Paper ID: 1866

“No, they did not”: Dialogue response dynamics in pre-trained language models

Sanghee J. Kim[1], **Lang Yu**[2] **& Allyson Ettinger**[1]

[1] Department of Linguistics, University of Chicago
[2] Meta

The 29th International Conference on Computational Linguistics (COLING), Oct 12-17, 2022, Gyeongju, Republic of Korea

Overview

- Analysis of pre-trained language models (PLMs) on linguistic competence
- PLMs' sensitivity to pragmatics & discourse information
- **Dialogue response dynamics**, focusing on **at-issueness**, and **ellipsis**

Dialogue response dynamics

“ The nurse, who has interest in French cuisine, adopted a rescue dog. ”

Dialogue response dynamics

“ The nurse, who has interest in French cuisine, adopted a rescue dog. ”

- Inside the main clause
- **At-issue** (main point)[1]

[1] Potts. (2005). *The logic of conventional implicatures.*

Dialogue response dynamics

“ The nurse, who has interest in French cuisine, adopted a rescue dog. ”

Dialogue response dynamics

“ The nurse, who has interest in French cuisine, adopted a rescue dog. ”

- Inside the embedded clause (or *appositive relative clause* (ARC))
- **Not-at-issue** (peripheral point)[1]

[1] Potts. (2005). *The logic of conventional implicatures.*

Dialogue response dynamics

Certain parts of an utterance are … [2-4]

{ At-issue / Not-at-issue } and { **more** / **less** } likely to receive response than others

[2] Amaral et al. (2007). *Ling. & Philosophy, 30*(6). [3] Koev. (2013). *Apposition and the structure of discourse*. [4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Dialogue response dynamics

Certain parts of an utterance are … [2-4]

{ **At-issue** / Not-at-issue } and { **more** / less } likely to receive response than others

[2] Amaral et al. (2007). *Ling. & Philosophy, 30*(6). [3] Koev. (2013). *Apposition and the structure of discourse*. [4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Dialogue response dynamics

Certain parts of an utterance are … [2-4]

{ At-issue / **Not-at-issue** } and { more / **less** } likely to receive response than others

[2] Amaral et al. (2007). *Ling. & Philosophy, 30*(6). [3] Koev. (2013). *Apposition and the structure of discourse*. [4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Dialogue response dynamics

Speaker:

"The nurse, who has interest in French cuisine, adopted a rescue dog."

Listener:

"**No**,

Dialogue response dynamics

Speaker:

"The nurse, who has interest in French cuisine, adopted a rescue dog."

Listener:

[*targeting at-issue*] "**No**, he didn't (adopt a rescue dog)."

Dialogue response dynamics

Speaker:

"The nurse, who has interest in French cuisine, adopted a rescue dog."

Listener:

[*targeting at-issue*] "**No**, he didn't (adopt a rescue dog)."

[*targeting not-at-issue*] "**No**, he doesn't (have interest in French cuisine)."

Dialogue response dynamics

Speaker:

"The nurse, who has interest in French cuisine, adopted a rescue dog."

Listener:

[*targeting at-issue*]	"**No**, he didn't (adopt a rescue dog)."
[*targeting not-at-issue*]	"**No**, he doesn't (have interest in French cuisine)."
[*targeting not-at-issue*]	"**Wait no**, he doesn't (have interest in French cuisine). "

Dialogue response dynamics

Dialogue response dynamics:

Interaction of response type and at-issue status of prior utterance

Humans are sensitive to the dynamics – are PLMs too?

Outline

Part 1	Dialogue response dynamics
Part 2	Experiments
Part 3	Error analysis
Part 4	Summary & Discussion

Part 2

Experiments

Header preference

Can PLMs prefer response headers
based on the type of content the response targets?

Header preference

Criterion #1 – Not-at-issue content

> Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he does not”.

Header preference

Criterion #1 – Not-at-issue content

Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he does not”.

Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he does not”.

Header preference

Criterion #1 – Not-at-issue content

P (Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he does not”. **)**

P (Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he does not”. **)**

Header preference

Criterion #1 – Not-at-issue content[4]

P (Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he does not”. **) – P (** Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he does not”. **) > 0**

[4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Header preference

Criterion #2 – At-issue content

> Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he did not”.

Header preference

Criterion #2 – At-issue content

Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he did not”.

Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he did not”.

Header preference

Criterion #2 – At-issue content[4]

P (Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “Wait no, he did not”. **) – P (** Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he did not”. **) ≈ 0**

[4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Header preference

- Template for test items:

NAME1 said, “**Context sentence**”, and NAME2 replied, “**Response sentence.**”

NP, who VP1, VP2.

HEADER$_{\{No/Wait\ no\}}$, PRONOUN AuxVerb not.

- VP1 and VP2 will always be targeted by different auxiliary verbs
- 6 AuxVerbs: ***is, was, does, did, has, could***

Header preference

Model Tested:

- Causal (unidirectional) language model (**CLM**):

  DistilGPT2[5]

- Maked language models (**MLMs**):

  BERT[6], **RoBERTa**[7], **XLM-ROBETa**[8], **DistilBERT**[9], **DistilRoBERTa**[9]

[5] HuggingFace. [6] Devlin et al. (2019). *NAACL-HLT.* [7] Liu et al. (2019). [8] Conneau et al. (2020). *LREC.* [9] Sanh et al. (2019).

Header preference

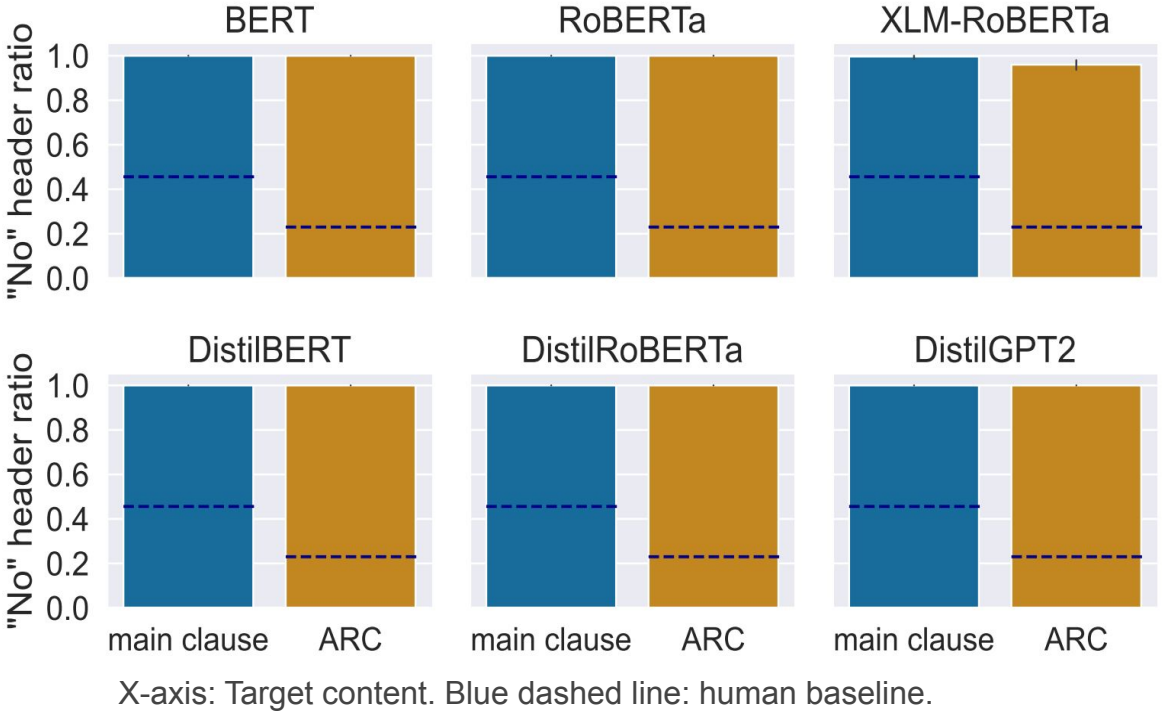
- Input sequence:

  Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “{No / Wait no}, he {did / does} not.”

- Measurement
  - CLM: Conditional log prob. of full sequence, normalized by length
  - MLM: Pseudo-log-likelihoods[e.g., 10]

[10] Salazar et al. (2020). *ACL*.

Header preference



X-axis: Target content. Blue dashed line: human baseline.

- “No” header > “Wait no” header regardless of target content

Header preference

Can PLMs prefer response headers
based on what type of content the response targets?

Not quite!

- “No” header – too frequent/strong a response?
- Atypical measurement with MLMs?

Target preference

Can PLMs prefer certain type of response content based on response header?

Target preference

Criterion #1 – With “No” header

(...), “No, he did not”.

*Targets **at-issue** content*

(...), “No, he does not”.

*Targets **not-at-issue** content*

Target preference

Criterion #1 – With “No” header

P ((…), “No, he did not”. **)**

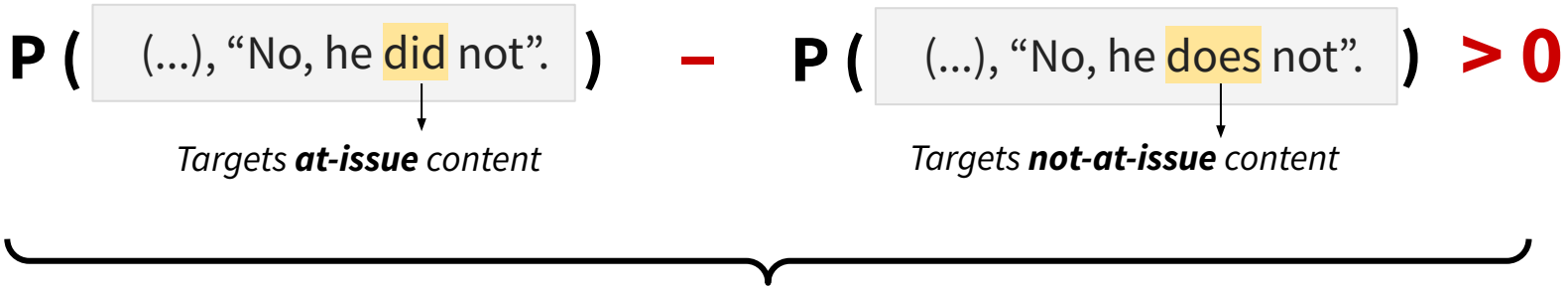
*Targets **at-issue** content*

P ((…), “No, he does not”. **)**

*Targets **not-at-issue** content*

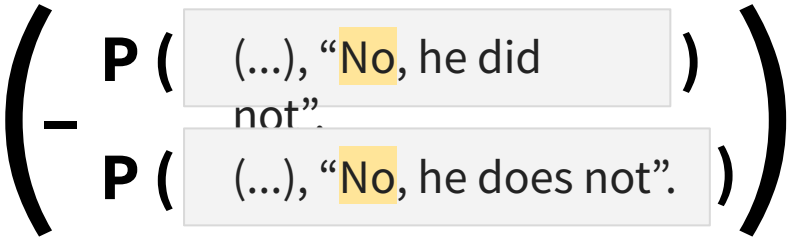
Target preference

Criterion #1 – With “No” header[4]



[4] Syrett & Koev. (2015). *J. of Sem, 32*(3).

Target preference



Target preference

Criterion #2 – With “Wait no” header

$$\left(P(\text{(...), “No, he did not”.}) - P(\text{(...), “No, he does not”.}) \right) > \left(P(\text{(...), “Wait no, he did not”.}) - P(\text{(...), “Wait no, he does not”.}) \right)$$

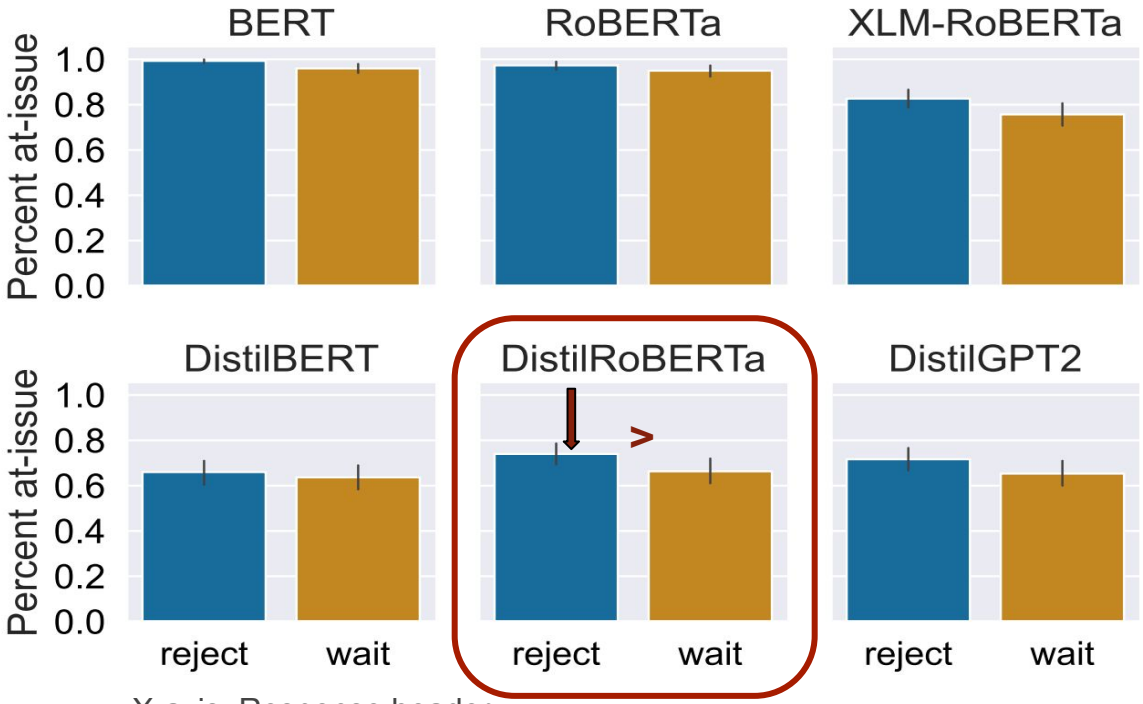
Target preference

- Input sequence:

  Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “{No / Wait no}, he CLM: {did / does} not.”
  MLM: [MASK] not.”

- Measurement
  - CLM: Conditional log prob. of full sequence, normalized by length
  - MLM: Probability of auxiliary verb at [MASK] position

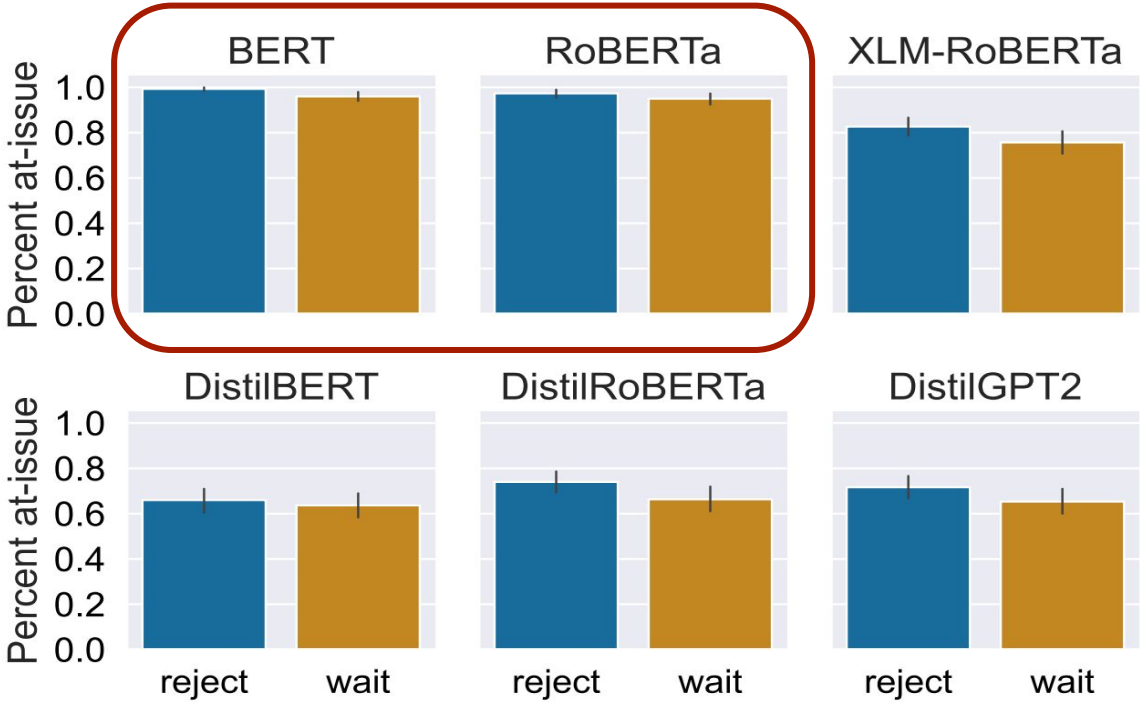
Target preference



X-axis: Response header.

- At-issue preference with “No”
- At-issue preference: “No” > “Wait no”

Target preference



X-axis: Response header.

- At-issue preference regardless of response header

Target preference

Can PLMs prefer certain type of response content based on response header?

It seems they can?

Target preference

However,

- Some models *always* prefer the at-issue content
- Recency effect?
  - Targeted at-issue content is always in the more recent position
  - Models simply targeting the most recent VP?

Conjunction

- Input sequence:

Marco said, “The nurse has interest in French cuisine and adopted a rescue dog,”
and Ellie replied, “{No / Wait no}, he CLM: {did / does} not.”
MLM: [MASK] not.”

Conjunction

- Input sequence:

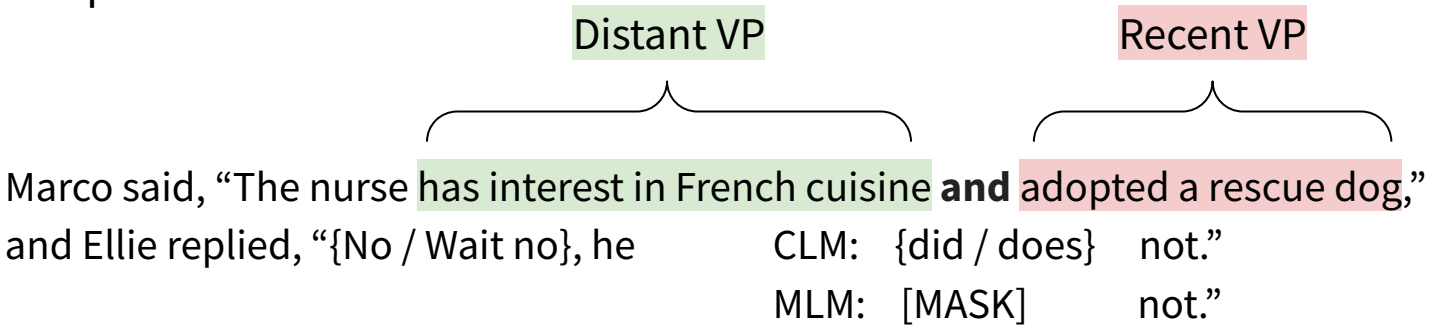
Both VPs are at-issue

Marco said, “The nurse has interest in French cuisine and adopted a rescue dog,”
and Ellie replied, “{No / Wait no}, he

CLM: {did / does} not.”
MLM: [MASK] not.”

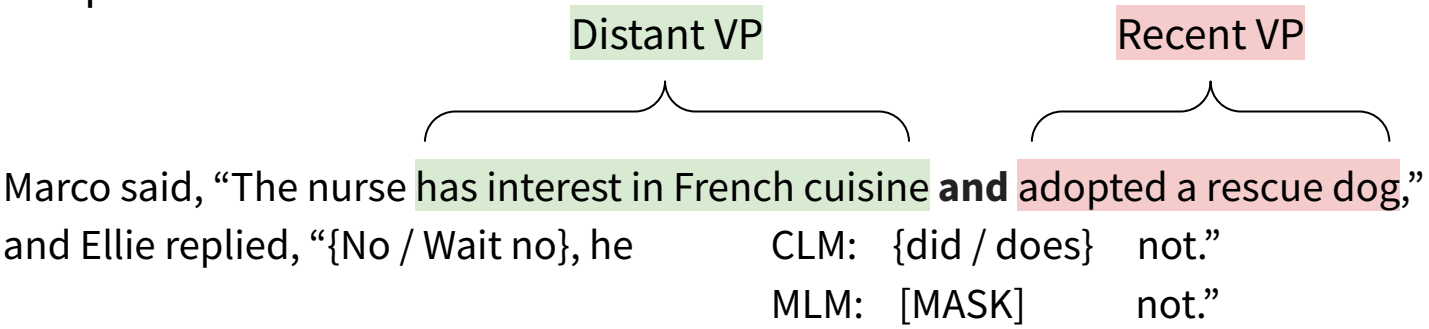
Conjunction

- Input sequence:



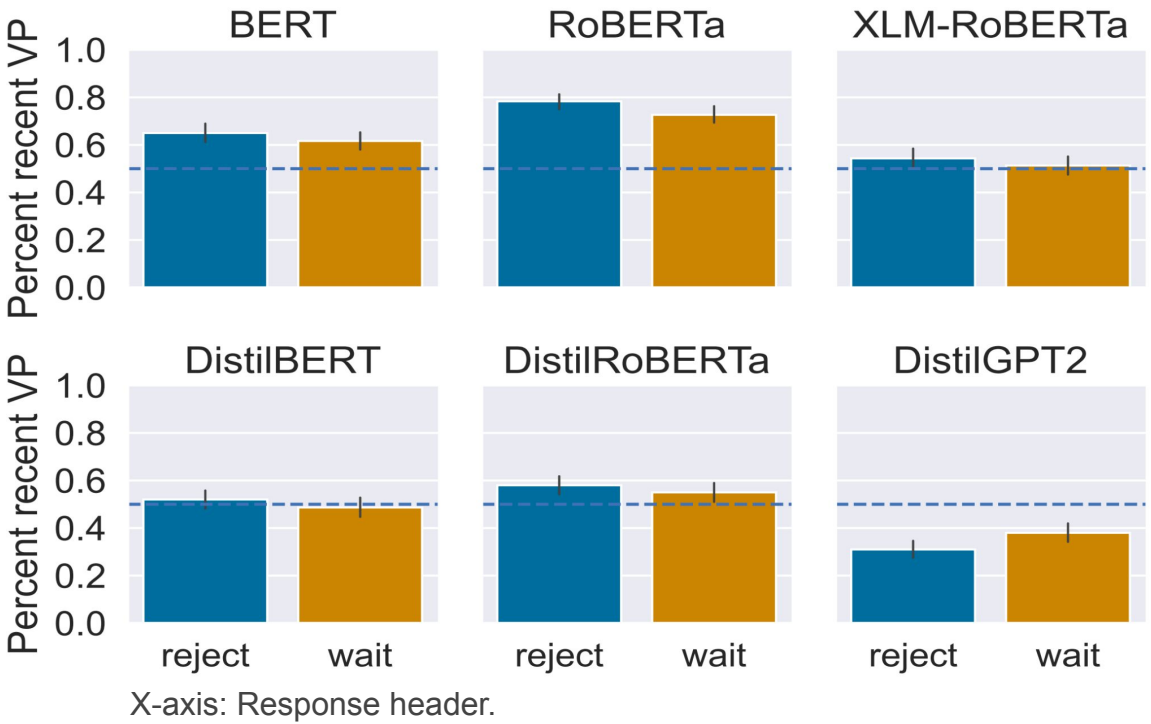
Conjunction

- Input sequence:



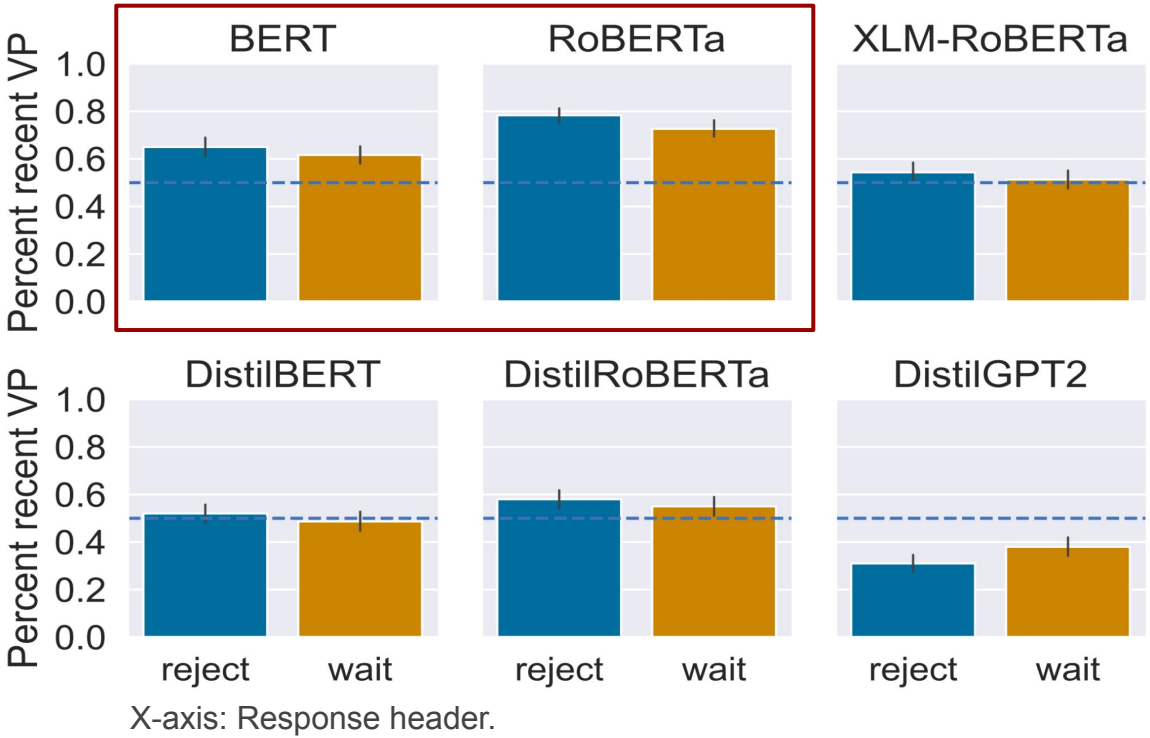
If Recency bias: **Preference for “… did not.” > 0.5**
If no Recency bias: **Preference for “… did not.” ≈ 0.5**

Conjunction

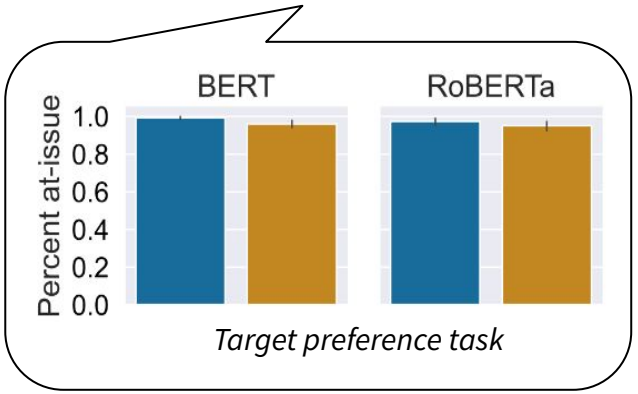


- Distilled models:
  - Around 50%
  - No strong recency bias

Conjunction



- BERT and RoBERTa
  - Trend of recency bias
  - But *weaker* than their at-issue preference



Conjunction

Is models' at-issueness preference guided by recency bias?

Cannot be explained by recency alone!

Probing

Are PLMs sensitive to the differing status of main vs. embedded clause?

Probing

- 3-class classification
    - Input: Token embeddings from the last hidden layer
    - Labels: (a) part of main clause (at-issue content)
      (b) part of embedded clause (not-at-issue content)
      (c) neither

Probing

Model	Accuracy (%)
BERT	99.9
RoBERTa	100
XLM-RoBERTa	99.2
DistilBERT	99.4
DistilGPT2	99.5
DistilRoBERTa	100

- Near perfect classification accuracy for all models

Probing

Are PLMs sensitive to the differing status of main clause vs. embedded clause content?

Yes!

Models are sensitive to the structural properties in dialogue dynamics

Ellipsis

Do PLMs have grasp of knowledge governing verb ellipsis?

Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “No, he didn’t ~~adopt a rescue dog~~.”

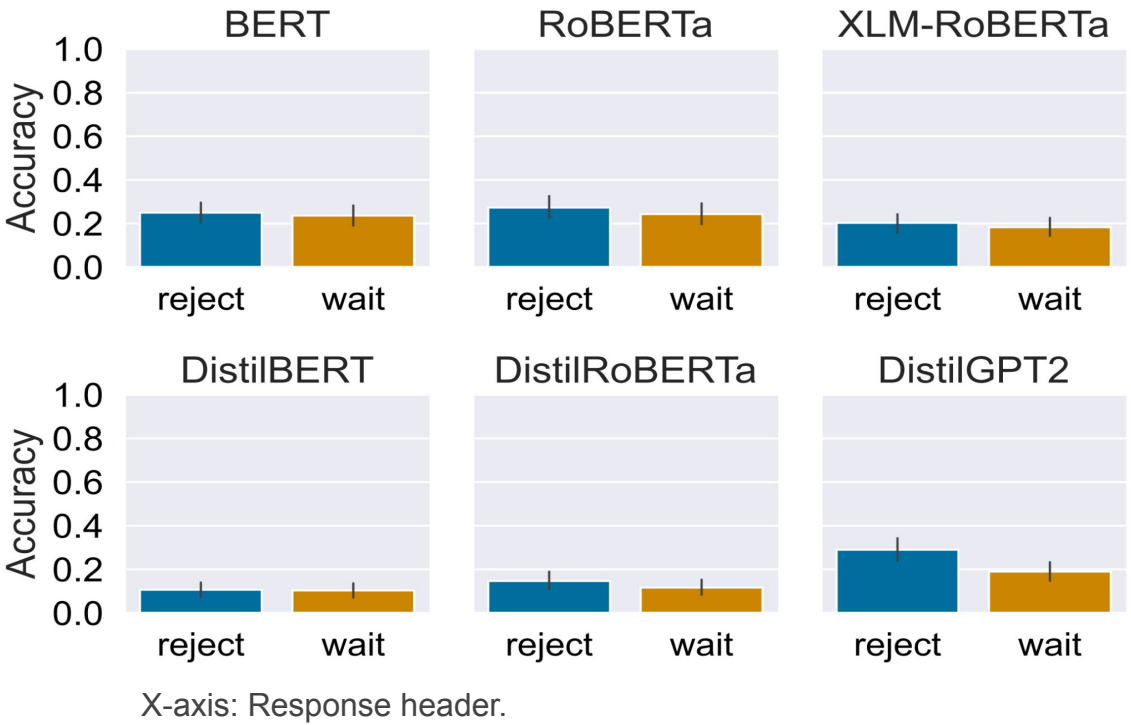
Ellipsis

- Input sequence:

  Marco said, “The nurse, who has interest in French cuisine, adopted a rescue dog,” and Ellie replied, “{No / Wait no}, he CLM: {did / does / is / was / has / would} not.”
  MLM: [MASK] not.”

- If Accurate: `AuxVerbs` for `VP1` and `VP2` are the top-2 in the model output

Ellipsis



- Sensitivity to the grammatical constraint on verb ellipsis is weak

Part 3

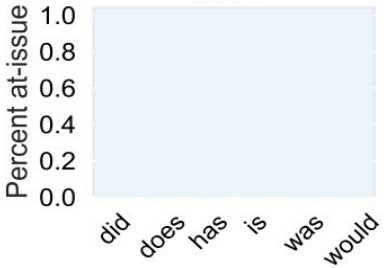
Error analysis

Error analysis in target preference task

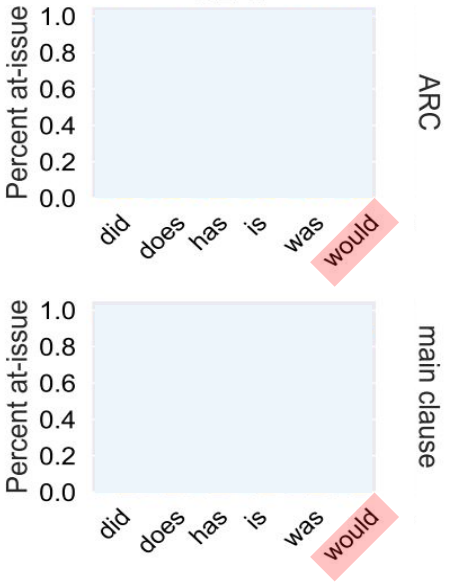
Are PLMs influenced by superficial factors on their performance in the target preference task?

Verb type effect?

Error analysis in target preference task



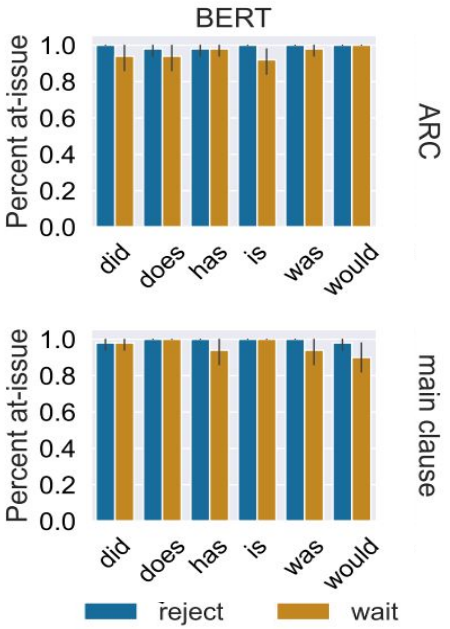
Error analysis in target preference task



*"Amaya, who **would** make pasta for dinner, was signing a song."*

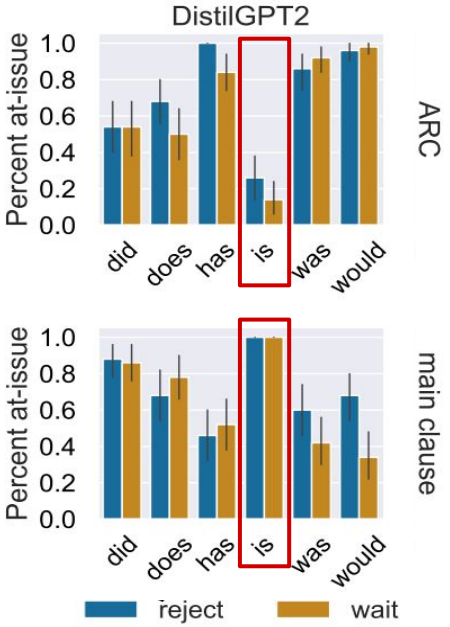
*"Wesley, who was angry, **would** have oatmeal for breakfast.*

Error analysis in target preference task



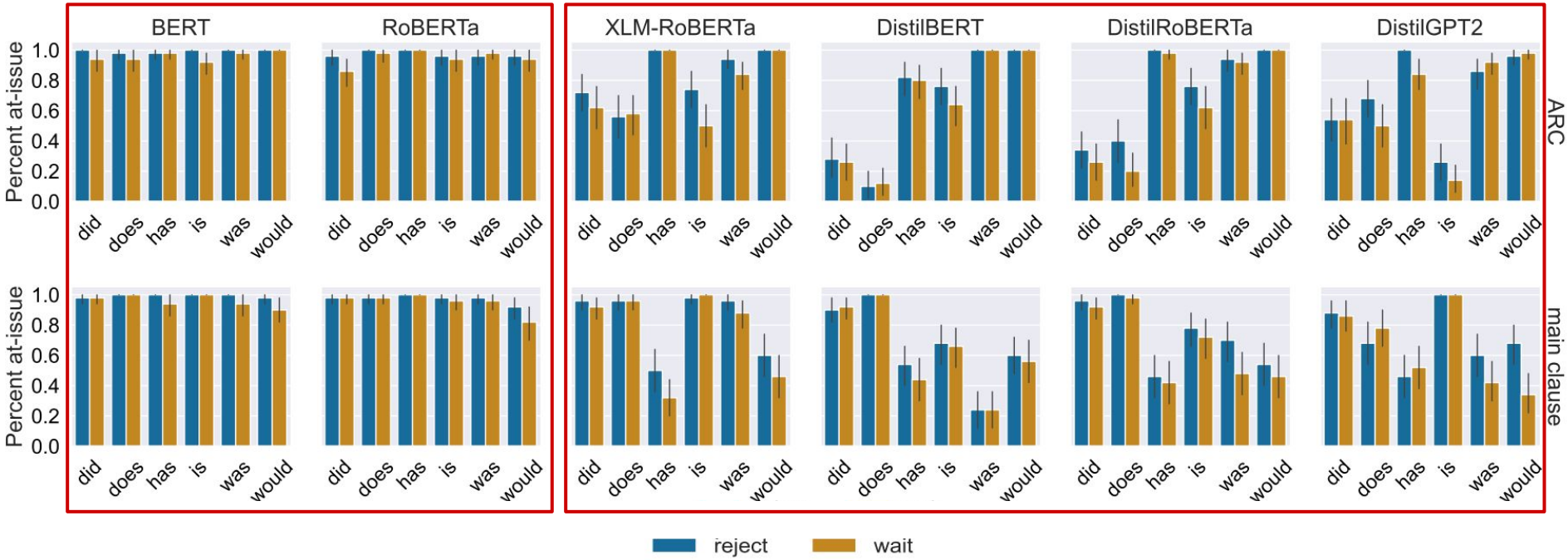
- Preference for at-issue content,
  with no major influence from verb type

Error analysis in target preference task



- Preference for at-issueness (main clause) is influenced by verb type

Error analysis in target preference task



Error analysis in target preference task

Are models influenced by superficial factors such as verb type on their sensitivity to dialogue response dynamics?

Distilled models are influenced by verb type (e.g., did, does, is)
BERT & RoBERTa aren't

Part 4

Summary & Discussion

Summary of findings

- General preference for targeting main clause
- Weak trends with capturing at-issue vs. not-at-issue contrast
- Sensitive to main vs. embedded clause distinction
- Limitation in grasping verb ellipsis
- Influence from superficial factor such as verb type

Discussion

- Discourse competence in standard PLMs is not sufficiently comprehensive
- Further understanding PLMs' sensitivity to pragmatics and discourse dynamics (e.g., [11-15])
- Current findings as foundational observations for dialogue-specific training

[11] Kurfalı & Östling. (2021). *RepL4NLP.* [12] Pilter & Nenkova. (2009). *ACL-IJCNLP.* [13] Patterson & Kehler. (2013). *EMNLP.* [14] Koto et al. (2021). *NAACL-HLT.*
[15] Pandia et al. (2021). *CoNLL.*

Acknowledgement

We are grateful to Ming Xiang, Shane Steinert-Threlkeld, Tal Linzen, Najoung Kim, and members of the UChicago CompLing Lab, for valuable comments and discussion.
We thank Kanishka Misra in particular for guidance on usage of the `minicons` library.

We also thank three anonymous reviewers for their thoughtful feedback and suggestions.

Thank you for listening!



Codes and material available at
`https://github.com/sangheek16/dialogue-response-dynamics`