., Hahn et al., 2022; Kuribayashi et al., 2022; Timkey and Linzen, 2023; De Varda and Marelli, 2024; Clark et al., 2025; Xu et al., 2026]

- Kuribayashi et al. [2022; EMNLP]:

Shorter is better



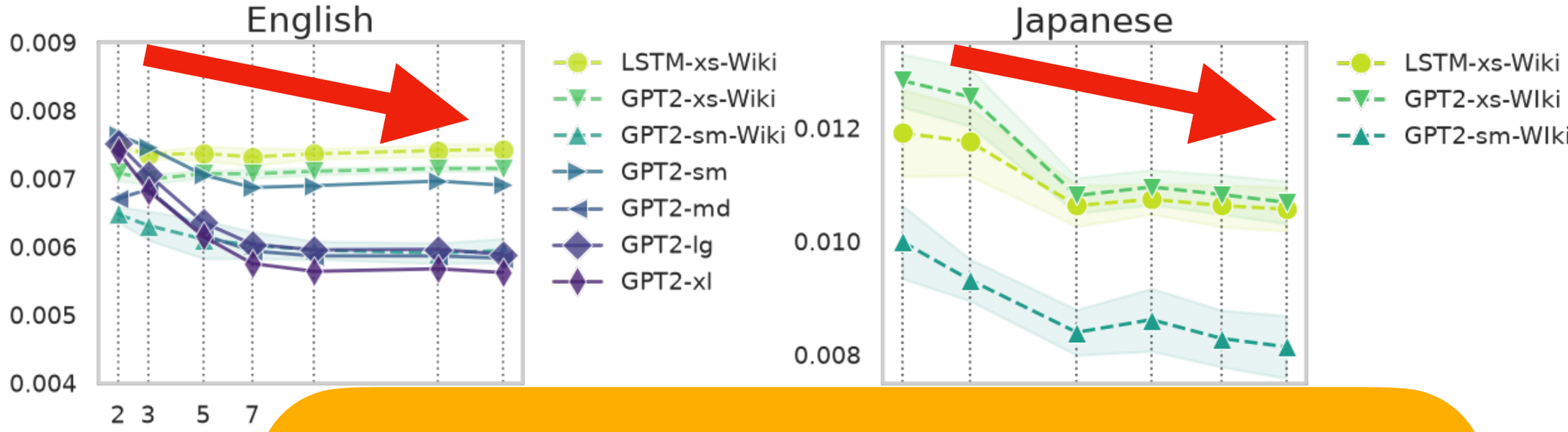
(Adapted from [Kuribayashi et al., 2022])

Limiting Context Makes LMs More "Human-like"

- Restricting models' access to prior context simulates human working memory limits [e.g., Hahn et al., 2022; Kuribayashi et al., 2022; Timkey and Linzen, 2023; De Varda and Marelli, 2024; Clark et al., 2025; Xu et al., 2026]

- Kuribayashi et al. [2022; EMNLP]:

Shorter is better



🤔 But humans are capable of grasping **long-distance discourse structures**?

But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]

But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]

The poor boy ____

But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]

The poor boy ____

- *fell*
- *started to cry*
- *caught a cold*
- *had no money*

But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]

Tom pushed John. The poor boy ____

- *fell*
- *started to cry*

- *caught a cold*
- *had no money*



But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]

Tom pushed John. The poor boy ____

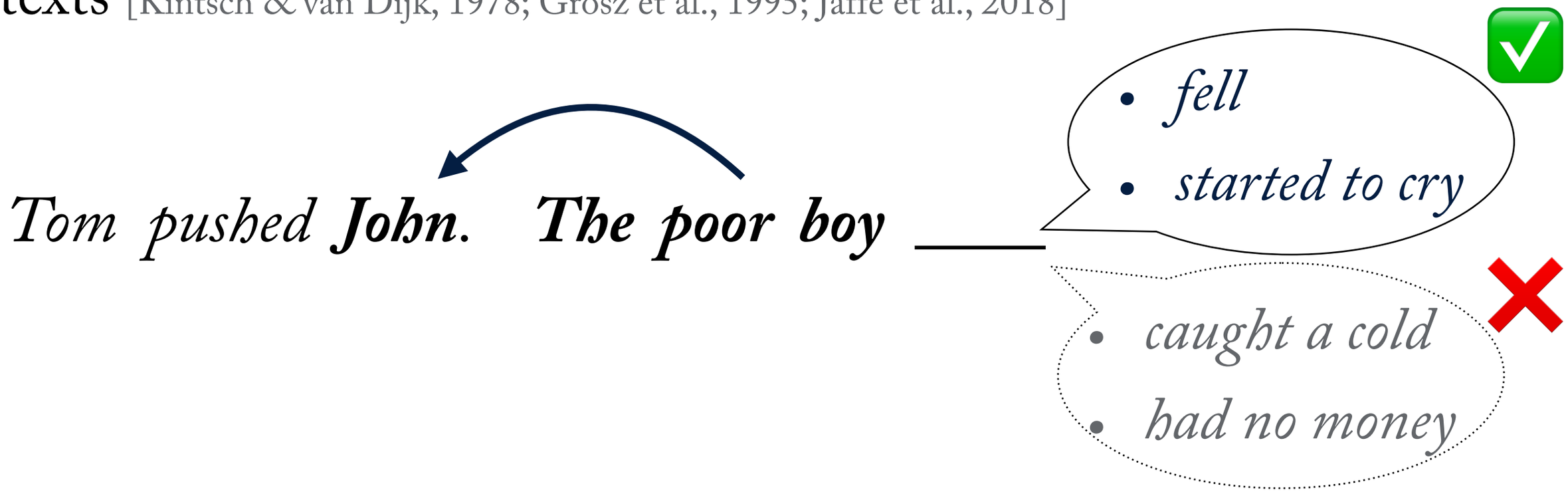


- *fell*
- *started to cry*

- *caught a cold*
- *had no money*

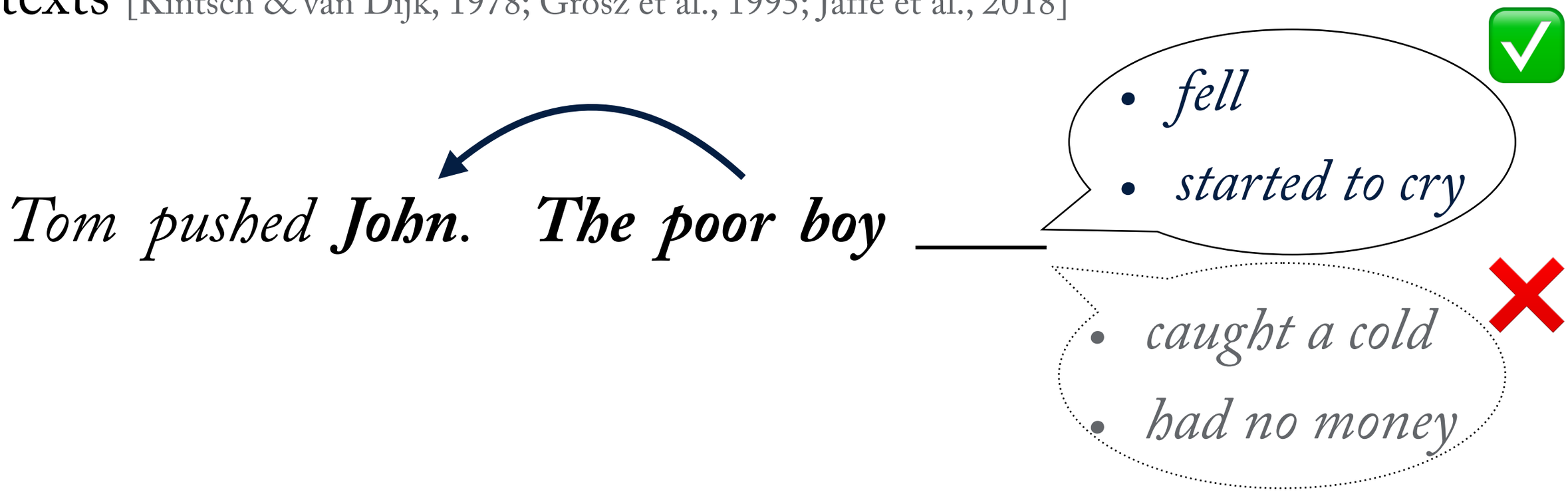
But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]



But Human Readers Use Long-range Context

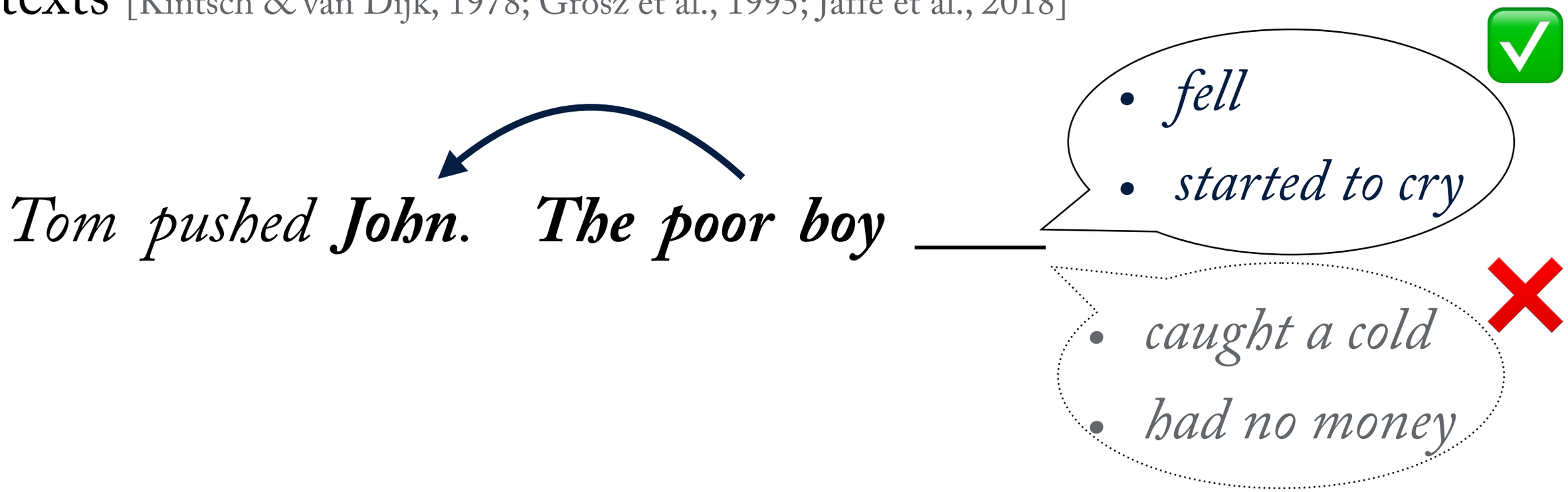
- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]



Q1: Is a shorter context always better?

But Human Readers Use Long-range Context

- Readers build and maintain discourse representations over long texts [Kintsch & van Dijk, 1978; Grosz et al., 1995; Jaffe et al., 2018]



Q1: Is a shorter context always better?

Q2: If a longer context helps, why does it help?

Roadmap

Question	How we answer it
Q1: Is a shorter context always better?	
Q2: If a longer context helps, what in it helps?	

Roadmap

Question	How we answer it
Q1: Is a shorter context always better?	Exp. 1: vary the context from 1 to 1,023 tokens
Q2: If a longer context helps, what in it helps?	

☕ Roadmap

Question	How we answer it
Q1: Is a shorter context always better?	Exp. 1: vary the context from 1 to 1,023 tokens
Q2: If a longer context helps, what in it helps?	Exp. 2: obscure entity relations in the context

Exp. 1: Vary Context from 1 to 1,023 Tokens

- LMs : GPT-2 small / medium / large / XL
- Context : only the last n tokens, $n = 1, \ldots, 1{,}023$ (28 sizes)

Exp. 1: Vary Context from 1 to 1,023 Tokens

- LMs : GPT-2 small / medium / large / XL
- Context : only the last n tokens, $n = 1, \ldots, 1{,}023$ (28 sizes)
- Evaluation : *Delta log-likelihood*

Exp. 1: Vary Context from 1 to 1,023 Tokens

- LMs : GPT-2 small / medium / large / XL
- Context : only the last n tokens, $n = 1, \ldots, 1{,}023$ (28 sizes)
- Evaluation : *Delta log-likelihood*
  gain in held-out log-likelihood for predicting RTs
  from adding **surprisal** to a baseline regression
  (a baseline model includes word length, unigram surprisal, and their spillover terms, and word positions)

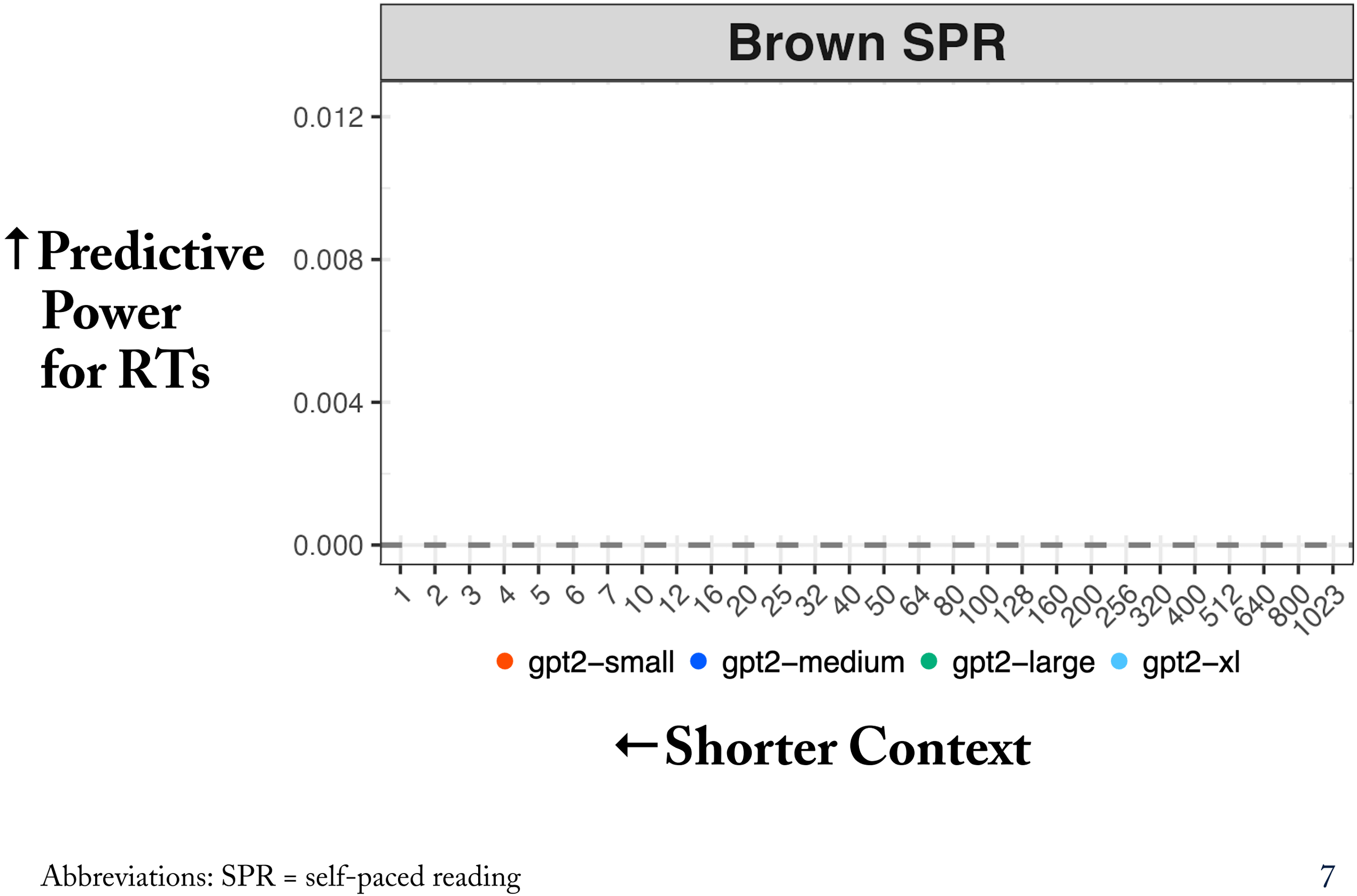
Exp. 1: Vary Context from 1 to 1,023 Tokens

- LMs : GPT-2 small / medium / large / XL
- Context : only the last n tokens, $n = 1, \ldots, 1{,}023$ (28 sizes)
- Evaluation : *Delta log-likelihood*
  gain in held-out log-likelihood for predicting RTs
  from adding **surprisal** to a baseline regression
  (a baseline model includes word length, unigram surprisal, and their spillover terms, and word positions)

Dataset	Measure	Tokens per text (mean)
Brown [Smith & Levy, 2013]	Self-paced reading	707
Natural Stories [Futrell et al., 2021]	Self-paced reading A-Maze	1,237
OneStop [Berzak et al., 2025]	Eye-Tracking*	751
Provo [Luke & Christianson, 2018]	Eye-Tracking*	60

*For eye-tracking data, we use (1) First Fixation (FF), (2) Gaze Duration (GD), and (3) Total Fixation time (TF)

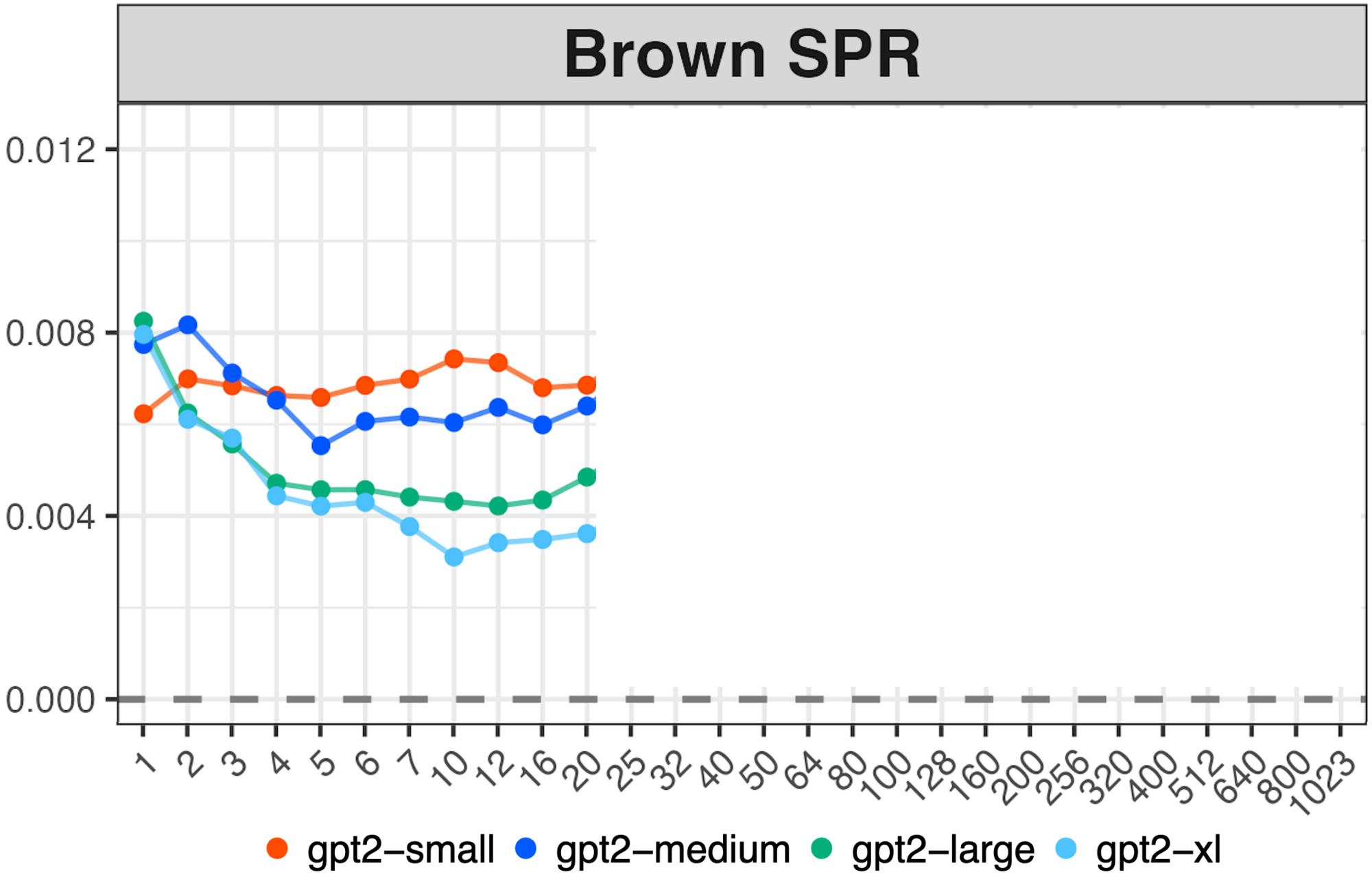
Exp. 1: Vary Context from 1 to 1,023 Tokens



Abbreviations: SPR = self-paced reading

Exp. 1: Vary Context from 1 to 1,023 Tokens

↑ Predictive Power for RTs

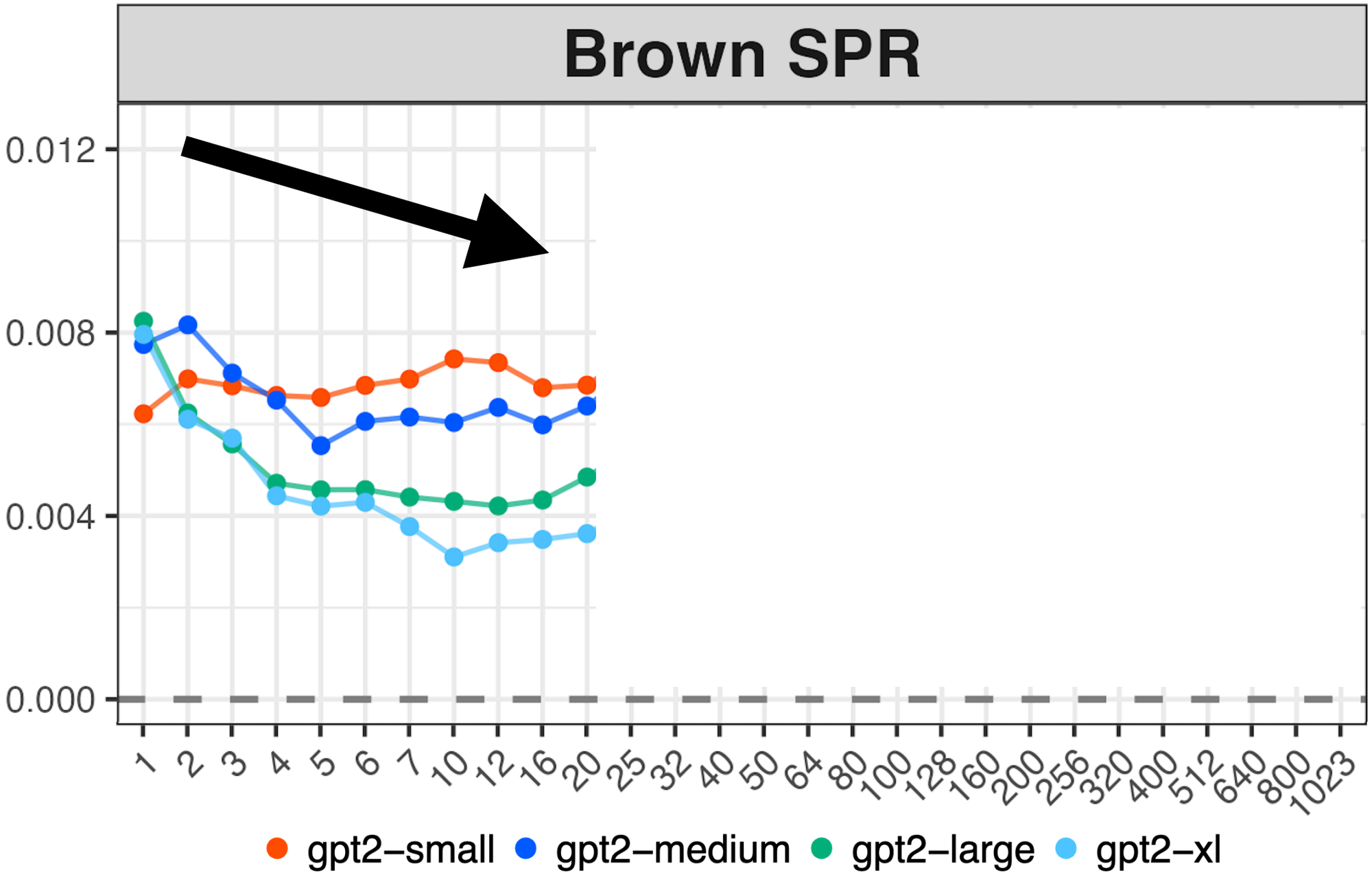


← Shorter Context

Abbreviations: SPR = self-paced reading

Exp. 1: Vary Context from 1 to 1,023 Tokens

↑ Predictive Power for RTs

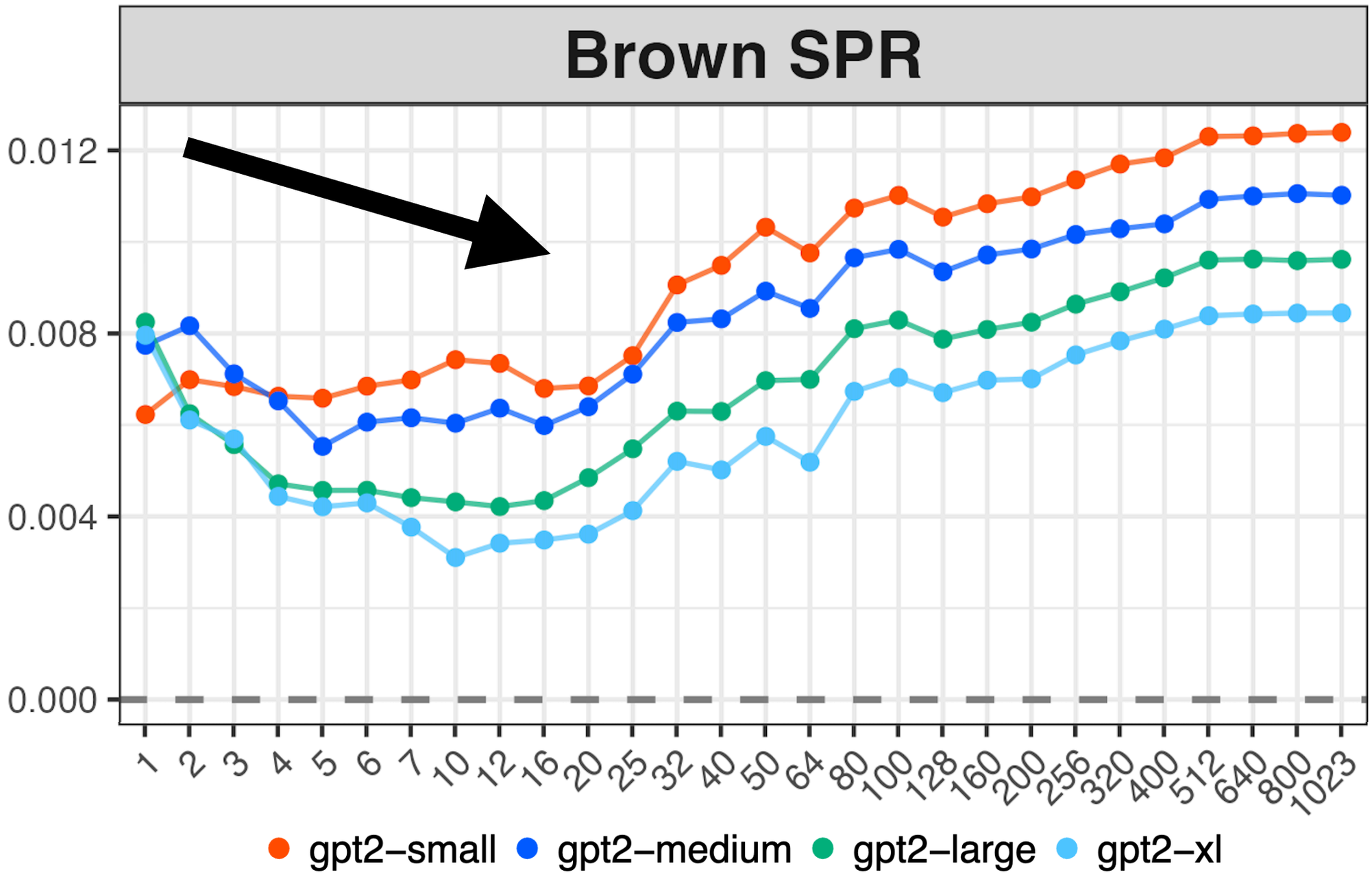


← Shorter Context

Abbreviations: SPR = self-paced reading

Exp. 1: Vary Context from 1 to 1,023 Tokens

↑ Predictive Power for RTs

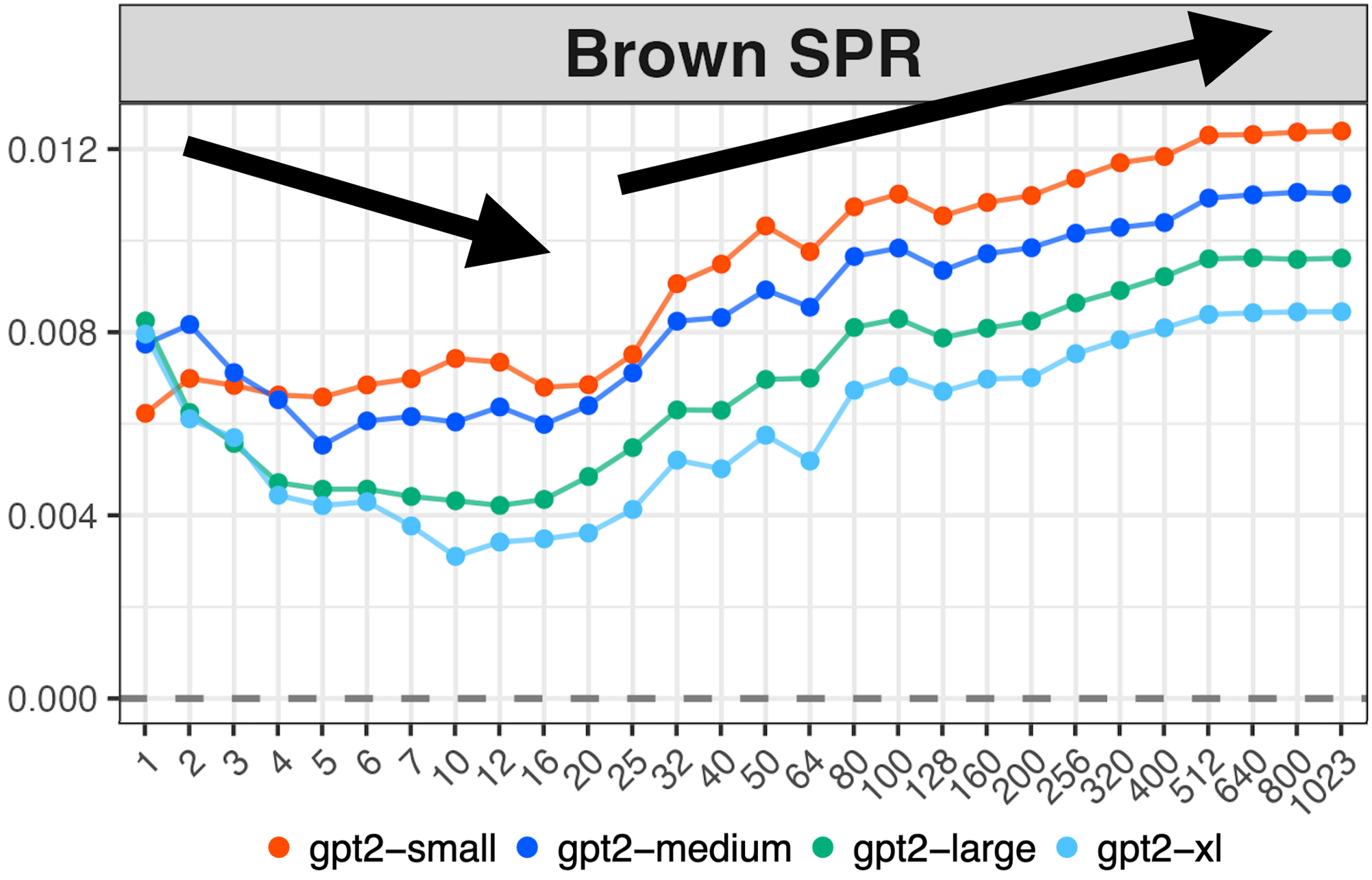


← Shorter Context

Abbreviations: SPR = self-paced reading

Exp. 1: Vary Context from 1 to 1,023 Tokens

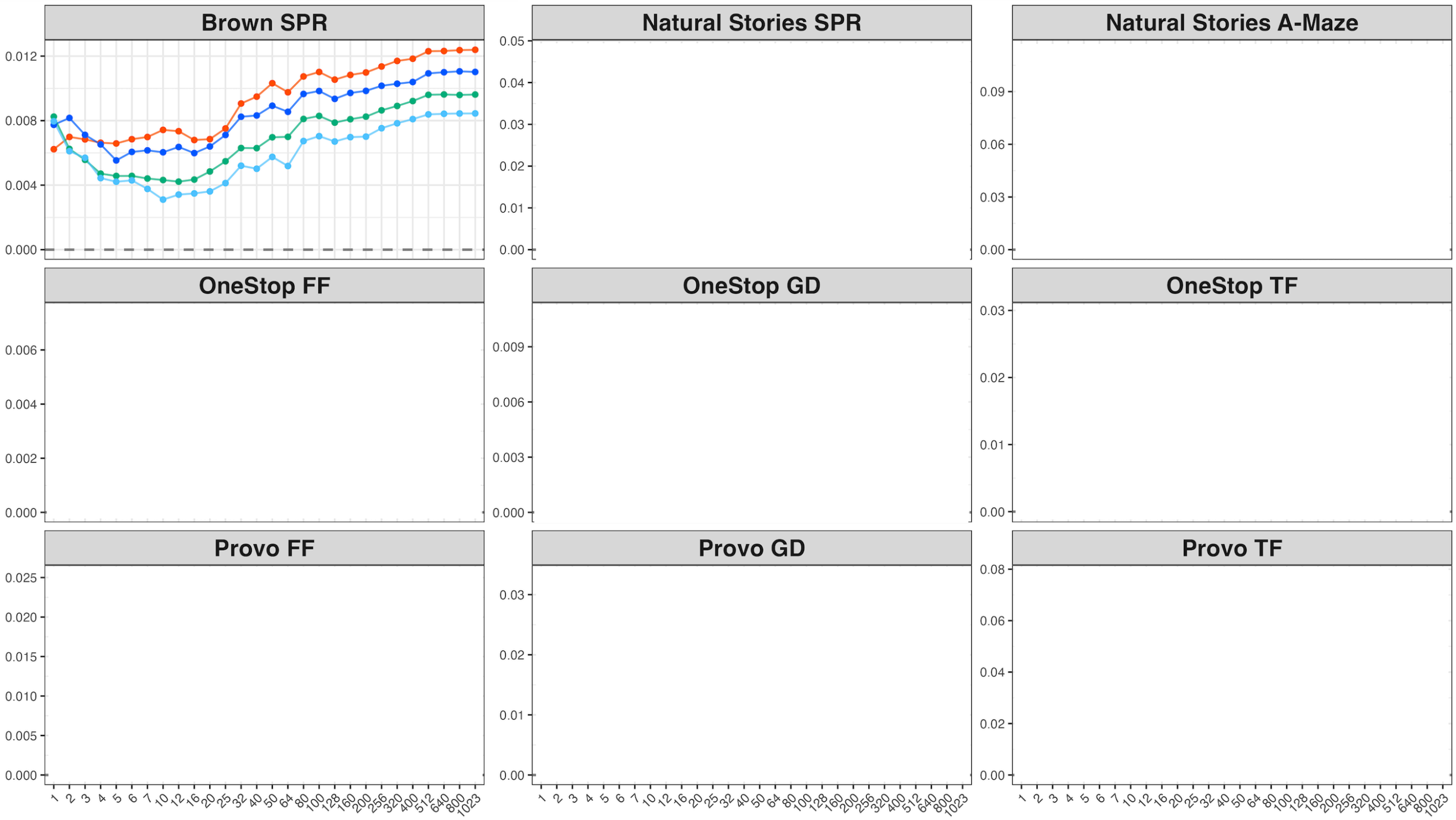
↑ Predictive Power for RTs



← Shorter Context

Abbreviations: SPR = self-paced reading

Exp. 1: Vary Context from 1 to 1,023 Tokens

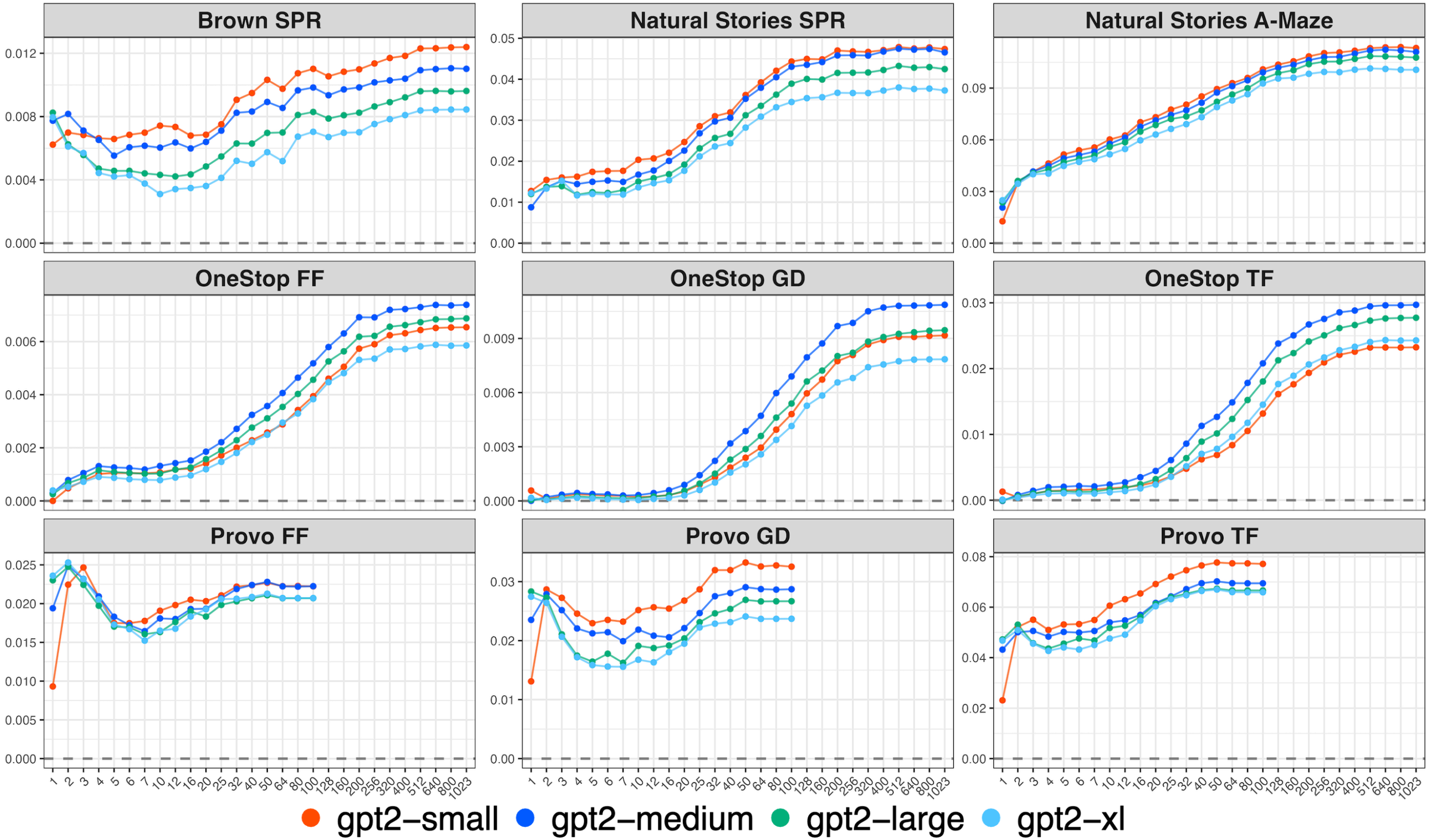


←Shorter Context

Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

Same pattern with GPT-Neo, GPT-J, and OPTs on Natural Stories SPR (Appendix C)

Exp. 1: Vary Context from 1 to 1,023 Tokens

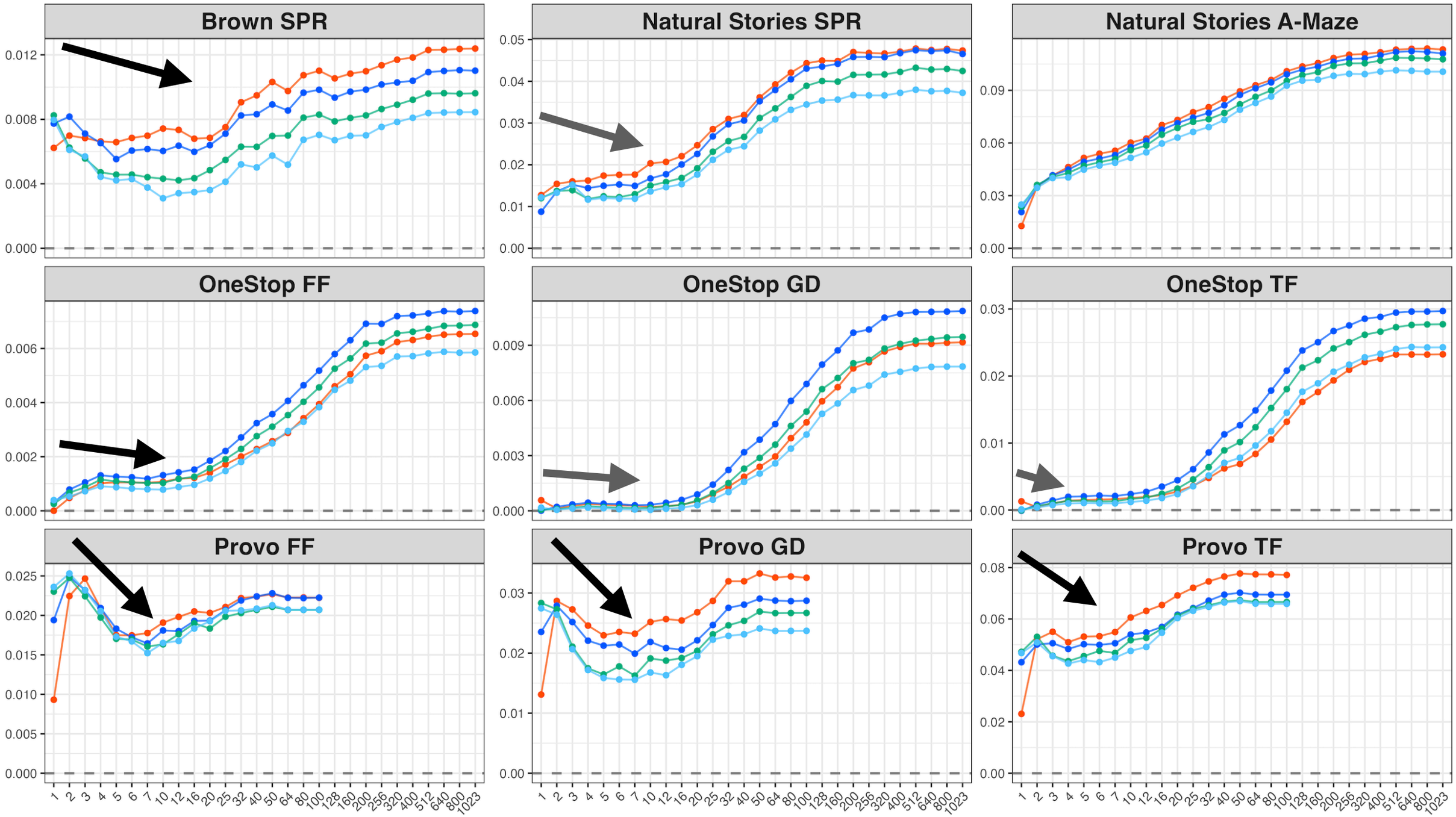


←Shorter Context

Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

Same pattern with GPT-Neo, GPT-J, and OPTs on Natural Stories SPR (Appendix C)

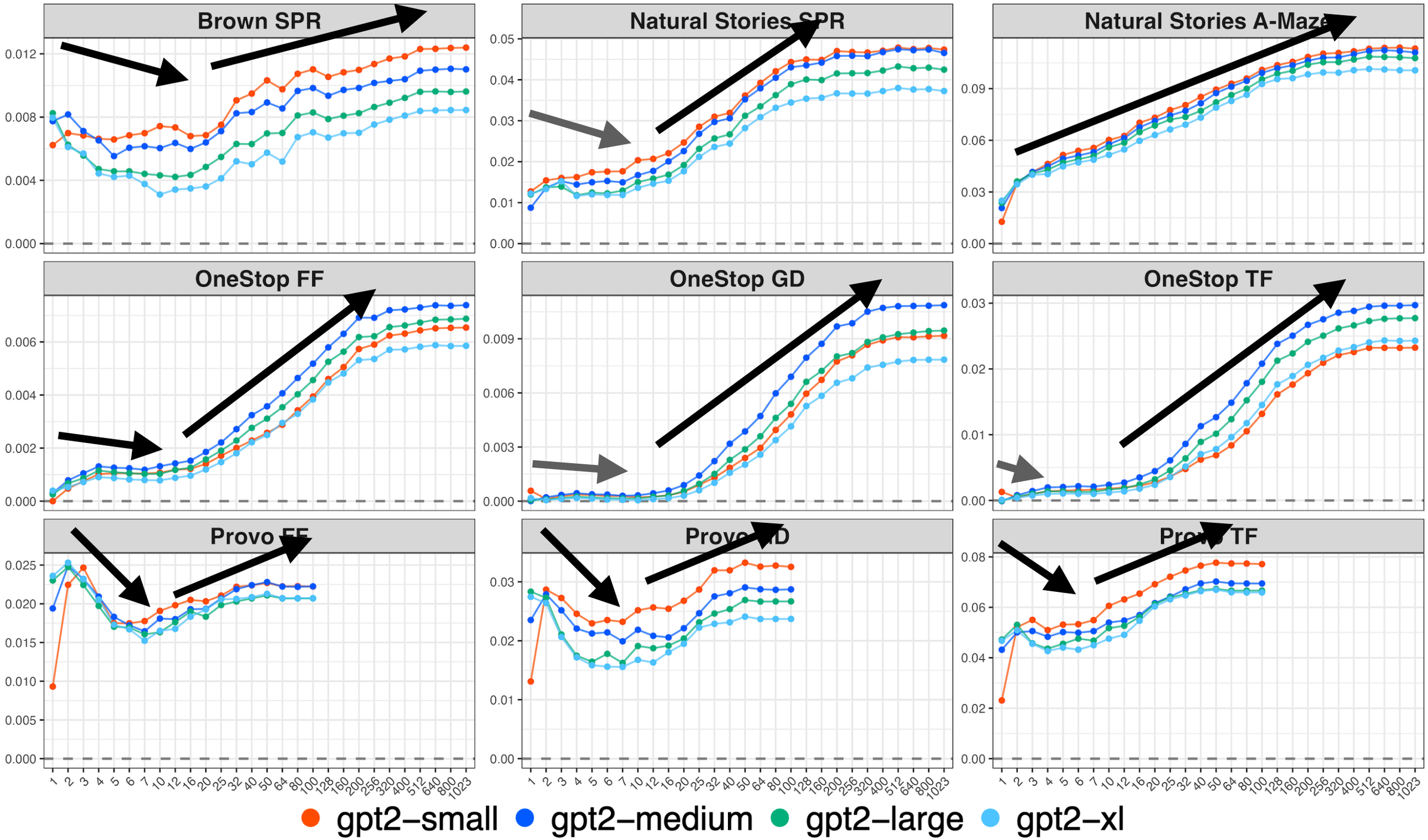
Exp. 1: Vary Context from 1 to 1,023 Tokens



Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

Same pattern with GPT-Neo, GPT-J, and OPTs on Natural Stories SPR (Appendix C)

Exp. 1: Vary Context from 1 to 1,023 Tokens

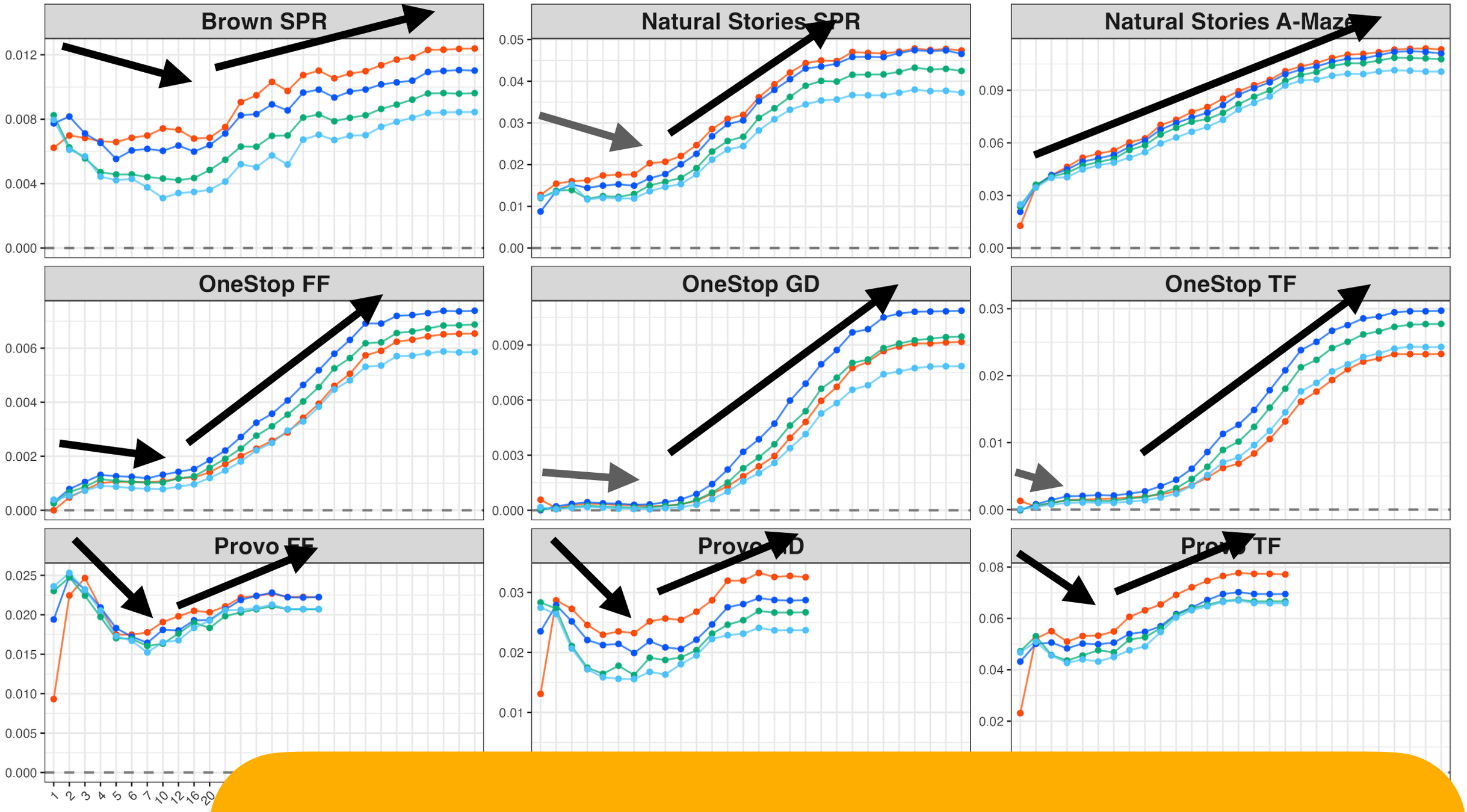


←Shorter Context

Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

Same pattern with GPT-Neo, GPT-J, and OPTs on Natural Stories SPR (Appendix C)

Exp. 1: Vary Context from 1 to 1,023 Tokens



😎 **U-/(S-) shaped curve:**
shorter-is-better locally, *longer-is-better* globally

Abbreviations:

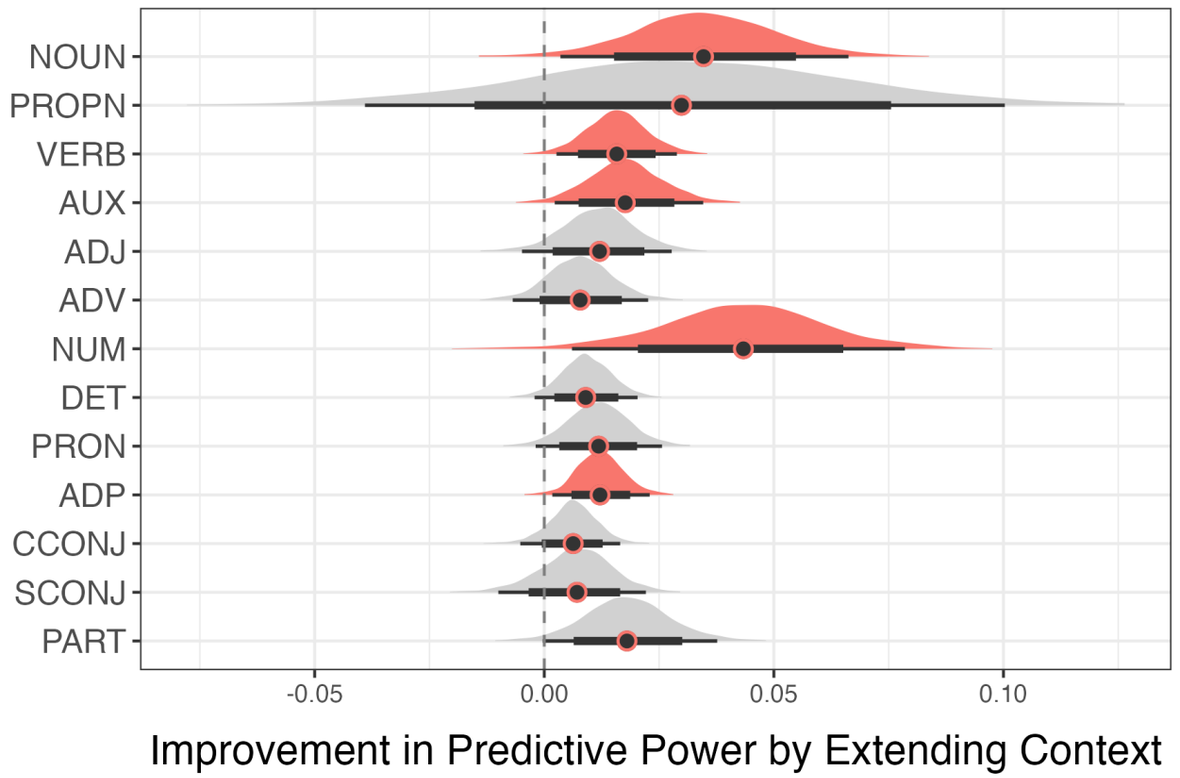
Same pattern with

Which Words Benefit from Long Context?

- Where does the fit to RTs improve with a **long** context, compared to a **short** context?

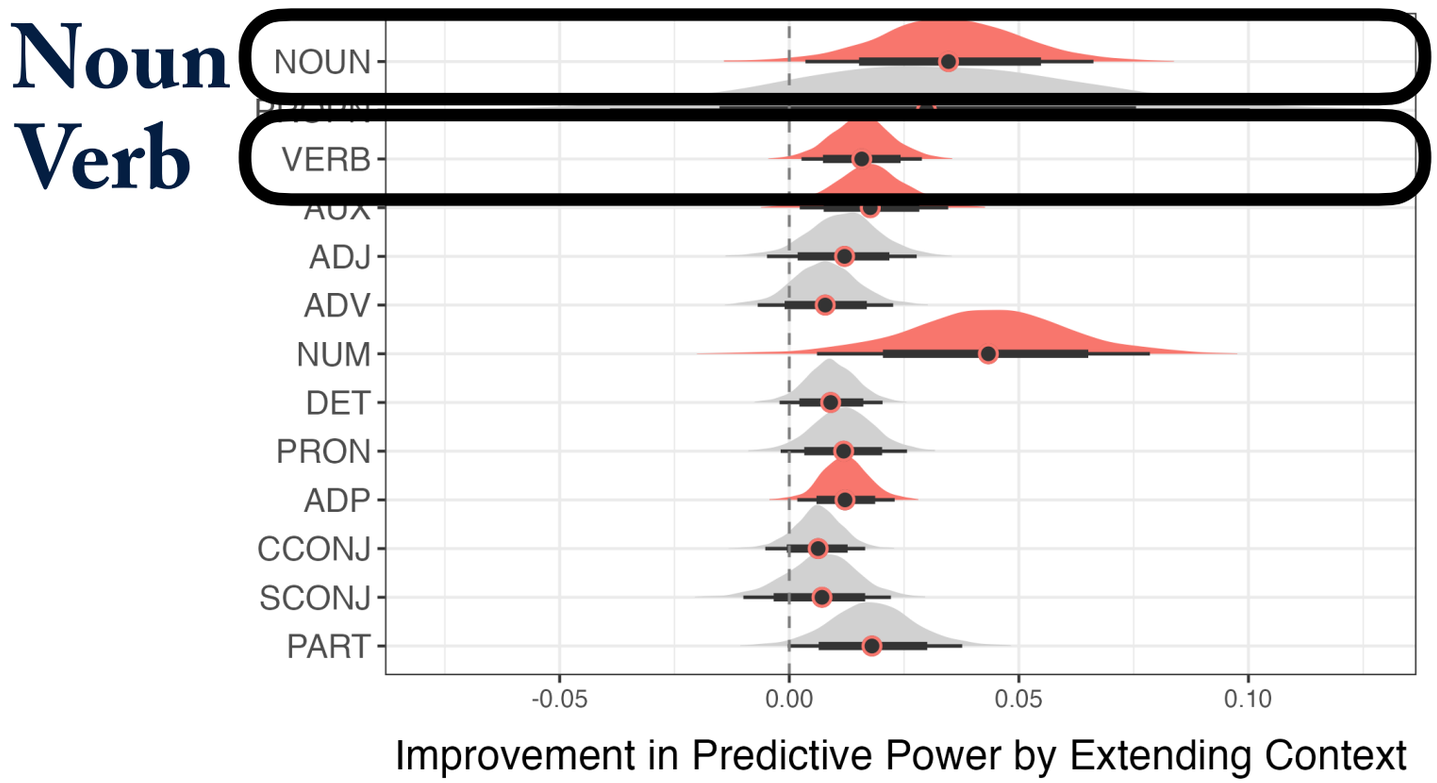
Which Words Benefit from Long Context?

- Where does the fit to RTs improve with a **long** context, compared to a **short** context?



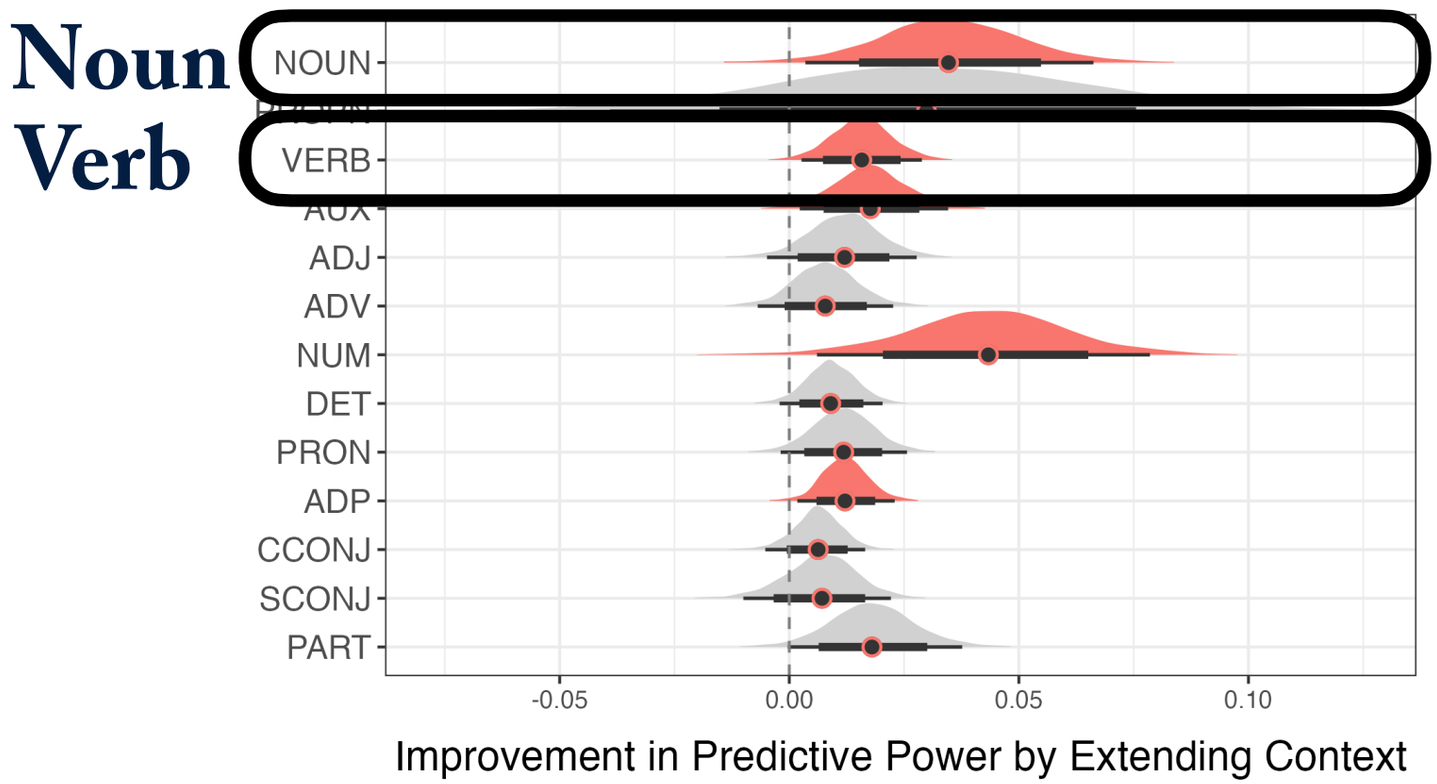
Which Words Benefit from Long Context?

- Where does the fit to RTs improve with a **long** context, compared to a **short** context?



Which Words Benefit from Long Context?

- Where does the fit to RTs improve with a **long** context, compared to a **short** context?



- Hypothesis: long context helps because the LMs track discourse (entity) relations:

Tom pushed ***John.*** ***The poor boy*** ____

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times

Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times

Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

Entity mentioned more than once

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times

Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

Entity mentioned more than once

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**

Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**

Customers hastily vacated their tables as ~~the tall buffalo hunter~~ **he** pushed open the doors and walked towards the bar.

Only ~~Blue Throat~~ **he** stayed where he was, his elbows supporting his massive frame. He leered at ~~the stranger~~ **him**.

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**
  - Singleton-Disrupted: pronominalize the entities that appear **only once**

Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

Entity mentioned only once

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**
  - Singleton-Disrupted: pronominalize the entities that appear **only once**

~~Customers~~ hastily vacated ~~their tables~~ as the tall buffalo hunter pushed open ~~the doors~~ and walked towards ~~the bar~~. Only Blue Throat stayed where he was, ~~his elbows~~ supporting ~~his massive frame~~. He leered at the stranger.

They **it** **them** **them** **they** **it**

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**
  - Singleton-Disrupted: pronominalize the entities that appear **only once**

  Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

- Only the context is modified:
  Both the predicted word and reading time are measured on the original word

Overall, 21,640 and 22,695 tokens are changed in Singleton-Disrupted and Coref-Disrupted, respectively 10

Exp. 2: Hide Coreference in the Context

- Give the LM a *counterfactual* context and measure the change in fit to human reading times
  - Coref-Disrupted: pronominalize the entities that appear **more than once**
  - Singleton-Disrupted: pronominalize the entities that appear **only once**

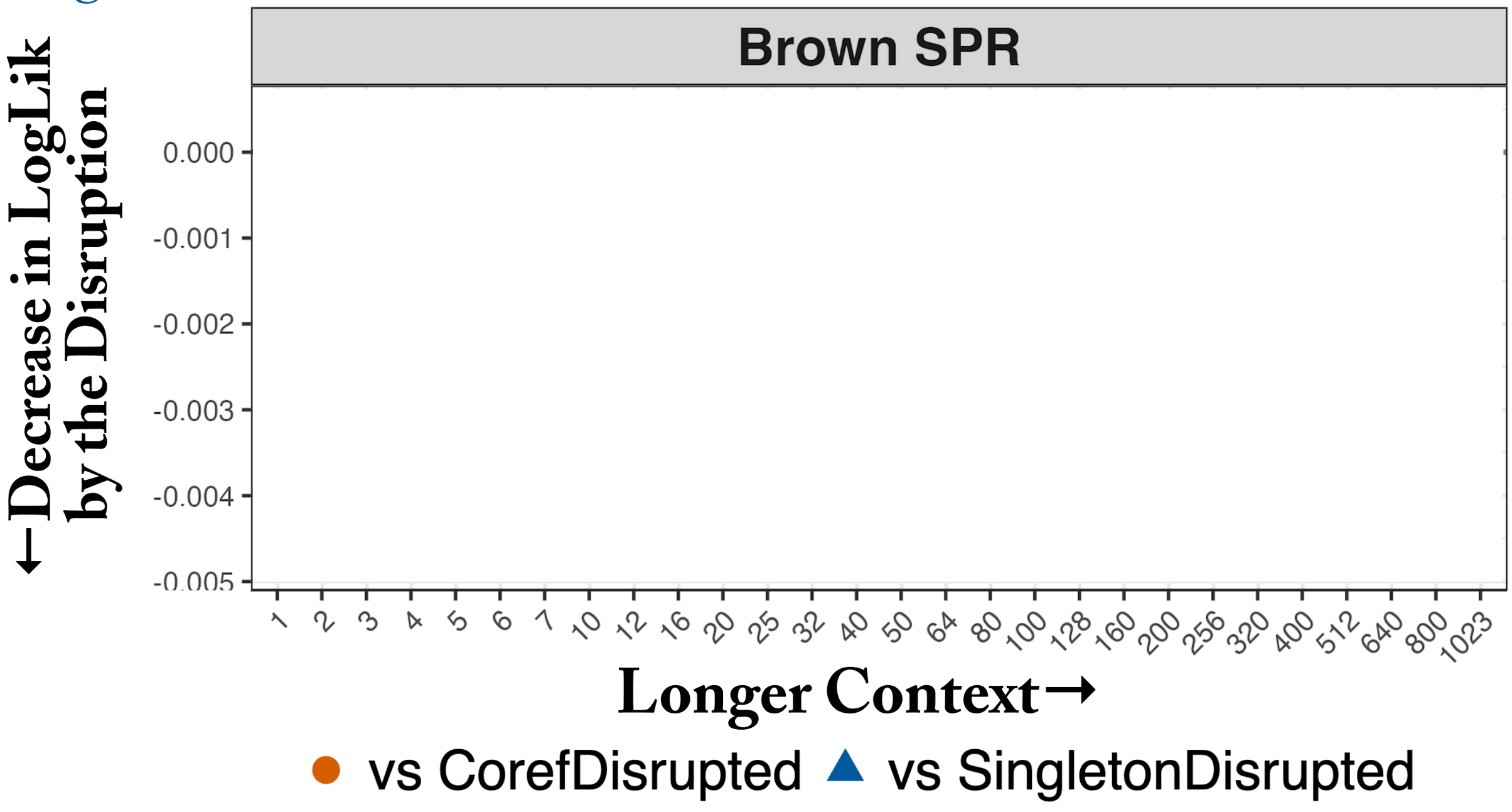
  Customers hastily vacated their tables as the tall buffalo hunter pushed open the doors and walked towards the bar. Only Blue Throat stayed where he was, his elbows supporting his massive frame. He leered at the stranger.

- Only the context is modified:
  Both the predicted word and reading time are measured on the original word
- Prediction:
  If entity tracking matters, the predictive power **drops more** in Coref-Disrupted than in Singleton-Disrupted, particulary in long contexts

Overall, 21,640 and 22,695 tokens are changed in Singleton-Disrupted and Coref-Disrupted, respectively 10

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

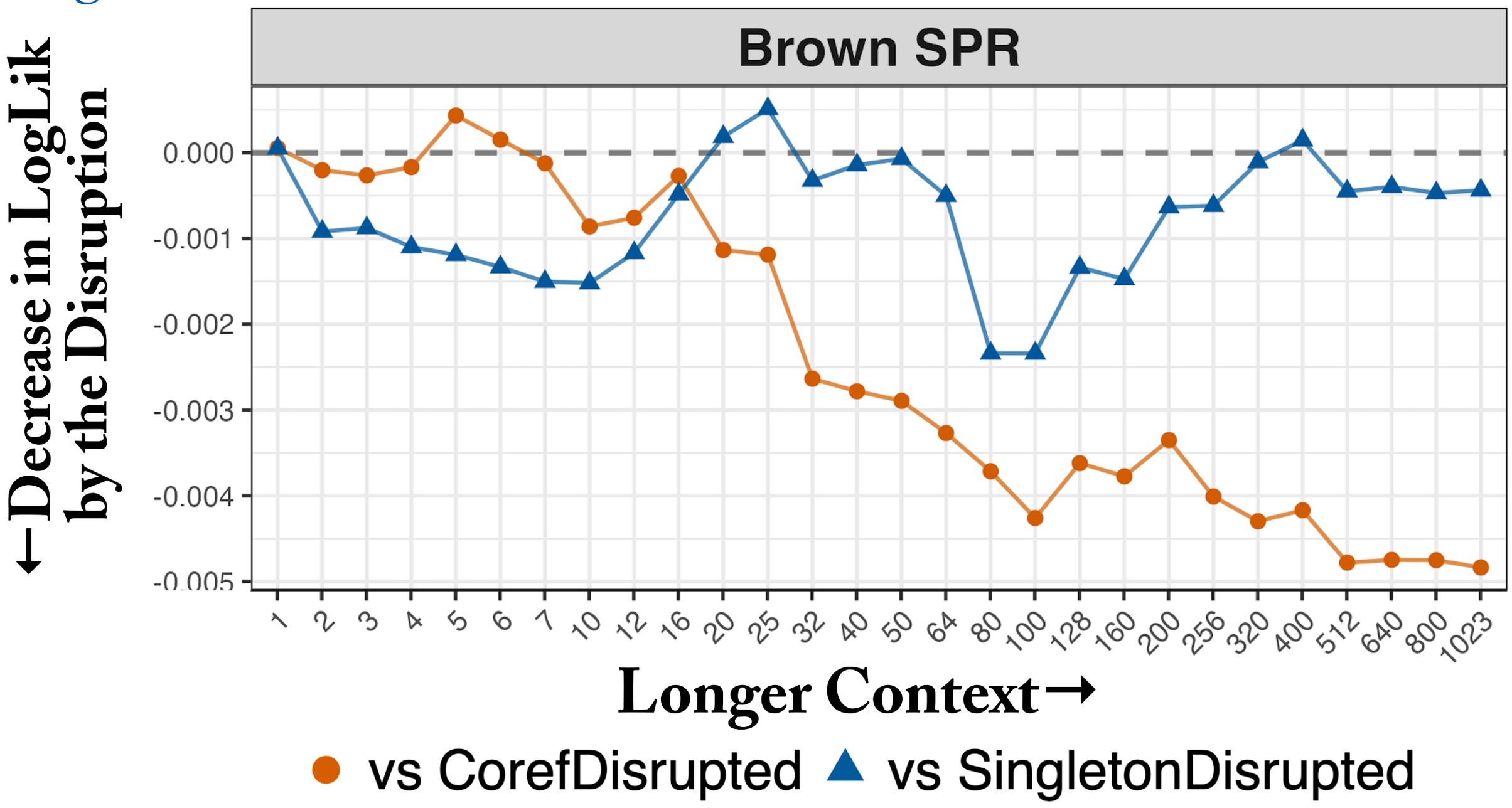


Abbreviations: SPR = self-paced reading

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

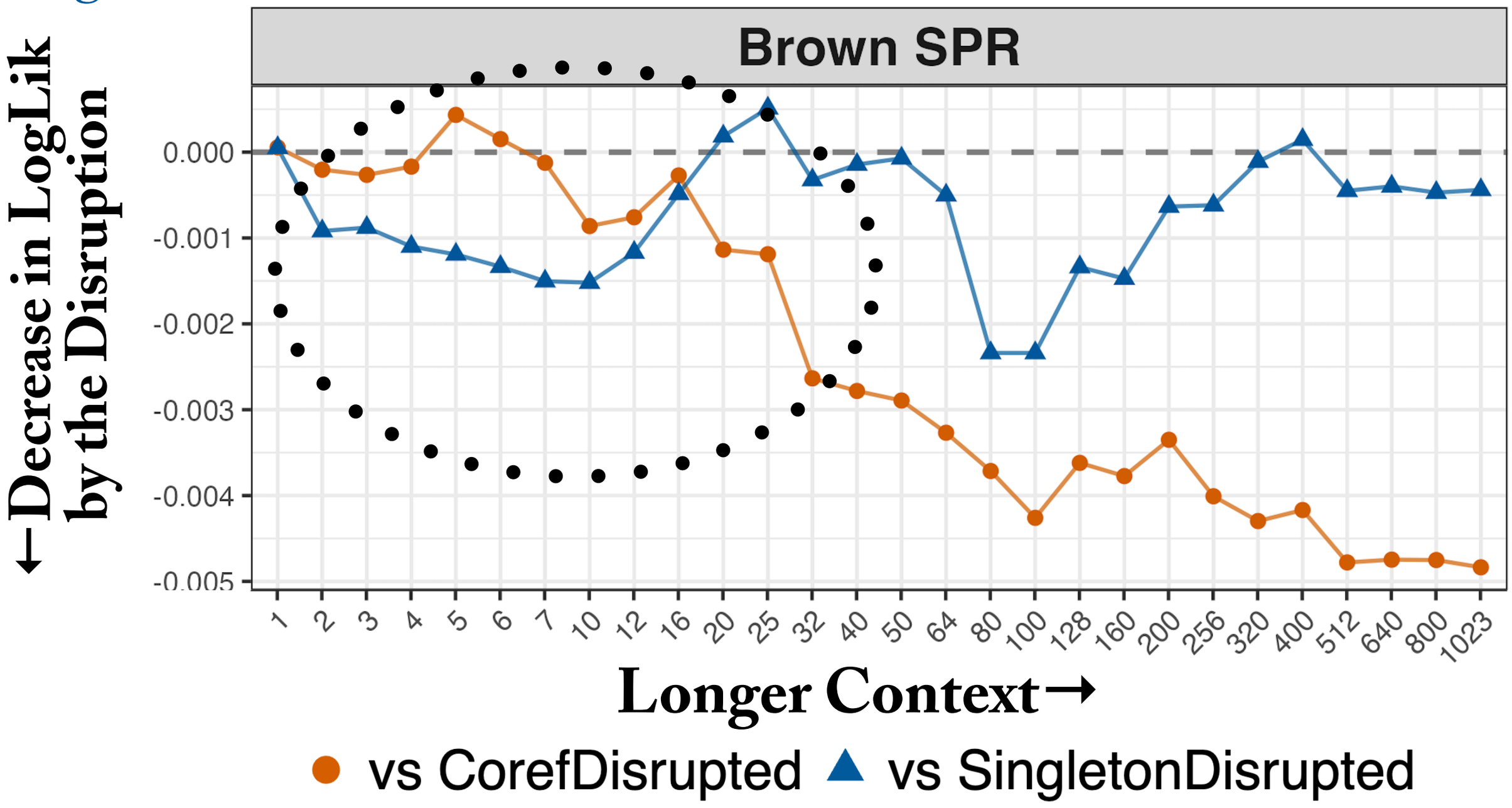


Abbreviations: SPR = self-paced reading

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

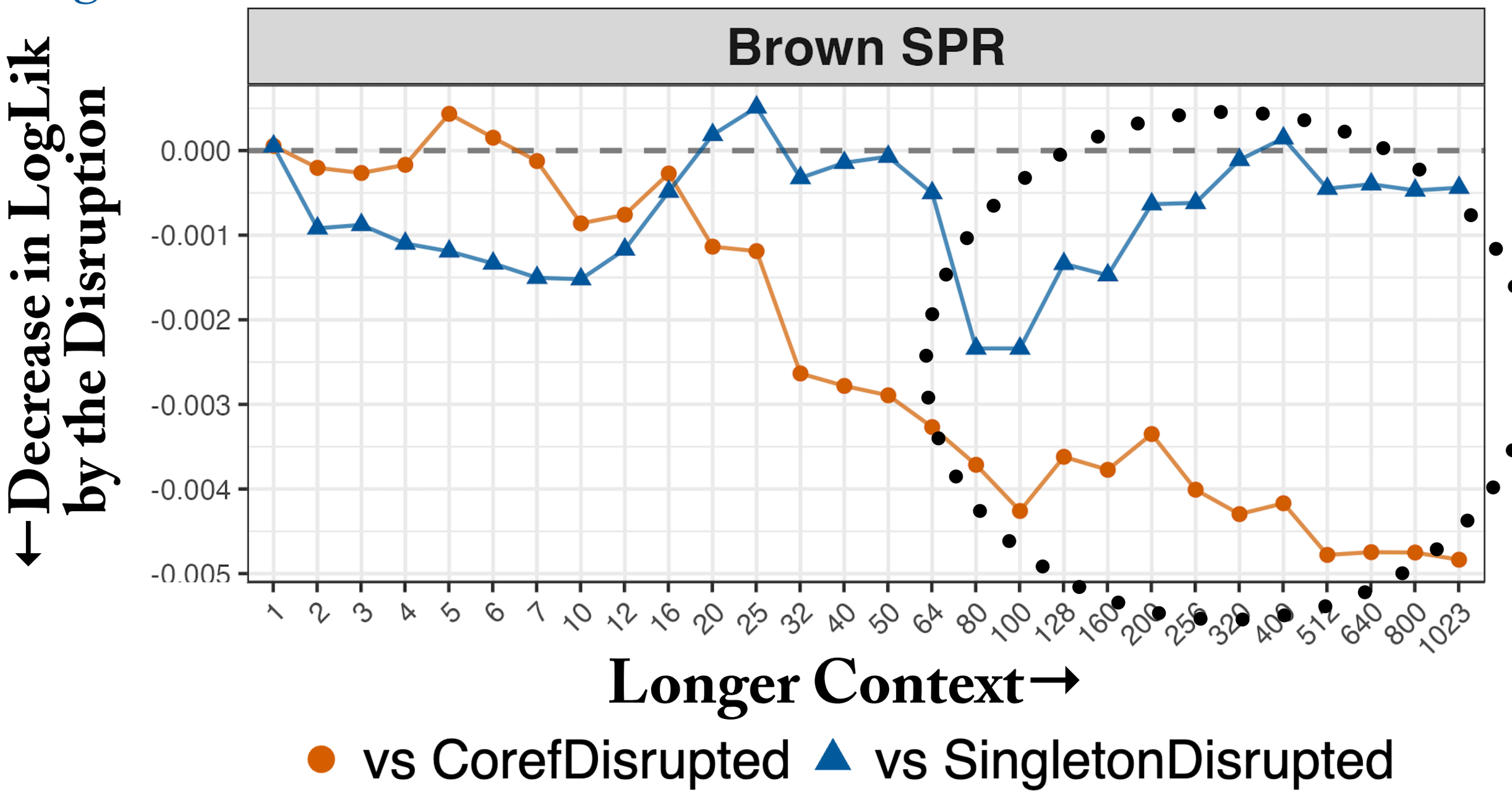


Abbreviations: SPR = self-paced reading

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

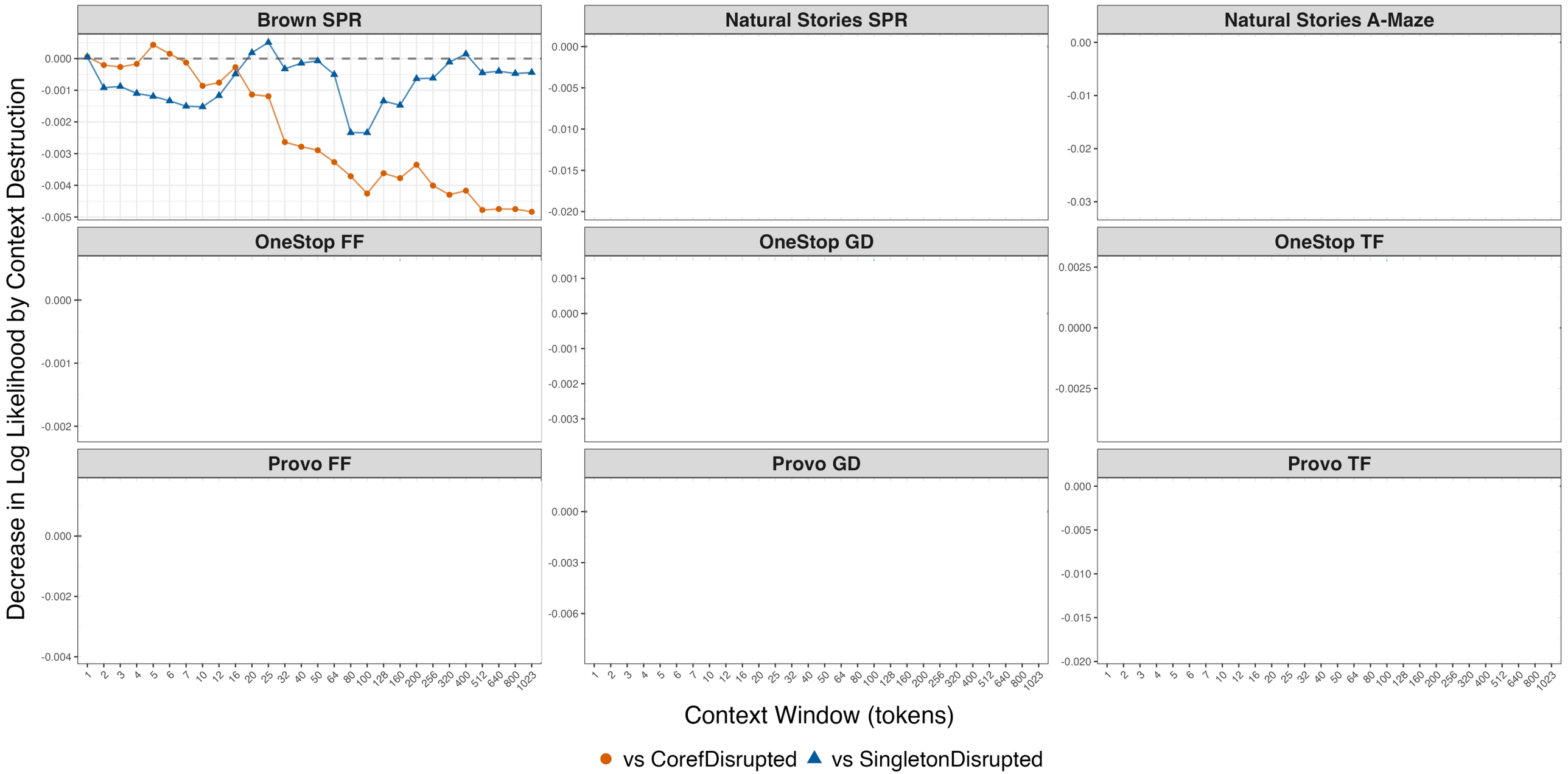


Abbreviations: SPR = self-paced reading

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

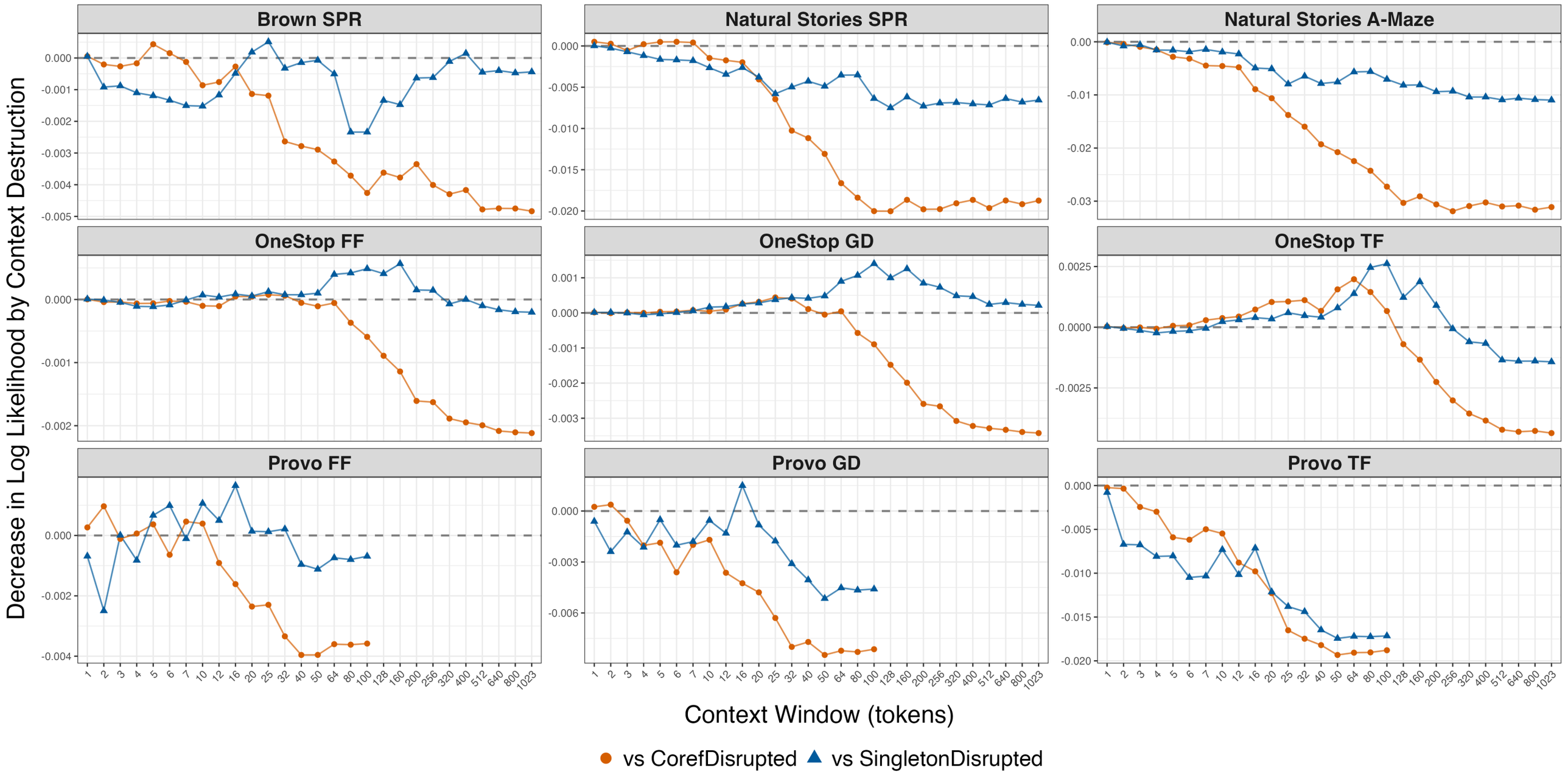


Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted : pronominalize the entities that appear **more than once**
- Singleton-Disrupted: pronominalize the entities that appear **only once**

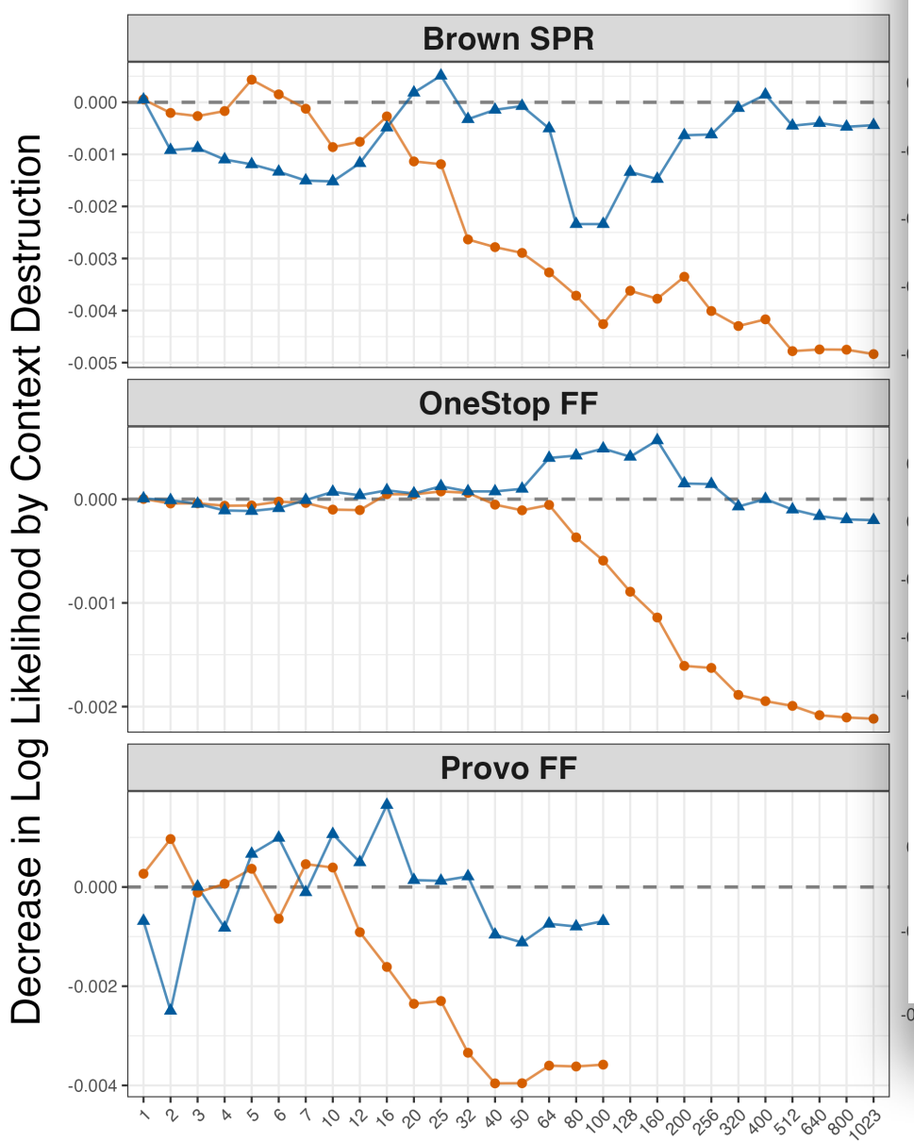


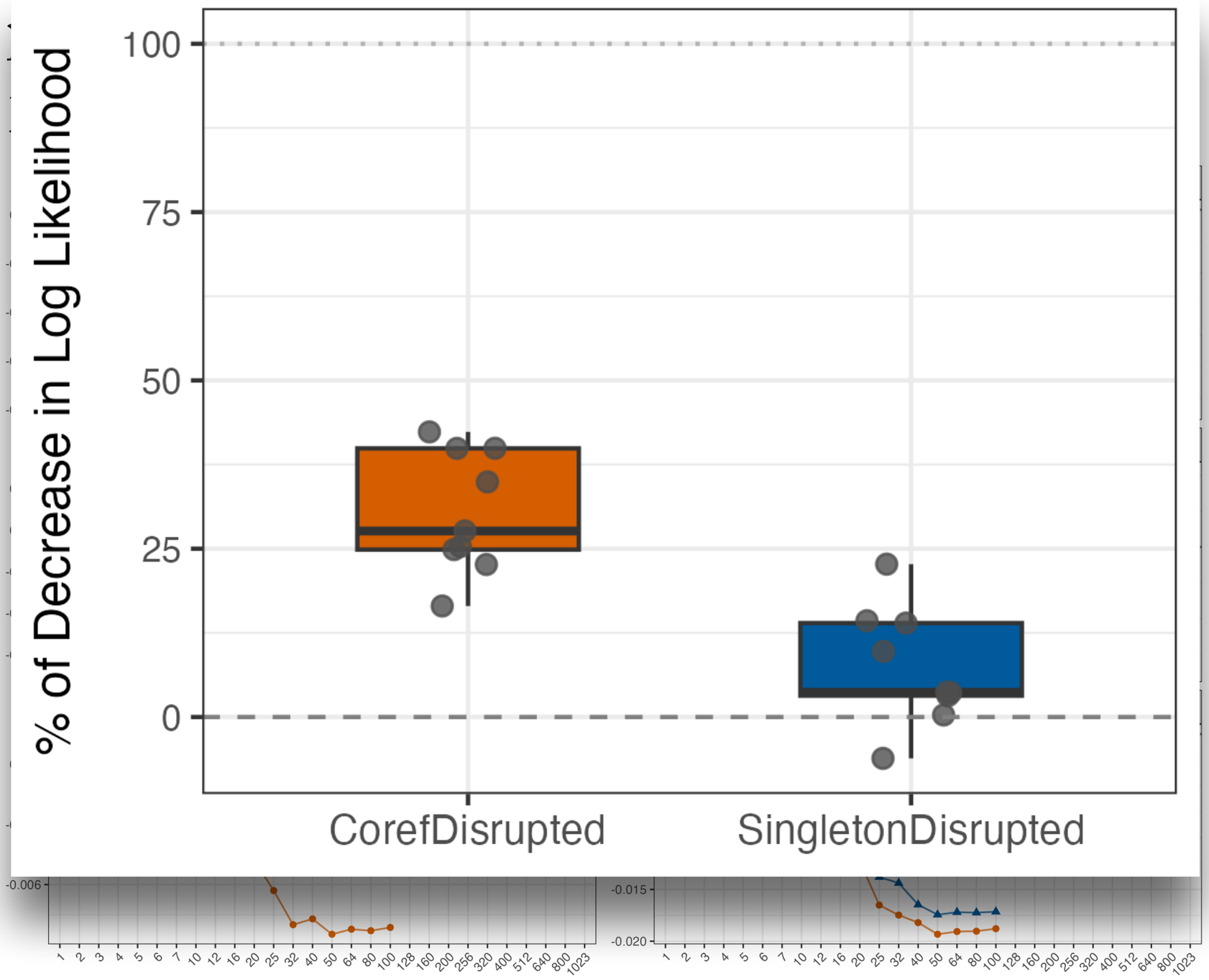
Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted :
- Singleton-Disrupted:



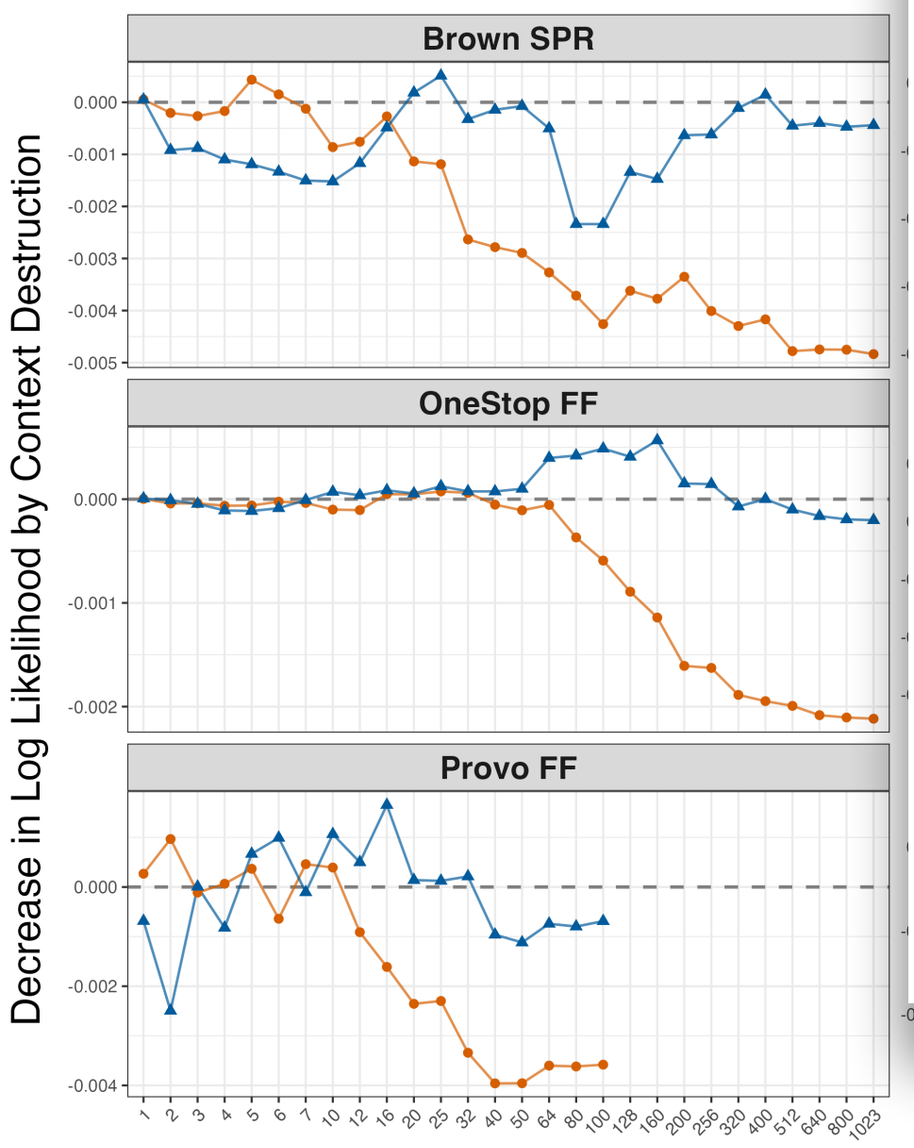


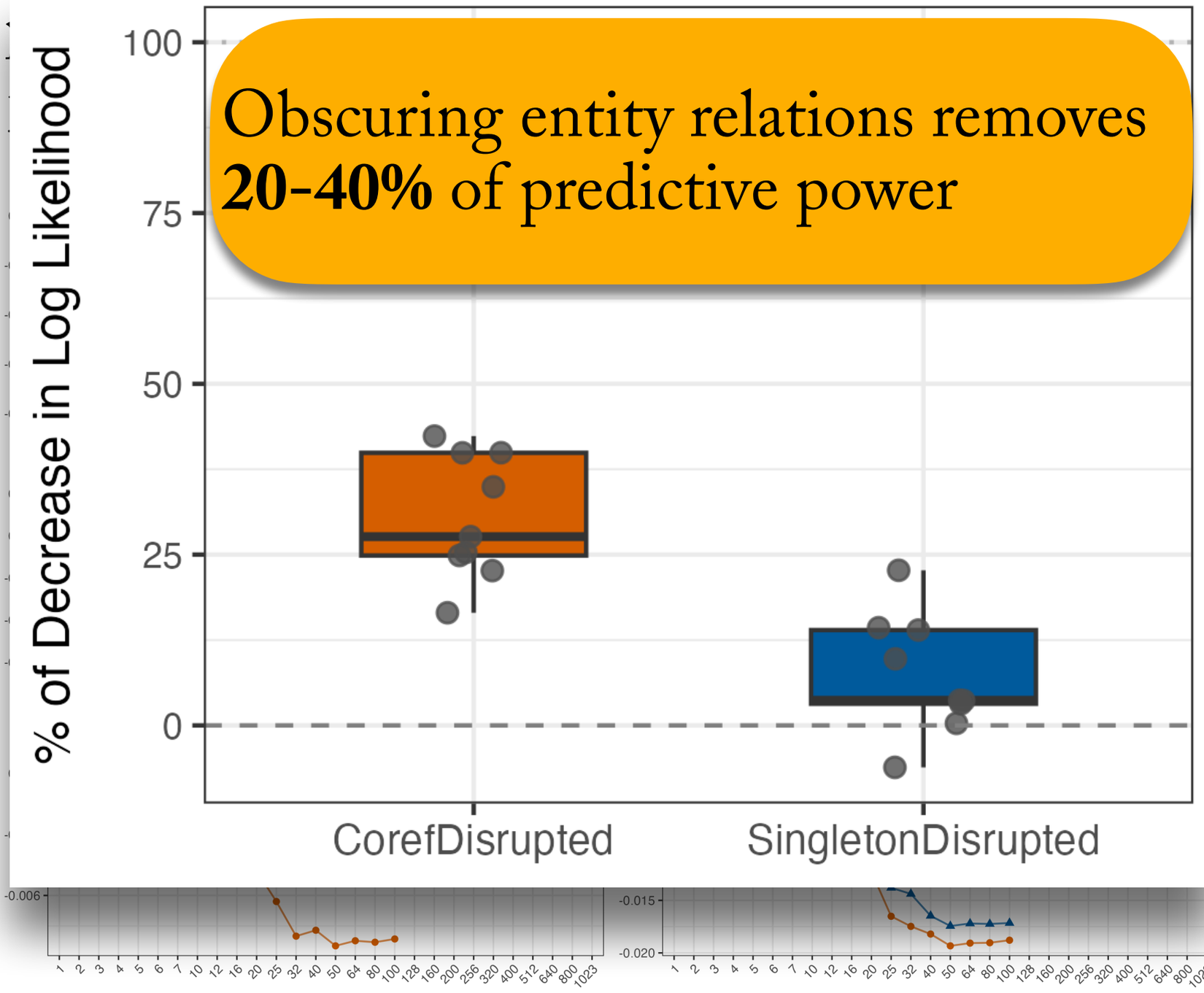
Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Exp. 2: Hide Coreference in the Context

- Coref-Disrupted :
- Singleton-Disrupted:





Abbreviations: SPR = self-paced reading; FF = first fixation; GD = gaze duration; TF = total fixation time

The results of GPT-2 small; same pattern for medium, large, and XL (Appendix D)

Take-home Messages

- Context length vs. the predictive power for RTs is U-shaped: *shorter-is-better* locally (<20 tokens), *longer-is-better* globally

Paper

Take-home Messages

- Context length vs. the predictive power for RTs is U-shaped: *shorter-is-better* locally (<20 tokens), *longer-is-better* globally
- Obscuring entity relations removes about 20–40% of the predictive power of long-context surprisal

Paper

Take-home Messages

- Context length vs. the predictive power for RTs is U-shaped: *shorter-is-better* locally (<20 tokens), *longer-is-better* globally
- Obscuring entity relations removes about 20–40% of the predictive power of long-context surprisal
- *Counterfactual* inference-time experiment: a general way to test what in the context matters

Paper