Supplementary Material:
How well do NLI models capture verb veridicality?

Alexis Ross
Harvard University
alexis_ross@college.harvard.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

A Does MNLI contain enough signal?

The fact that the BERT-based model's prediction performance is poor may be due to the fact that it has not been given sufficient training examples to be able to learn about the particular projection patterns associated with verbs. E.g. it might be representing the information needed to inform their signature, but not given enough training examples such that it knows to/how to use it.

To investigate this, we first count the raw number of times that the lemmatized verb followed by the appropriate particle appear in MNLI train (e.g. how many times does *"know that"* or *"go to"* appear in the premise or hypothesis). Then, we look for particular constructions of interest–i.e. 1) the verb lemma appears in the premise; 2) it appears as a verb (according to the parser); and 3) it appears as the root of a subtree that begins with the verb-particle construction. Table 2 gives the full counts for how often each verb appears at all and within a relevant construction within MNLI.

It is worth noting that all 137 verbs in our dataset appear at least once in MNLI train. The mean number of occurrences is 412 (σ=1,047). *"have to"* occurs the most often (7,459 times); *"advise that"* occurs the least often (1 time). Nearly all[1] verbs also occur in the setting that we care about (i.e. as the root of a verb phrase in the premise). On average, each verb occurs 212 times ($\sigma = 564$) in the desired setting, or 56% ($\mu = 0.56, \sigma = 0.25$) of the times it occurs.

Second, we perform a small annotation ourselves on the cases when the verb appears in the desired setting, to assess whether the inferences made do in fact require reasoning about the projection signature of the verb. Specifically, we take a sample of 100 p/h pairs, sampled uniformly across verb types, and have one of the authors judge whether or not each inference depends on the verbs projective behavior. We consider the inference to "depend on" the projective behavior if the truth of the complement clause in the premise is relevant to the hypothesis. Table 1 gives examples from MNLI train in which this is and is not the case.

In our sample of 100, we find that 21 of the inferences do in fact require making some assumption about projectivity.[2] We marked 18 examples as "not applicable" since the verb did not actually appear in the construction we wanted it to (i.e. due to faulty parses, mostly from the telephone domain). In the remaining 61 examples, the verb construction was present but not related to the inference for which the model was being trained.

Thus, a very back-of-the-envelop computation would suggest that, on average, 21% of the on average 212 relevant occurrences of each verb type–i.e. around 40 p/h pairs per verb type–could be useful training examples for the phenomenon we study here. Our analysis suggests that this amount of training is sufficient for the model to learn non-trivial aspects of projective behavior of these verbs, but not to "master" such inferences. Future work will be required to determined what the minimum number/proportion of examples is that is needed for currently models to learn the correct behavior. How many examples should be required is a normative question.

B Annotation Guidelines

We are researchers from the `Institution`, trying to help computers understand language! Understanding language allows computers to do use-

[1] Except for *"advise that"*, which appears as the root of a verb phrase in the hypothesis

[2] Note this does not necessarily mean that the judgment which yields the correct label is consistent with formal semantic analysis of the verb's signature, and in fact it often is not.

Requires reasoning about veridicality
p: Orrin Hatch has to acknowledge that Lee is well qualified for the post he seeks. h: Lee is completely unqualified for the post.
p: Rubin agnostics will be glad to learn that there are two weak spots in his record. h: Rubin has a very strong record
Does not require reasoning about veridicality
p : But after being forewarned, if it happens again, I'll be hard to convince that you were in the right. h: I won't believe you if it happens again.
p: Finally, CBO did not attempt to price the relocation of personnel to a central location. h: CBO didn't want to relocate any of their personnel.

Table 1: Examples of MNLI train examples that do (top two) and do not (bottom two) require reasoning about veridicality.

ful things like answer our questions or summarize news articles for us. While common sense reasoning is easy for people, it is very very hard for computers. Please help us by using your common sense to decide if sentences are more likely to be true or false.

Note: Many of these sentences have been automatically generated and some many be difficult to interpret. Please select the ”does not make sense” option if you feel you cannot make any reasonable inference of what the sentence is supposed to mean. You will never be punished for choosing this option.

Instructions. For each pair of sentences, assume that the first sentence is true, describes a real scenario, or expresses an opinion. Using your best judgement, indicate how likely it is, on a scale of 1 to 5, that the second sentence is also true, describes the same scenario, or expresses the same opinion.

1. Your answers should be based only on information which is stated or implied by the first sentence. Even if the second sentence seems like it is reasonably true in general, you should only choose 4 or 5 if the truth of the second sentence can be inferred entirely from the first sentence. E.g. for the sentence pair below, you should choose 3 since the first sentence alone provides us no way of knowing that Greece is in Europe (even though our world knowledge tells us that the second sentence is true).

Example 1
Greece is a country.
Greece is a European country.
CORRECT ANSWER: 3, not necessarily true or necessarily false.

2. The order of the sentences is important. For example, when the same sentences as above are provided in the opposite order, the correct answer changes to 5: definitely true.

Example 2
Greece is a European country.
Greece is a country.
CORRECT ANSWER: 5, definitely true.

3. It is okay to make reasonable assumptions. If the first sentence makes it highly unlikely that the second is true, indicate so by choosing option 1 or 2. Try to interpret the sentences as you would if you heard them in a real life conversation. It is okay to make reasonable assumptions that you believe most people would make. E.g. it is okay to assume that a bomb is not a fake bomb in the below context.

Example 3
The terrorists were collecting materials to build a bomb.
The terrorists were collecting materials to build a fake bomb.
CORRECT ANSWER: 1, definitely NOT true. or 2, probably NOT true.

Example 4
The man failed to submit the report on time
The man submitted the report on time
CORRECT ANSWER: 1, definitely NOT true. or 2, probably NOT true.

4. When in doubt, you should err on the side of uncertainty. If the context of the first sentence makes it reasonably possible for the second to be either true or false, choose 3.

Example 5

Secretary Clinton is the expected nominee.

Secretary Clinton is the nominee.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

Example 6

The police have arrested a suspected murderer.

The police have arrested a murderer.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

5. Take the entire sentence into account. Remember that we are interested in the information communicated by the sentence as a whole. Inserting the same words might lead to different answers in different contexts.

Example 7 She is a potential candidate for the Senate.

She is a candidate for the Senate.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

Example 8

They are talking about potential candidates for the Senate.

They are talking about candidates for the Senate.

CORRECT ANSWER: 4, probably true or 5, definitely true.

Keep in mind, we are predominantly interested in understanding whether the second sentence communicates the same information as the first, or if it adds or removes important information.

C Verbs Excluded by Annotation Filters

See Table 4 for number of contexts excluded per verb type.

To measure inter-rater agreement, for each example and each of the three raters assigned to the example, we calculated the correlation between that rater's score and the averaged score of the other two raters. The Spearman correlation among raters, averaged across the three raters for each example, was 0.78 for positive contexts and 0.74 for negative contexts.

D Counterfactual Results by Verb

See below tables (Table 5) for KL divergence between baseline and target distributions that result from substitutions involving each verb type.

have to	7444	0.64	want to	6947	0.48	need to	5294	0.31
like to	2808	0.48	think that	2702	0.61	know that	2093	0.73
try to	1897	0.61	come to	1652	0.45	say that	1451	0.44
get to	1229	0.49	believe that	1156	0.59	seem to	1065	0.78
tend to	915	0.61	plan to	838	0.13	ensure that	672	0.92
continue to	626	0.71	see that	621	0.69	show that	613	0.20
attempt to	612	0.14	mean that	571	0.36	love to	509	0.43
feel that	502	0.55	state that	464	0.03	find that	447	0.72
return to	375	0.32	claim that	363	0.20	work to	347	0.22
wish to	331	0.53	report that	328	0.06	agree that	307	0.57
prefer to	304	0.46	note that	301	0.36	refuse to	299	0.15
begin to	292	0.59	happen to	287	0.70	help to	284	0.30
mean to	277	0.38	suggest that	272	0.83	felt that	269	0.52
hope that	269	0.43	choose to	263	0.60	be that	259	0.83
understand that	251	0.79	saw that	238	0.66	realize that	233	0.77
claim to	230	0.17	argue that	219	0.75	use to	219	0.51
seek to	209	0.37	appear to	207	0.78	assume that	206	0.73
hope to	206	0.47	start to	189	0.44	remember that	185	0.78
move to	181	0.36	fail to	177	0.69	expect to	175	0.63
decide to	163	0.63	turn to	152	0.53	estimate that	139	0.22
doubt that	135	0.11	intend to	124	0.60	indicate that	117	0.87
hear that	116	0.62	wish that	113	0.32	say to	112	0.54
admit that	111	0.76	recognize that	108	0.86	notice that	105	0.71
learn to	102	0.74	manage to	100	0.72	struggle to	95	0.03
aim to	93	0.16	conclude that	88	0.78	wait to	87	0.41
prove that	86	0.49	concern that	83	0.00	agree to	74	0.65
promise to	73	0.25	require that	71	0.76	fear that	69	0.39
worry that	64	0.47	demand that	62	0.19	expect that	59	0.71
suspect that	56	0.64	pretend to	55	0.44	learn that	54	0.72
forget to	52	0.63	deserve to	51	0.29	request that	50	0.48
hold that	49	0.18	decline to	48	0.31	prove to	48	0.50
mention that	45	0.47	tell that	43	0.21	demonstrate that	41	0.88
give that	40	0.60	imply that	40	0.90	comment that	39	0.00
insist that	39	0.77	strive to	38	0.63	confirm that	37	0.49
explain that	37	0.81	discover that	33	0.73	serve to	33	0.73
recommend that	28	0.86	ask to	27	0.67	seem that	27	0.78
dare to	26	0.46	proceed to	24	1.12	add that	23	0.39
complain that	23	0.52	propose to	23	0.91	reveal that	23	0.78
determine that	22	0.82	predict that	21	0.71	contend that	20	0.75
observe that	20	0.90	reply that	19	0.63	assert that	17	0.88
cease to	17	0.18	provide that	17	1.06	remain to	16	0.94
declare that	14	0.43	threaten to	14	0.64	write that	14	0.86
decide that	13	0.46	exist to	13	0.69	acknowledge that	12	0.50
announce that	11	0.82	appear that	10	0.60	prepare to	10	0.90
propose that	8	0.75	speculate that	8	1.50	convince that	3	1.00
warn that	3	1.00	advise that	1	0.00			

Table 2: For each verb construction: 1) how many times did it appear at all in MNLI? and 2) in what percentage did it appear in a format like that we test for, i.e. as the root of a verb phrase involving the relevant complement type?

	Correlation		Overlap	
	Pos.	Neg.	Pos.	Neg.
$+/+$	0.52	0.7	0.17	0.11
$+/-$	0.55	0.69	0.16	0.15
$-/+$	0.47	0.61	0.18	0.12
$\circ/+$	0.81	0.67	0.15	0.12
$\circ/-$	0.72	0.54	0.14	0.11
$-/\circ$	0.49	0.53	0.13	0.11
$+/\circ$	0.46	0.64	0.15	0.13
$\circ/\circ$	0.73	0.66	0.15	0.13

Table 3: Inter-rater reliability by verb signature. For each signature, the following metrics were calculated: 1) correlation (Spearman), 2) absolute overlap. Correlation was calculated between the score of each rater (3 per example) and the averaged scores of the other two raters. Absolute overlap was calculated for each pair of raters. These metrics were averaged across examples and raters.

Excluded: allow to (0/12), ask that (0/12), take to (0/12), suppose to (0/12), require to (0/12), relate to (0/10), refer to (0/12), leave to (0/12), lead to (0/12), know to (0/12), fund to (0/10), force to (0/12), do to (0/12), determine to (0/12), design to (0/12), consider to (0/12), believe to (0/12), think to (0/12). **Rest:** attempt to (28/40), manage to (30/40), know that (32/40), recommend that (4/12), demand that (4/12), require that (4/12), come to (5/12), request that (5/12), forget to (15/22), get to (6/12), ask to (6/12), hope to (18/24), realize that (34/40), try to (34/40), prepare to (6/12), pretend to (6/12), wait to (6/12), give that (7/12), dare to (8/13), return to (6/11), threaten to (7/12), exist to (7/12), remain to (7/12), be that (7/12), use to (7/12), deserve to (7/12), report that (8/12), hope that (20/24), hold that (8/12), happen to (8/12), refuse to (36/40), move to (8/12), need to (8/12), prefer to (8/12), expect to (8/12), discover that (8/12), propose that (8/12), tend to (8/12), turn to (8/12), claim to (8/12), work to (8/12), declare that (8/12), wish that (8/12), announce that (9/12), ensure that (9/12), claim that (9/12), propose to (9/12), help to (9/12), say to (9/12), advise that (8/11), wish to (9/12), have to (9/12), struggle to (9/12), choose to (10/12), add that (10/12), cease to (10/12), demonstrate that (10/12), learn to (10/12), understand that (10/12), appear to (10/12), mean to (10/12), want to (10/12), agree to (10/12), hear that (9/11), seem to (10/12), provide that (10/12), predict that (10/12), recognize that (10/12), contend that (10/12), strive to (10/12), feel that (10/12), fear that (10/12), fail to (10/12), explain that (10/12), seek to (10/12), doubt that (10/12), prove that (10/12), note that (10/12), prove to (11/12), think that (11/12), remember that (17/18), tell that (11/12), suspect that (11/12), start to (11/12), saw that (11/12), serve to (11/12), see that (11/12), seem that (11/12), acknowledge that (11/12), learn that (11/12), promise to (11/12), felt that (11/12), admit that (11/12), aim to (11/12), appear that (11/12), argue that (11/12), assert that (11/12), assume that (11/12), believe that (11/12), concern that (11/12), confirm that (11/12), convince that (11/12), decide that (11/12), decide to (11/12), estimate that (11/12), write that (11/12), insist that (11/12), plan to (11/12), intend to (11/12), observe that (11/12), imply that (11/12), indicate that (11/12), mention that (14/15), say that (12/12), begin to (12/12), love to (12/12), mean that (12/12), notice that (12/12), comment that (10/10), warn that (12/12), agree that (12/12), worry that (12/12), like to (12/12), conclude that (12/12), complain that (12/12), reply that (12/12), find that (12/12), suggest that (12/12), continue to (12/12), state that (12/12), speculate that (12/12), proceed to (10/10), decline to (12/12), determine that (12/12), reveal that (12/12), expect that (12/12), show that (12/12)

Table 4: The effects of our exclusion criteria by verb. Parentheses denote the number of contexts included post exclusion out of the total number of contexts we obtained Turk annotations for.

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
admit that	$D^*\|\|D_{vt}$	0.03	0.29	0.13	1.57
(+/+)	$D^*\|\|D$	0.02	0.10	0.00	0.06
notice that	$D^*\|\|D_{vt}$	0.01	0.03	0.12	0.22
(+/+)	$D^*\|\|D$	0.03	0.37	0.00	0.04
observe that	$D^*\|\|D_{vt}$	0.01	0.36	0.08	1.28
(+/+)	$D^*\|\|D$	0.03	0.13	0.00	0.07
reveal that	$D^*\|\|D_{vt}$	0.03	0.22	0.17	1.92
(+/+)	$D^*\|\|D$	0.03	0.13	0.00	0.06
see that	$D^*\|\|D_{vt}$	0.02	0.14	0.11	0.02
(+/+)	$D^*\|\|D$	0.02	0.00	0.00	0.05
remember that	$D^*\|\|D_{vt}$	0.03	0.65	0.15	1.31
(+/+)	$D^*\|\|D$	0.02	0.06	0.00	0.04
learn that	$D^*\|\|D_{vt}$	0.01	0.79	0.15	2.05
(+/+)	$D^*\|\|D$	0.03	0.30	0.00	0.05
understand that	$D^*\|\|D_{vt}$	0.02	0.03	0.14	1.84
(+/+)	$D^*\|\|D$	0.03	0.42	0.01	0.03
find that	$D^*\|\|D_{vt}$	0.03	1.46	0.16	3.10
(+/+)	$D^*\|\|D$	0.02	0.12	0.00	0.04
note that	$D^*\|\|D_{vt}$	0.01	0.36	0.06	1.20
(+/+)	$D^*\|\|D$	0.03	0.12	0.01	0.10
know that	$D^*\|\|D_{vt}$	0.03	0.13	0.20	1.77
(+/+)	$D^*\|\|D$	0.02	0.24	0.00	0.02
recognize that	$D^*\|\|D_{vt}$	0.02	0.15	0.13	1.53
(+/+)	$D^*\|\|D$	0.03	0.26	0.00	0.06
realize that	$D^*\|\|D_{vt}$	0.03	0.07	0.20	1.98
(+/+)	$D^*\|\|D$	0.03	0.42	0.00	0.03
acknowledge that	$D^*\|\|D_{vt}$	0.05	0.06	0.04	0.20
(+/+)	$D^*\|\|D$	0.02	0.12	0.00	0.06
discover that	$D^*\|\|D_{vt}$	0.09	0.44	0.05	1.64
(+/+)	$D^*\|\|D$	0.02	0.07	0.00	0.06
start to	$D^*\|\|D_{vt}$	0.00	0.42	0.60	0.86
(+/−)	$D^*\|\|D$	0.16	0.05	0.00	0.00
get to	$D^*\|\|D_{vt}$	0.00	0.19	0.66	0.91
(+/−)	$D^*\|\|D$	0.16	0.09	0.00	0.00
manage to	$D^*\|\|D_{vt}$	0.00	0.05	0.80	0.03
(+/−)	$D^*\|\|D$	0.16	0.10	0.00	0.00
begin to	$D^*\|\|D_{vt}$	0.00	0.01	0.60	0.08
(+/−)	$D^*\|\|D$	0.16	0.03	0.00	0.00
use to	$D^*\|\|D_{vt}$	0.01	0.40	0.50	0.96
(+/−)	$D^*\|\|D$	0.12	0.04	0.00	0.00
come to	$D^*\|\|D_{vt}$	0.00	0.13	0.52	0.99
(+/−)	$D^*\|\|D$	0.16	0.12	0.00	0.00
dare to	$D^*\|\|D_{vt}$	0.07	1.00	0.49	0.68
(+/−)	$D^*\|\|D$	0.08	0.07	0.00	0.00
serve to	$D^*\|\|D_{vt}$	0.00	0.27	0.56	1.12
(+/−)	$D^*\|\|D$	0.16	0.08	0.00	0.00
learn to	$D^*\|\|D_{vt}$	0.00	0.20	0.69	0.34
(+/−)	$D^*\|\|D$	0.16	0.01	0.00	0.01
fail to	$D^*\|\|D_{vt}$	0.00	0.01	1.67	1.36
(−/+)	$D^*\|\|D$	1.37	1.21	0.01	0.00
forget to	$D^*\|\|D_{vt}$	0.06	0.11	1.08	5.46
(−/+)	$D^*\|\|D$	0.92	1.36	0.01	0.01
mean to	$D^*\|\|D_{vt}$	0.03	1.04	0.45	1.05
(○/+)	$D^*\|\|D$	0.15	0.00	0.00	0.00
predict that	$D^*\|\|D_{vt}$	0.00	3.00	0.05	3.59
(○/+)	$D^*\|\|D$	0.01	0.04	0.00	0.08
explain that	$D^*\|\|D_{vt}$	0.01	0.06	0.12	1.43
(○/+)	$D^*\|\|D$	0.03	0.31	0.00	0.07
add that	$D^*\|\|D_{vt}$	0.01	0.03	0.09	1.16
(○/+)	$D^*\|\|D$	0.03	0.38	0.00	0.07
warn that	$D^*\|\|D_{vt}$	0.07	0.08	0.10	1.15
(○/+)	$D^*\|\|D$	0.01	0.22	0.00	0.08
suspect that	$D^*\|\|D_{vt}$	0.71	0.04	0.07	0.19
(○/+)	$D^*\|\|D$	0.10	0.09	0.01	0.03

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
attempt to	$D^*\|\|D_{vt}$	0.25	0.07	0.68	0.01
(○/−)	$D^*\|\|D$	0.07	0.09	0.00	0.00
decline to	$D^*\|\|D_{vt}$	0.54	0.03	6.26	0.33
(−/○)	$D^*\|\|D$	1.32	0.55	0.03	0.01
refuse to	$D^*\|\|D_{vt}$	0.38	0.03	7.57	1.44
(−/○)	$D^*\|\|D$	1.52	0.96	0.05	0.01
remain to	$D^*\|\|D_{vt}$	0.50	0.01	0.16	0.06
(−/○)	$D^*\|\|D$	0.12	0.05	0.00	0.01
help to	$D^*\|\|D_{vt}$	0.01	1.46	0.60	0.83
(+/○)	$D^*\|\|D$	0.12	0.02	0.00	0.00
have to	$D^*\|\|D_{vt}$	0.01	0.19	0.66	1.26
(+/○)	$D^*\|\|D$	0.14	0.10	0.00	0.01
tend to	$D^*\|\|D_{vt}$	0.00	0.20	0.54	0.90
(+/○)	$D^*\|\|D$	0.16	0.10	0.00	0.00
confirm that	$D^*\|\|D_{vt}$	0.05	0.17	0.05	2.05
(+/○)	$D^*\|\|D$	0.02	0.16	0.00	0.07
demonstrate that	$D^*\|\|D_{vt}$	0.04	0.04	0.10	0.05
(+/○)	$D^*\|\|D$	0.02	0.05	0.00	0.03
show that	$D^*\|\|D_{vt}$	0.03	0.35	0.11	0.90
(+/○)	$D^*\|\|D$	0.02	0.05	0.00	0.04
determine that	$D^*\|\|D_{vt}$	0.05	0.87	0.09	2.67
(+/○)	$D^*\|\|D$	0.01	0.11	0.01	0.07
ensure that	$D^*\|\|D_{vt}$	0.05	0.44	0.06	0.87
(+/○)	$D^*\|\|D$	0.01	0.01	0.01	0.08
move to	$D^*\|\|D_{vt}$	0.01	0.10	0.54	0.12
(○/○)	$D^*\|\|D$	0.13	0.02	0.00	0.00
intend to	$D^*\|\|D_{vt}$	0.01	0.34	0.55	0.95
(○/○)	$D^*\|\|D$	0.15	0.06	0.00	0.00
seek to	$D^*\|\|D_{vt}$	0.35	0.16	0.51	0.99
(○/○)	$D^*\|\|D$	0.06	0.12	0.00	0.00
agree to	$D^*\|\|D_{vt}$	0.01	0.00	0.63	0.00
(○/○)	$D^*\|\|D$	0.14	0.00	0.00	0.00
expect to	$D^*\|\|D_{vt}$	0.13	0.01	0.56	0.00
(○/○)	$D^*\|\|D$	0.08	0.00	0.00	0.00
happen to	$D^*\|\|D_{vt}$	0.00	0.03	0.52	0.06
(○/○)	$D^*\|\|D$	0.16	0.12	0.00	0.00
claim to	$D^*\|\|D_{vt}$	0.02	0.03	0.54	0.04
(○/○)	$D^*\|\|D$	0.15	0.12	0.00	0.00
wish to	$D^*\|\|D_{vt}$	0.21	0.07	0.01	0.01
(○/○)	$D^*\|\|D$	0.06	0.08	0.00	0.00
wait to	$D^*\|\|D_{vt}$	0.55	5.45	0.39	1.32
(○/○)	$D^*\|\|D$	0.02	1.05	0.00	0.02
love to	$D^*\|\|D_{vt}$	0.02	0.01	0.62	0.18
(○/○)	$D^*\|\|D$	0.12	0.25	0.00	0.00
work to	$D^*\|\|D_{vt}$	0.02	0.35	0.53	0.03
(○/○)	$D^*\|\|D$	0.13	0.14	0.00	0.00
say to	$D^*\|\|D_{vt}$	0.02	0.00	0.52	0.05
(○/○)	$D^*\|\|D$	0.15	0.03	0.00	0.00
prefer to	$D^*\|\|D_{vt}$	0.11	0.04	0.51	0.15
(○/○)	$D^*\|\|D$	0.06	0.04	0.00	0.00
hope to	$D^*\|\|D_{vt}$	0.07	0.31	0.05	1.31
(○/○)	$D^*\|\|D$	0.16	0.08	0.00	0.00
prove to	$D^*\|\|D_{vt}$	0.00	0.01	0.77	0.04
(○/○)	$D^*\|\|D$	0.16	0.12	0.01	0.01
seem to	$D^*\|\|D_{vt}$	0.00	0.18	0.57	1.14
(○/○)	$D^*\|\|D$	0.16	0.11	0.01	0.00
try to	$D^*\|\|D_{vt}$	0.42	0.26	0.66	1.20
(○/○)	$D^*\|\|D$	0.06	0.11	0.00	0.00
continue to	$D^*\|\|D_{vt}$	0.00	0.00	0.58	0.12
(○/○)	$D^*\|\|D$	0.16	0.08	0.00	0.00
need to	$D^*\|\|D_{vt}$	0.24	0.27	0.24	0.94
(○/○)	$D^*\|\|D$	0.06	0.07	0.00	0.01
want to	$D^*\|\|D_{vt}$	0.18	0.47	0.03	0.97
(○/○)	$D^*\|\|D$	0.04	0.04	0.00	0.00

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
like to	$D^*\|D_{vt}$	0.08	0.01	0.59	0.01
(◦/◦)	$D^*\|D$	0.11	0.00	0.00	0.00
appear to	$D^*\|D_{vt}$	0.00	0.17	0.58	0.95
(◦/◦)	$D^*\|D$	0.16	0.11	0.00	0.00
plan to	$D^*\|D_{vt}$	0.04	0.01	0.56	0.07
(◦/◦)	$D^*\|D$	0.14	0.10	0.00	0.00
choose to	$D^*\|D_{vt}$	0.00	0.20	0.63	0.81
(◦/◦)	$D^*\|D$	0.16	0.10	0.00	0.00
decide to	$D^*\|D_{vt}$	0.00	0.20	0.74	0.81
(◦/◦)	$D^*\|D$	0.16	0.10	0.01	0.01
prepare to	$D^*\|D_{vt}$	0.03	0.00	0.47	0.02
(◦/◦)	$D^*\|D$	0.14	0.02	0.00	0.00
pretend to	$D^*\|D_{vt}$	1.02	0.04	0.55	0.01
(◦/◦)	$D^*\|D$	0.04	0.07	0.00	0.00
proceed to	$D^*\|D_{vt}$	0.00	0.19	0.55	1.11
(◦/◦)	$D^*\|D$	0.16	0.11	0.00	0.00
return to	$D^*\|D_{vt}$	0.00	0.10	0.59	0.03
(◦/◦)	$D^*\|D$	0.16	0.03	0.00	0.00
cease to	$D^*\|D_{vt}$	0.33	0.19	0.17	4.72
(◦/◦)	$D^*\|D$	1.07	1.19	0.01	0.00
propose to	$D^*\|D_{vt}$	0.11	0.17	0.50	0.88
(◦/◦)	$D^*\|D$	0.14	0.11	0.00	0.00
deserve to	$D^*\|D_{vt}$	0.15	0.00	0.55	0.09
(◦/◦)	$D^*\|D$	0.07	0.13	0.00	0.00
exist to	$D^*\|D_{vt}$	0.01	0.46	0.57	1.03
(◦/◦)	$D^*\|D$	0.13	0.03	0.00	0.00
aim to	$D^*\|D_{vt}$	0.03	0.23	0.54	1.05
(◦/◦)	$D^*\|D$	0.14	0.09	0.00	0.00
strive to	$D^*\|D_{vt}$	0.09	0.01	0.48	0.06
(◦/◦)	$D^*\|D$	0.11	0.09	0.01	0.00
threaten to	$D^*\|D_{vt}$	0.05	0.41	0.10	1.07
(◦/◦)	$D^*\|D$	0.18	0.04	0.00	0.00
promise to	$D^*\|D_{vt}$	0.03	0.00	0.54	0.10
(◦/◦)	$D^*\|D$	0.15	0.04	0.00	0.00
ask to	$D^*\|D_{vt}$	0.43	0.02	0.42	0.04
(◦/◦)	$D^*\|D$	0.22	0.02	0.00	0.01
turn to	$D^*\|D_{vt}$	0.00	0.04	0.57	0.00
(◦/◦)	$D^*\|D$	0.16	0.03	0.00	0.00
struggle to	$D^*\|D_{vt}$	2.04	0.10	5.62	0.60
(◦/◦)	$D^*\|D$	0.55	0.35	0.01	0.01
reply that	$D^*\|D_{vt}$	0.02	0.04	0.09	1.46
(◦/◦)	$D^*\|D$	0.03	0.37	0.00	0.04
hope that	$D^*\|D_{vt}$	0.01	0.15	0.36	0.42
(◦/◦)	$D^*\|D$	0.39	0.05	0.02	0.06
decide that	$D^*\|D_{vt}$	0.04	0.02	0.08	0.41
(◦/◦)	$D^*\|D$	0.02	0.08	0.01	0.06
imply that	$D^*\|D_{vt}$	0.01	0.01	0.05	0.42
(◦/◦)	$D^*\|D$	0.06	0.13	0.01	0.06
declare that	$D^*\|D_{vt}$	0.02	1.10	0.08	2.97
(◦/◦)	$D^*\|D$	0.02	0.12	0.00	0.09
prove that	$D^*\|D_{vt}$	0.06	0.02	0.03	0.27
(◦/◦)	$D^*\|D$	0.02	0.05	0.00	0.02
assert that	$D^*\|D_{vt}$	0.05	0.04	0.09	0.45
(◦/◦)	$D^*\|D$	0.01	0.10	0.00	0.07
feel that	$D^*\|D_{vt}$	0.05	0.06	0.06	0.85
(◦/◦)	$D^*\|D$	0.01	0.38	0.01	0.03
expect that	$D^*\|D_{vt}$	0.30	0.11	0.10	0.15
(◦/◦)	$D^*\|D$	0.01	0.03	0.00	0.04
insist that	$D^*\|D_{vt}$	0.09	0.37	0.11	0.22
(◦/◦)	$D^*\|D$	0.02	0.06	0.01	0.09
announce that	$D^*\|D_{vt}$	0.03	0.03	0.08	0.24
(◦/◦)	$D^*\|D$	0.02	0.05	0.01	0.08
worry that	$D^*\|D_{vt}$	0.03	0.34	0.12	0.18
(◦/◦)	$D^*\|D$	0.19	0.41	0.01	0.05

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
hold that	$D^*\|D_{vt}$	0.08	0.00	0.06	0.11
(◦/◦)	$D^*\|D$	0.01	0.05	0.00	0.05
complain that	$D^*\|D_{vt}$	0.02	0.02	0.05	0.16
(◦/◦)	$D^*\|D$	0.00	0.38	0.01	0.06
propose that	$D^*\|D_{vt}$	0.37	0.94	0.07	3.25
(◦/◦)	$D^*\|D$	0.03	0.21	0.00	0.06
request that	$D^*\|D_{vt}$	0.69	2.15	0.04	0.91
(◦/◦)	$D^*\|D$	0.13	0.03	0.01	0.07
speculate that	$D^*\|D_{vt}$	0.34	0.46	0.04	1.21
(◦/◦)	$D^*\|D$	0.48	0.13	0.01	0.09
recommend that	$D^*\|D_{vt}$	0.21	1.16	0.04	0.79
(◦/◦)	$D^*\|D$	0.02	0.04	0.01	0.05
seem that	$D^*\|D_{vt}$	0.08	1.95	0.08	4.69
(◦/◦)	$D^*\|D$	0.00	0.27	0.00	0.04
demand that	$D^*\|D_{vt}$	0.39	1.81	0.03	0.75
(◦/◦)	$D^*\|D$	0.05	0.02	0.01	0.06
saw that	$D^*\|D_{vt}$	0.02	1.09	0.11	1.53
(◦/◦)	$D^*\|D$	0.02	0.00	0.00	0.05
state that	$D^*\|D_{vt}$	0.05	0.94	0.05	3.75
(◦/◦)	$D^*\|D$	0.02	0.17	0.01	0.10
argue that	$D^*\|D_{vt}$	0.17	0.28	0.06	1.39
(◦/◦)	$D^*\|D$	0.01	0.11	0.01	0.09
indicate that	$D^*\|D_{vt}$	0.05	0.47	0.05	3.09
(◦/◦)	$D^*\|D$	0.01	0.33	0.01	0.08
report that	$D^*\|D_{vt}$	0.03	2.09	0.07	3.22
(◦/◦)	$D^*\|D$	0.02	0.04	0.01	0.06
suggest that	$D^*\|D_{vt}$	0.04	0.49	0.04	3.41
(◦/◦)	$D^*\|D$	0.07	0.45	0.01	0.07
think that	$D^*\|D_{vt}$	0.28	0.56	0.07	3.43
(◦/◦)	$D^*\|D$	0.01	0.48	0.01	0.05
believe that	$D^*\|D_{vt}$	0.00	0.30	0.05	0.68
(◦/◦)	$D^*\|D$	0.01	0.06	0.00	0.03
be that	$D^*\|D_{vt}$	0.01	3.84	0.12	3.82
(◦/◦)	$D^*\|D$	0.03	0.03	0.00	0.00
say that	$D^*\|D_{vt}$	0.06	0.03	0.07	0.87
(◦/◦)	$D^*\|D$	0.01	0.30	0.01	0.10
agree that	$D^*\|D_{vt}$	0.05	0.38	0.04	1.50
(◦/◦)	$D^*\|D$	0.02	0.10	0.00	0.05
felt that	$D^*\|D_{vt}$	0.07	0.68	0.09	3.38
(◦/◦)	$D^*\|D$	0.01	0.38	0.01	0.03
conclude that	$D^*\|D_{vt}$	0.04	0.03	0.11	0.29
(◦/◦)	$D^*\|D$	0.02	0.06	0.00	0.04
claim that	$D^*\|D_{vt}$	0.17	0.04	0.07	0.72
(◦/◦)	$D^*\|D$	0.00	0.18	0.01	0.07
mean that	$D^*\|D_{vt}$	0.03	0.09	0.06	0.00
(◦/◦)	$D^*\|D$	0.02	0.04	0.01	0.05
assume that	$D^*\|D_{vt}$	0.05	0.06	0.05	0.07
(◦/◦)	$D^*\|D$	0.13	0.05	0.01	0.08
require that	$D^*\|D_{vt}$	0.18	0.91	0.04	0.83
(◦/◦)	$D^*\|D$	0.01	0.03	0.01	0.01
estimate that	$D^*\|D_{vt}$	0.10	0.04	0.07	0.48
(◦/◦)	$D^*\|D$	0.01	0.07	0.01	0.09
comment that	$D^*\|D_{vt}$	0.03	0.07	0.07	1.36
(◦/◦)	$D^*\|D$	0.02	0.24	0.01	0.07
advise that	$D^*\|D_{vt}$	0.13	0.14	0.07	0.11
(◦/◦)	$D^*\|D$	0.00	0.05	0.00	0.07
doubt that	$D^*\|D_{vt}$	0.02	0.01	1.07	1.68
(◦/◦)	$D^*\|D$	1.36	0.69	0.01	0.08
give that	$D^*\|D_{vt}$	0.01	1.27	0.07	1.77
(◦/◦)	$D^*\|D$	0.03	0.09	0.01	0.05
concern that	$D^*\|D_{vt}$	0.59	0.05	0.06	0.02
(◦/◦)	$D^*\|D$	0.07	0.17	0.01	0.06
convince that	$D^*\|D_{vt}$	0.12	0.10	0.17	0.45
(◦/◦)	$D^*\|D$	0.00	0.12	0.00	0.03

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
fear that	$D^* \| D_{vt}$	0.12	0.31	0.79	1.21
(○/○)	$D^* \| D$	0.41	0.28	0.01	0.06
write that	$D^* \| D_{vt}$	0.07	0.31	0.03	1.41
(○/○)	$D^* \| D$	0.03	0.12	0.00	0.07
tell that	$D^* \| D_{vt}$	0.02	0.18	0.09	1.47
(○/○)	$D^* \| D$	0.03	0.14	0.01	0.07
hear that	$D^* \| D_{vt}$	0.02	0.93	0.11	2.03
(○/○)	$D^* \| D$	0.02	0.24	0.01	0.05
contend that	$D^* \| D_{vt}$	0.15	0.25	0.07	0.56
(○/○)	$D^* \| D$	0.00	0.10	0.01	0.10
appear that	$D^* \| D_{vt}$	0.05	1.16	0.09	1.37
(○/○)	$D^* \| D$	0.01	0.28	0.00	0.01
provide that	$D^* \| D_{vt}$	0.08	1.69	0.07	0.81
(○/○)	$D^* \| D$	0.01	0.01	0.01	0.04
wish that	$D^* \| D_{vt}$	0.22	0.14	1.21	0.22
(○/○)	$D^* \| D$	0.91	0.01	0.00	0.03
mention that	$D^* \| D_{vt}$	0.06	0.14	0.06	1.64
(○/○)	$D^* \| D$	0.01	0.18	0.01	0.10

Table 5: Comparison (KL divergence) of post-manipulation prediction distribution to target verb distribution (D_{vt}, top row) and baseline distribution (D, bottom row). High similarity to D_{vt} suggests the model changed its predictions in response to the manipulation.