How well do NLI models capture verb veridicality?

Alexis Ross and Ellie Pavlick

Harvard University *Brown University*

Work started at JSALT*

Veridicality: An Introduction

- A context is considered ***veridical*** when the propositions it contains are taken to be true, even if not explicitly asserted.

Veridicality: An Introduction

- A context is considered ***veridical*** when the propositions it contains are taken to be true, even if not explicitly asserted.

Examples: *He does not **know** that the answer is 5.* → *The answer is 5.*

Veridicality: An Introduction

- A context is considered ***veridical*** when the propositions it contains are taken to be true, even if not explicitly asserted.

Examples: *He does not **know** that the answer is 5.* → *The answer is 5.*
*He does not **think** that the answer is 5.* ↛ *The answer is 5.*

Veridicality: An Introduction

- A context is considered ***veridical*** when the propositions it contains are taken to be true, even if not explicitly asserted.

Examples: *He does not **know** that the answer is 5.* → *The answer is 5.*
*He does not **think** that the answer is 5.* ↛ *The answer is 5.*

- Understanding veridicality remains an open problem for computational models of natural language inference (Rudinger et al, 2018).

Veridicality: An Introduction

- A context is considered ***veridical*** when the propositions it contains are taken to be true, even if not explicitly asserted.

  Examples: *He does not **know** that the answer is 5.* → *The answer is 5.*
  *He does not **think** that the answer is 5.* ↛ *The answer is 5.*

- Understanding veridicality remains an open problem for computational models of natural language inference (Rudinger et al, 2018).
- We focus on **veridicality within verb-complement** constructions, for which we see past work as taking two approaches:
  - **Lexical Semantic**
  - **Speaker-Meaning**

Verb Veridicality: **Lexical Semantic** Approach

- Veridicality is a lexical semantic property of verbs.
- Each verb can be thought of as having a unique **two-bit signature**, which specifies the types of inferences it licenses–positive (+), negative (-), or neutral (◦) inferences–in positive and negative environments (Karttunen, 2012).
- These signatures apply to all contexts.

Verb Veridicality: **Lexical Semantic** Approach

- A verb having the signature −/◦ means that the *negation* of its complement projects in positive environments, and neither its complement nor its negation projects in negative environments.

(−/◦) “refuse to”

He **refused to** do the same. → ¬ He did the same.
He **did not refuse** to do the same. ↛ He did the same.

Verb Veridicality: **Speaker-Meaning** Approach

- Inferences involving veridicality rely heavily on non-lexical information and are better understood as a graded, **pragmatic** phenomenon.

Verb Veridicality: **Speaker-Meaning** Approach

- Inferences involving veridicality rely heavily on non-lexical information and are better understood as a graded, **pragmatic** phenomenon.
- Inferences may diverge from what is predicted by verb signature based on the context:

Verb Veridicality: **Speaker-Meaning** Approach

- Inferences involving veridicality rely heavily on non-lexical information and are better understood as a graded, **pragmatic** phenomenon.
- Inferences may diverge from what is predicted by verb signature based on the context:

Example: *refuse to* (−/○)

Lexical Semantic:	*He did **not refuse to** speak.*	↛	*He spoke.*
Speaker Meaning:	*He did **not refuse to** speak.*	→	*He spoke.*

Our Work

- **Question 1:** To what extent do the speaker and sentence meaning approaches diverge in their predictions of human inferences?

Our Work

- **Question 1:** To what extent do the speaker and sentence meaning approaches diverge in their predictions of human inferences?
- **Question 2:** Do neural models of natural language inference (NLI) learn to make correct inferences about verb veridicality?

Our Work

- **Question 1:** To what extent do the speaker and sentence meaning approaches diverge in their predictions of human inferences?
- **Question 2:** Do neural models of natural language inference (NLI) learn to make correct inferences about verb veridicality?
    - This work assumes the **speaker-meaning** approach: Models which consistently *mirror human inferences* about veridicality in context can be said to understand veridicality.
        - Context significantly affects veridicality judgments (de Marneffe et al., 2012)
        - NLI datasets such as SNLI and MNLI take crowdsourced approaches to inference judgments (Bowman et al., 2015; Williams et al., 2018)

Our Work: 3 Main Contributions

1. Collect a **new NLI evaluation set** of 1,500 sentence pairs involving verb-complement constructions.

Our Work: 3 Main Contributions

1. Collect a **new NLI evaluation set** of 1,500 sentence pairs involving verb-complement constructions.
2. Discuss new **analysis of human judgements of veridicality**.

Our Work: 3 Main Contributions

1. Collect a **new NLI evaluation set** of 1,500 sentence pairs involving verb-complement constructions.
2. Discuss new **analysis of human judgements of veridicality**.
3. Evaluate the state-of-the-art **BERT model** on these inferences.

(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {`*“to”*`|`*“that”*`}`

(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C> and <¬S, C>`, with the sentence (S) and complement (C).

(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C>` and `<¬S, C>`, with the sentence (S) and complement (C).
    - Original sentence: *“He knows that the answer is 5.”*
    - `<S, C>:`
        - (P): He knows that the answer is 5.
        - (H): The answer is 5.
    - `<¬S, C>:`
        - (P): He does **not** know that the answer is 5.
        - (H): The answer is 5.

(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C> and <¬S, C>`, with the sentence (S) and complement (C).
    - Original sentence: *“He knows that the answer is 5.”*
    - `<S, C>:`
        - (P): He knows that the answer is 5.
        - (H): The answer is 5.
    - `<¬S, C>:`
        - (P): He does **not** know that the answer is 5.
        - (H): The answer is 5.

(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C> and <¬S, C>`, with the sentence (S) and complement (C).
- Collect human judgments on **Amazon Mechanical Turk**: 3 raters label entailment on a 5-point scale from -2 to 2, and we take the mean for each pair.

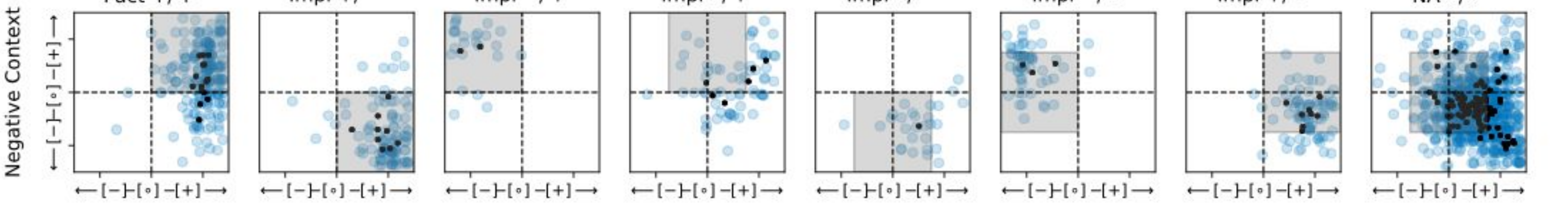
(1) Approach

- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C> and <¬S, C>`, with the sentence (S) and complement (C).
- Collect human judgments on **Amazon Mechanical Turk**: 3 raters label entailment on a 5-point scale from -2 to 2, and we take the mean for each pair.
- Resulting dataset: 1,500 unique contexts, 137 verbs, 8 signatures.

(1) Approach

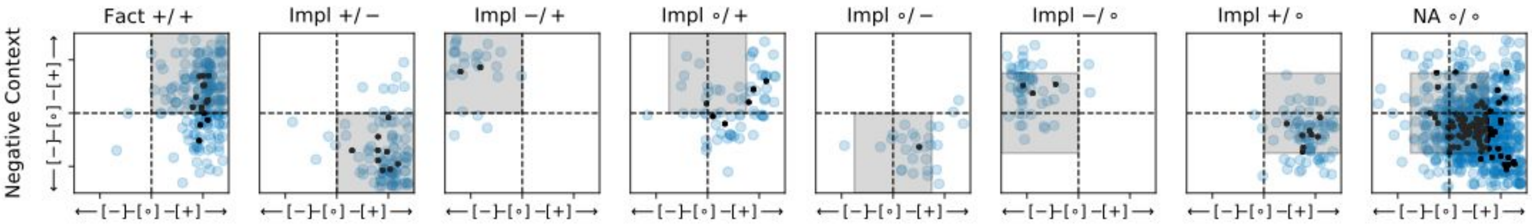
- Collect sentences from MultiNLI (Williams et al, 2018) that contain any verb-complement construction matching: `verb {“to”|“that”}`
- For each original sentence from MNLI, create **2 new premise, hypothesis pairs,** `<S, C>` `and` `<¬S, C>`, with the sentence (S) and complement (C).
- Collect human judgments on **Amazon Mechanical Turk**: 3 raters label entailment on a 5-point scale from -2 to 2, and we take the mean for each pair.
- Resulting dataset: 1,500 unique contexts, 137 verbs, 8 signatures.
- Compare human judgements with the predictions made by BERT NLI model.

(2) Analysis of Human Judgements



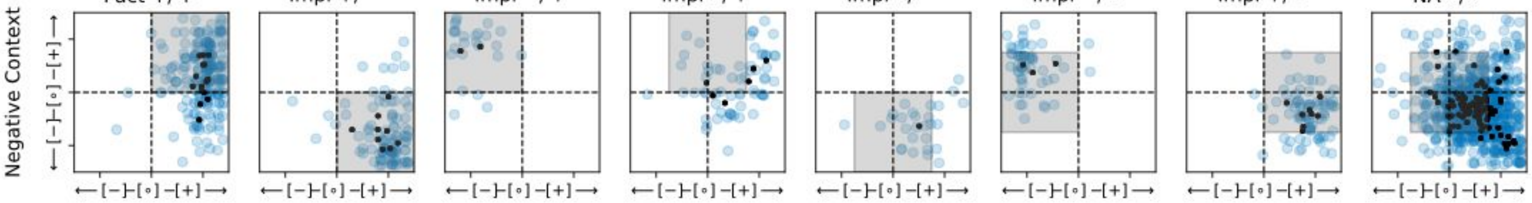
- Outliers:
  - **Factives** (Box 1): find that, reveal that, see that
  - **Implicatives** (Box 2): add that, explain that, warn that.

(2) Analysis of Human Judgements



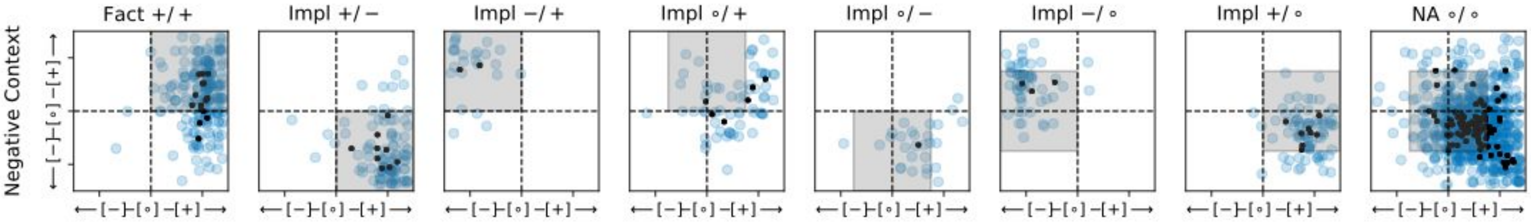
- *Averaged* across all contexts, verbs tend to behave as expected given their lexical semantic signature.

(2) Analysis of Human Judgements



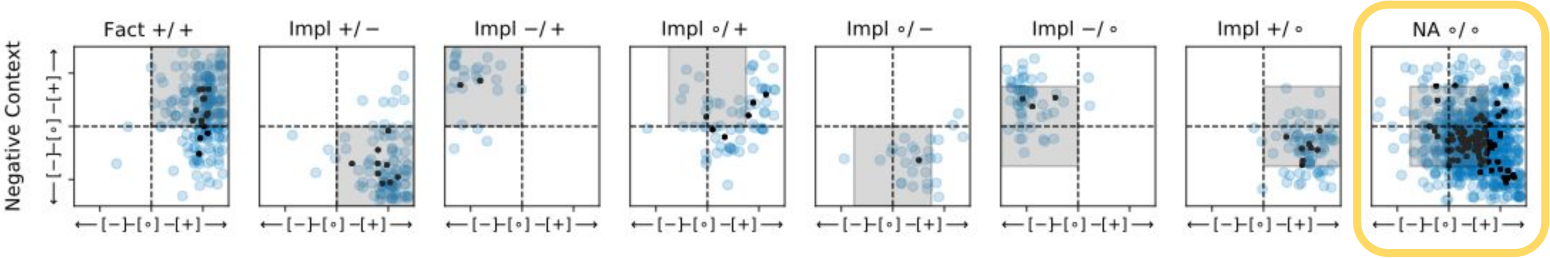
- *Averaged* across all contexts, verbs tend to behave as expected given their lexical semantic signature
- However, we observed two trends providing evidence that veridicality judgments rely heavily on **contextual** features.
  - Veridicality Bias
  - Within-Verb Variation

(2) Human Trend 1: **Veridicality Bias**



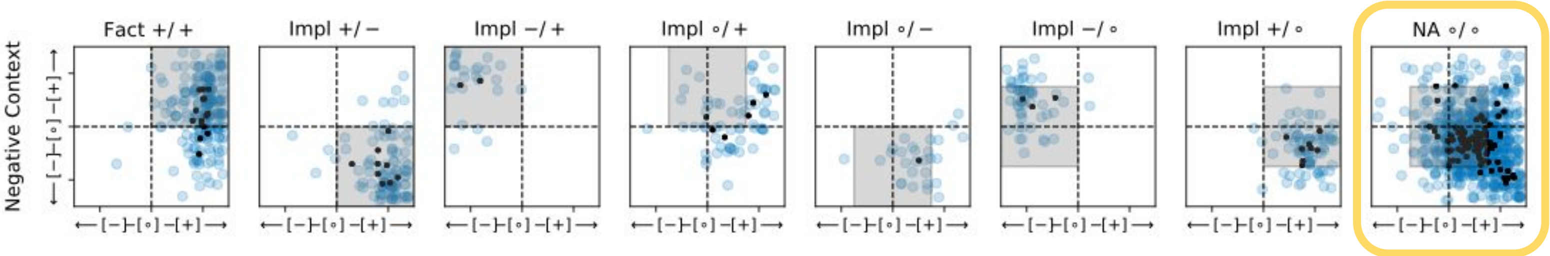
- **Veridicality Bias:** Inferences about complements are often made (positive or negative), even in environments when the expectation is that the verb is non-veridical (∘ signature).

(2) Human Trend 1: **Veridicality Bias**



- **Veridicality Bias:** Inferences about complements are often made (positive or negative), even in environments when the expectation is that the verb is non-veridical (◦ signature).

(2) Human Trend 1: **Veridicality Bias**



- **Veridicality Bias:** Inferences about complements are often made (positive or negative), even in environments when the expectation is that the verb is non-veridical (∘ signature).

- Example: Verb with (∘ / ∘) signature behaves like (+ / −)

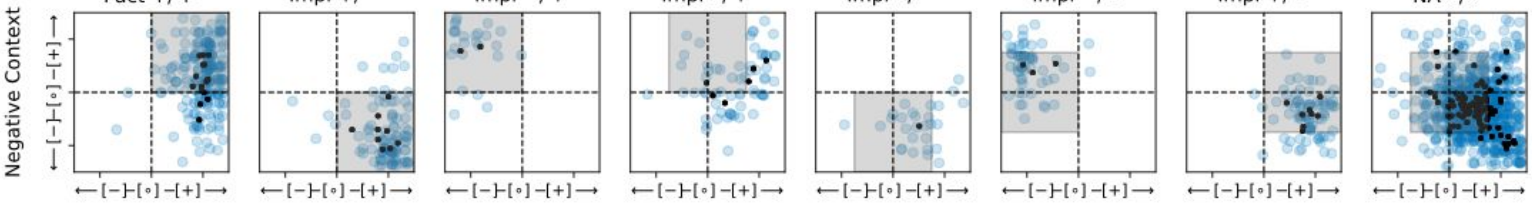
(+) (1.7) The GAO has **indicated that** it is unwilling to compromise.

→ It is unwilling to compromise.

(−) (-1.0) The GAO has **not indicated that** it is unwilling to compromise.

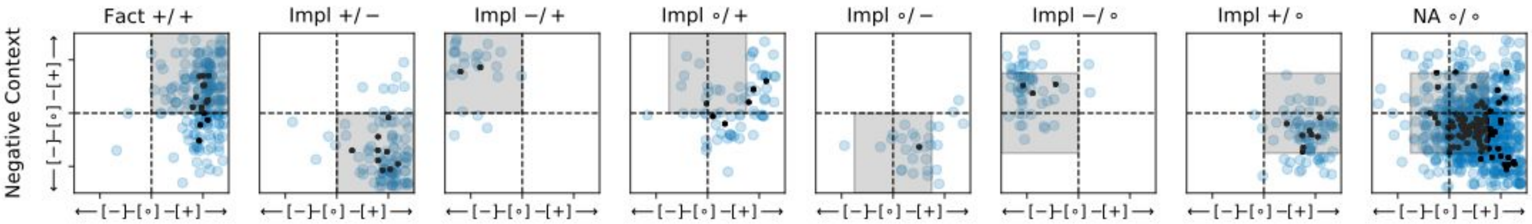
→ ¬ It is unwilling to compromise.

(2) Human Trend 2: **Within-Verb Variation**



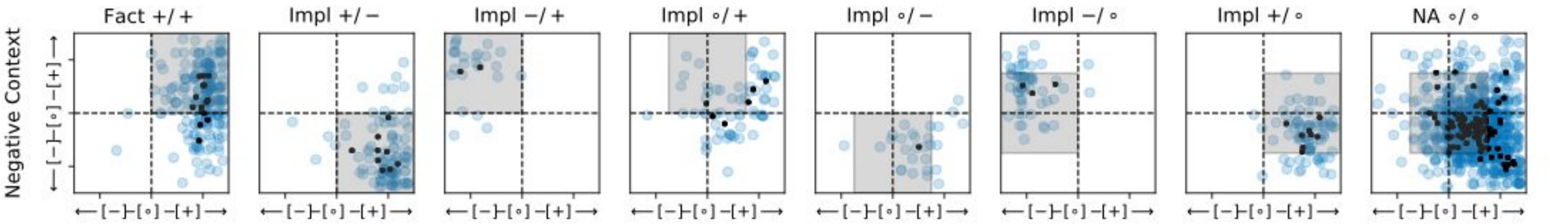
- **Within-Verb Variation:** Signatures provide a weak signal for predicting inferences in individual sentences.
- Within each signature, there is high variance across contexts, in all cases spanning at least 2 points on the -2 to 2 scale.

(2) Human Trend 2: **Within-Verb Variation**



- **Within-Verb Variation:** Signatures provide a weak signal for predicting inferences in individual sentences.
- Within each signature, there is high variance across contexts, in all cases spanning at least 2 points on the -2 to 2 scale.
  - In an ordinary least squares regression, verb signature alone explained only a small amount of this variation.

(2) Human Trend 2: **Within-Verb Variation**



- Example: Factive verb “know that” (+/+) behaves differently in contexts.

(+) (1.7) Everyone **knows that** the CPI is the most accurate.

→ The CPI is the most accurate.

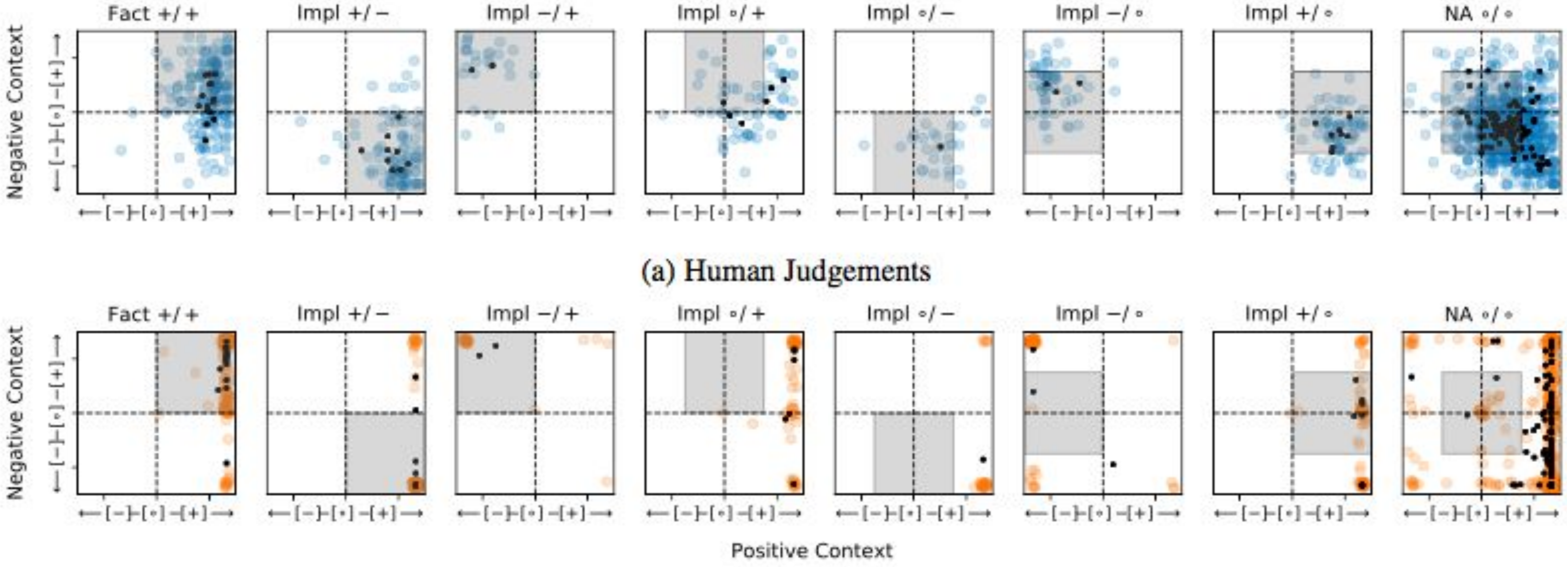
(+) (1.7) Everyone **does not know that** the CPI is the most accurate.

→ The CPI is the most accurate.

(+) (0.7) I **know that** I was born to succeed. → I was born to succeed.

(∘) (0.3) I **do not know that** I was born to succeed. ↛ I was born to succeed.

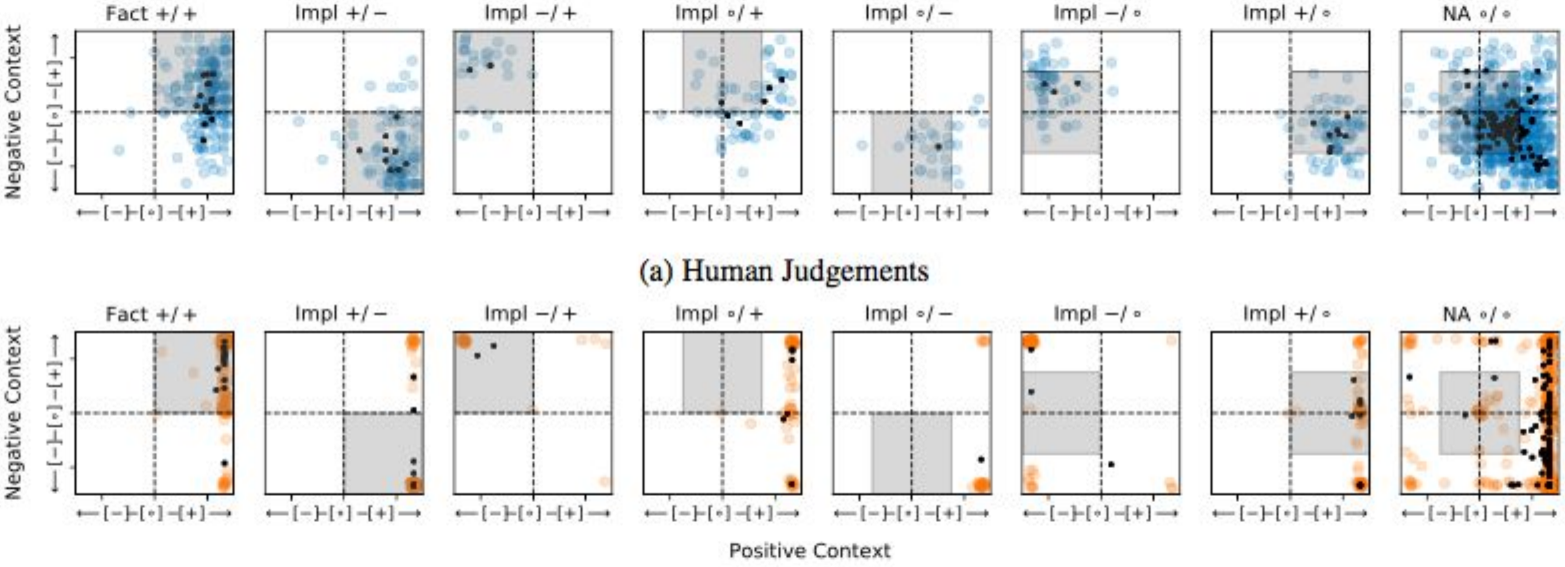
(3) BERT Predictions



(a) Human Judgements

(b) Model Predictions

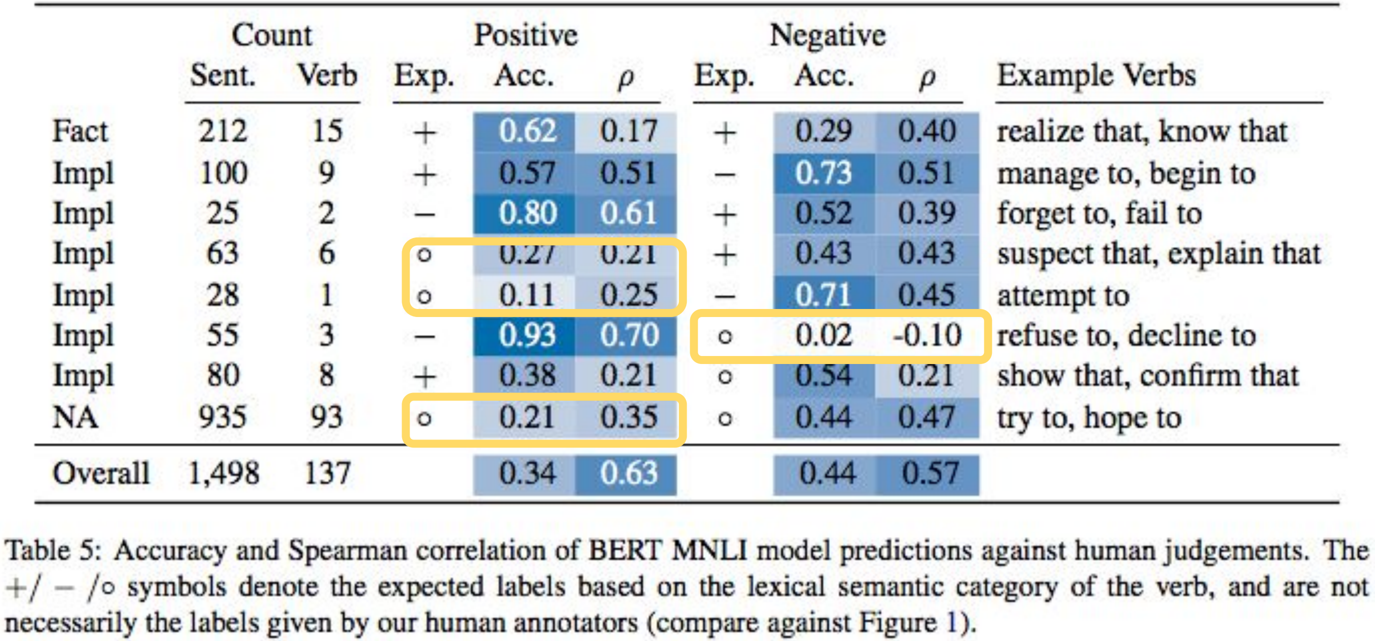
(3) BERT Predictions: **Exaggerated Veridicality Bias**



(a) Human Judgements

(b) Model Predictions

(3) BERT Predictions: **Exaggerated Veridicality Bias**



	Count		Positive			Negative			
	Sent.	Verb	Exp.	Acc.	ρ	Exp.	Acc.	ρ	Example Verbs
Fact	212	15	+	0.62	0.17	+	0.29	0.40	realize that, know that
Impl	100	9	+	0.57	0.51	−	0.73	0.51	manage to, begin to
Impl	25	2	−	0.80	0.61	+	0.52	0.39	forget to, fail to
Impl	63	6	○	0.27	0.21	+	0.43	0.43	suspect that, explain that
Impl	28	1	○	0.11	0.25	−	0.71	0.45	attempt to
Impl	55	3	−	0.93	0.70	○	0.02	-0.10	refuse to, decline to
Impl	80	8	+	0.38	0.21	○	0.54	0.21	show that, confirm that
NA	935	93	○	0.21	0.35	○	0.44	0.47	try to, hope to
Overall	1,498	137		0.34	0.63		0.44	0.57	

Table 5: Accuracy and Spearman correlation of BERT MNLI model predictions against human judgements. The +/ − /○ symbols denote the expected labels based on the lexical semantic category of the verb, and are not necessarily the labels given by our human annotators (compare against Figure 1).

(3) BERT Predictions: **Counterfactual Analysis**

- **Question:** Are the above-observed trends in BERT’s predictions driven predominantly by lexical priors (the presence of a specific verb), or are they sensitive to other aspects of a verb’s context?

(3) BERT Predictions: **Counterfactual Analysis**

- **Question:** Are the above-observed trends in BERT's predictions driven predominantly by lexical priors (the presence of a specific verb), or are they sensitive to other aspects of a verb's context?
- Replace the main verb in the **sentence** with the target verb and observe whether predictions change.

(3) BERT Predictions: **Counterfactual Analysis**

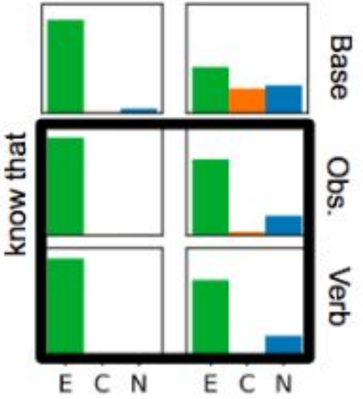
- **Question:** Are the above-observed trends in BERT's predictions driven predominantly by lexical priors (the presence of a specific verb), or are they sensitive to other aspects of a verb's context?
- Replace the main verb in the **sentence** with the target verb and observe whether predictions change.

(S) He **attempted to** overcome the sensation.
(C) He overcame the sensation.

(S*) He **tried to** overcome the sensation.
(C) He overcame the sensation.

(3) BERT Predictions: **Counterfactual Analysis**

- **Question:** Are the above-observed trends in BERT's predictions driven predominantly by lexical priors (the presence of a specific verb), or are they sensitive to other aspects of a verb's context?
- Replace the main verb in the **sentence** with the target verb and observe whether predictions change.
- **Results:** BERT's predictions are largely driven by individual verb types (i.e. "know that").



original

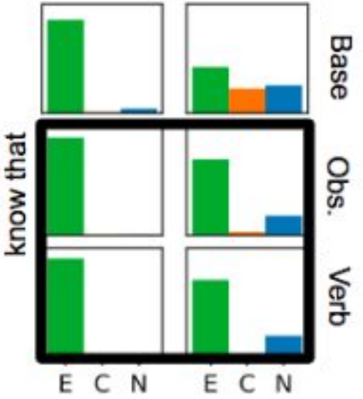
replacement

"know that"

(3) BERT Predictions: **Counterfactual Analysis**

- **Question:** Are the above-observed trends in BERT’s predictions driven predominantly by lexical priors (the presence of a specific verb), or are they sensitive to other aspects of a verb’s context?
- Replace the main verb in the **sentence** with the target verb and observe whether predictions change.
- **Results:** BERT’s predictions are largely driven by individual verb types (i.e. “know that”).

  Other Results: BERT’s predictions are sensitive to the syntactic placement of these verb types. (See paper for more information.)



original

replacement

“know that”

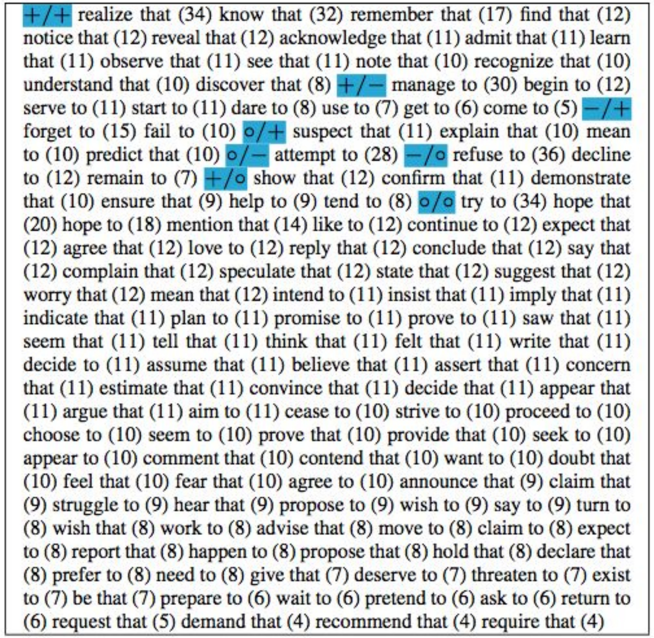
Conclusions

- Contextual factors influence human inference patterns for verb veridicality.
    - Within-Verb Variation
    - Veridicality Bias
- BERT amplifies the veridicality bias exhibited by humans.
- BERT’s predictions are driven by lexical cues (particular verbs).

Questions?

The dataset is publicly available at: `https://github.com/alexisjihyeross/verb_veridicality`

Dataset



+/+ realize that (34) know that (32) remember that (17) find that (12) notice that (12) reveal that (12) acknowledge that (11) admit that (11) learn that (11) observe that (11) see that (11) note that (10) recognize that (10) understand that (10) discover that (8) +/− manage to (30) begin to (12) serve to (11) start to (11) dare to (8) use to (7) get to (6) come to (5) −/+ forget to (15) fail to (10) ○/+ suspect that (11) explain that (10) mean to (10) predict that (10) ○/− attempt to (28) −/○ refuse to (36) decline to (12) remain to (7) +/○ show that (12) confirm that (11) demonstrate that (10) ensure that (9) help to (9) tend to (8) ○/○ try to (34) hope that (20) hope to (18) mention that (14) like to (12) continue to (12) expect that (12) agree that (12) love to (12) reply that (12) conclude that (12) say that (12) complain that (12) speculate that (12) state that (12) suggest that (12) worry that (12) mean that (12) intend to (11) insist that (11) imply that (11) indicate that (11) plan to (11) promise to (11) prove to (11) saw that (11) seem that (11) tell that (11) think that (11) felt that (11) write that (11) decide to (11) assume that (11) believe that (11) assert that (11) concern that (11) estimate that (11) convince that (11) decide that (11) appear that (11) argue that (11) aim to (11) cease to (10) strive to (10) proceed to (10) choose to (10) seem to (10) prove that (10) provide that (10) seek to (10) appear to (10) comment that (10) contend that (10) want to (10) doubt that (10) feel that (10) fear that (10) agree to (10) announce that (9) claim that (9) struggle to (9) hear that (9) propose to (9) wish to (9) say to (9) turn to (8) wish that (8) work to (8) advise that (8) move to (8) claim to (8) expect to (8) report that (8) happen to (8) propose that (8) hold that (8) declare that (8) prefer to (8) need to (8) give that (7) deserve to (7) threaten to (7) exist to (7) be that (7) prepare to (6) wait to (6) pretend to (6) ask to (6) return to (6) request that (5) demand that (4) recommend that (4) require that (4)

Table 2: 137 verbs belonging to 8 signatures. Parentheses denote number of contexts in which each verb appears in our final, annotated dataset (§4)

.