

# Detecting Stance in Media on Global Warming

Yiwei Luo<sup>1</sup> Dallas Card<sup>2</sup> Dan Jurafsky<sup>1,2</sup>

Stanford University

<sup>1</sup>Department of Linguistics <sup>2</sup>Department of Computer Science

{yiweil, dcard, jurafsky}@stanford.edu

## Abstract

Citing opinions is a powerful yet understudied strategy in argumentation. For example, an environmental activist might say, “Leading scientists agree that global warming is a serious concern,” framing a clause which affirms their own stance (*that global warming is serious*) as an opinion endorsed (*[scientists] agree*) by a reputable source (*leading*). In contrast, a global warming denier might frame the same clause as the opinion of an *untrustworthy* source with a predicate connoting *doubt*: “*Mistaken* scientists *claim* [...]” Our work studies opinion-framing in the global warming (GW) debate,<sup>1</sup> an increasingly partisan issue that has received little attention in NLP. We introduce **Global Warming Stance Dataset (GWSD)**, a dataset of stance-labeled GW sentences, and train a BERT classifier to study novel aspects of argumentation in how different sides of a debate represent their own and each other’s opinions. From 56K news articles, we find that similar linguistic devices for self-affirming and opponent-doubting discourse are used across GW-accepting and skeptic media, though GW-skeptical media shows more opponent-doubt. We also find that authors often characterize sources as hypocritical, by ascribing opinions expressing the author’s *own* view to source entities known to publicly endorse the *opposing* view. We release our stance dataset, model, and lexicons of framing devices for future work on opinion-framing and the automatic detection of GW stance.

## 1 Introduction

Ascribing opinions to other people is a powerful yet understudied strategy in argumentation.

<sup>1</sup>Throughout, we use the term *debate* to refer to the existence of contrasting opinions about GW expressed in the media; it is important to emphasize that there is virtually 100% consensus among scientists regarding the reality of anthropogenic global warming (Powell, 2017).

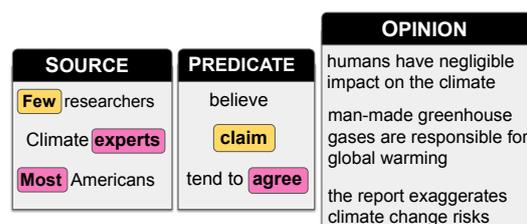


Figure 1. Examples of SOURCE, PREDICATE, and OPINION components, and within components, examples of **affirming** and **doubting** framing devices.

For example, an environmental activist might say, “*Leading* scientists *agree* that global warming is serious,” whereas a global warming denier could say, “*Mistaken* scientists *claim* that global warming is serious.” In both these examples, the embedded clause (*that global warming is serious*) is presented as an opinion belonging to a source entity (*scientists*). However, differences in the choice of predicate (*agree* vs. *claim*) and in how the source is described lead to very different interpretations. We henceforth refer to the use of such [ENTITY] [EXPRESS] [STATEMENT] sentences as **opinion-framing**, and to the respective components as the SOURCE, PREDICATE, and OPINION (see Fig. 1).

Despite its pervasiveness in argumentative discourse, opinion-framing is understudied as a persuasive strategy. This paper studies opinion-framing in the media coverage of global warming (GW), an increasingly partisan issue in the United States (Pew Research Center, 2020) that has received little attention in NLP despite its real world urgency. We focus on acts of opinion-framing representing *self-affirming* and *opponent-doubting* discourses, i.e., discourse affirming one’s own OPINIONS (embedded clauses ascribed to a SOURCE, as depicted in Fig. 1) and discourse casting doubt on

the other side’s. Studying such discourses requires a way to identify the stance of a given OPINION with respect to GW, but this is a challenging task.

To this end, we introduce **GWSD: Global Warming Stance Dataset**, a dataset for detecting and analyzing GW stance in text. We collect human judgments of GW stance for 2K sentences with Amazon Mechanical Turk (AMT)<sup>2</sup> and use our dataset to train a BERT-based classifier that achieves 75% accuracy (competitive with human performance) for GW stance detection. Extending prior work in NLP and linguistics, we develop lexicons of affirming and doubting framing devices with respect to the PREDICATE that embeds the OPINION (e.g., *know* vs. *claim*) and the SOURCE to which the opinion is ascribed (e.g., *a peer-reviewed study* vs. *a misleading paper*) (see Fig. 1).

We then apply our model and lexicons to study two questions about opinion-framing in argumentation: **Q1**: Do different sides of a debate (in this case, GW-accepting and GW-skeptical media) show symmetry in their use of self-affirming and opponent-doubting discourse? We might expect some similarities (e.g., the use of *agree* to frame OPINIONS expressing one’s own side’s stance, or the use of *claim* to cast doubt on OPINIONS from the opposing side), but given inherent asymmetries in the nature of the GW debate, it is not clear whether such strategies will be found across sides to equal extents.

Second, since opinion-framing is a way of putting words into someone’s mouth, we also ask **Q2**: In cases where OPINIONS are ascribed to a named entity with a known (public) stance, does the stance of the OPINION match the expected stance of the named entity?

Applying our model to a set of 500K OPINIONS (**Op<sub>full</sub>**) extracted from 56K GW articles, we find that GW-skeptical media engages in comparatively more opponent-doubt, though both sides of the debate show more self-affirmation overall, and use similar sets of framing devices for each respective discourse type. We also find that opinion-framing does indeed ascribe OPINIONS differing from the overt views of entities to those entities nonetheless, as part of a rhetorical strategy of ascribing hypocrisy: authors portray their *own* OPINION as

being held (in private) by figures who endorse the *opposite* OPINION (in public).

Our contributions are the following:

1. **GWSD**, a dataset of 2K sentences from GW news with annotations for stance.
2. A weighted extension of BERT competitive with human performance for classifying the stance of a sentence with respect to GW.
3. Lexicons of affirming and doubting PREDICATES (e.g., *know*, *claim*) and SOURCE modifiers (e.g., *peer-reviewed*, *misleading*).
4. Analyses on a set of 500K opinions from GW news to illustrate the utility of our dataset and lexicons for studying opinion-framing.

We release our dataset, model, and lexicons as part of this paper.<sup>3</sup>

## 2 Related work

Our work is related to social psychology research on persuasion (Cialdini, 1993; Orji et al., 2015) and recent NLP research on argumentation, such as predicting argument convincingness (Habernal and Gurevych, 2016; Simpson and Gurevych, 2018) and studying discourse-level and non-linguistic features predictive of persuasion (Yang and Kraut, 2017; Zhang et al., 2016). The latter’s work on self- vs. opponent-coverage is particularly relevant to the GW debate and we apply a similar categorization to the stance of ascribed opinions.

Also relevant is the literature on factuality and speaker commitment (de Marneffe et al., 2011; Soni et al., 2014; Werner et al., 2015; Rudinger et al., 2018; Jiang and de Marneffe, 2019), and relatedly, work studying how words can express subjectivity or bias (Riloff and Wiebe, 2003; Recasens et al., 2013; Pryzant et al., 2020). Our current paper builds upon previous work by examining such triggers as opinion-framing devices in an argumentation context, where biases related to people’s prior beliefs may interact with the lexical effects of these words.

Opinion-framing can be thought of as a special case of the broader phenomenon of framing as discussed in the communications and political science literatures (Entman, 2006; Lakoff and Ferguson, 2006; Chong and Druckman, 2007), as well as in NLP (Tsur et al., 2015; Field et al., 2018; Roy and

<sup>2</sup>We also experimented with tweets from GW-activists/skeptics and headlines from extreme conservative/liberal outlets as potential sources of softly stance-labeled sentences, but found that classifiers trained on these data perform poorly on news discourse.

<sup>3</sup><https://github.com/yiweiluo/GWStance>

Goldwasser, 2020). Both phenomena serve to emphasize particular aspects of an issue, and are often used with the intent to influence perception of that issue. Our attention to the component of SOURCE in instances of opinion-framing is also informed by communications research on the *messenger effect* (that people’s perceptions of a message may depend heavily on the message source) (Bolsen et al., 2019; Myrick and Evans Comfort, 2020; Fielding et al., 2020; Esposito et al., 2013). Furthermore, our interest in predicates of opinion attribution is inspired by communications studies examining how the choice of predicate (*say vs. assert*) can encode journalist stance (Caldas-Coulthard, 2002) and bias audience perception of the quoted entity (Gidengil and Everitt, 2003). Finally, our dataset contribution builds on Mohammad et al. (2016), who created the first climate change stance task and dataset.

### 3 GWSD: A dataset for GW stance

To enable our study of opinion-framing, and to facilitate further work on stance, we create a new publicly-available dataset of OPINION spans extracted from GW news articles (described in §3.1) that we have annotated with stance judgements using AMT (§3.2). To investigate potential annotator biases, we study the impact of annotator characteristics on their perception of stance (with approval from our Institutional Review Board) (§3.3), and combine ratings so as to infer a distribution over stance labels for each span while accounting for bias (§3.4), which we release along with the raw annotations.

#### 3.1 Extracting sentences for the dataset

Our base dataset consists of OPINION spans extracted from 56K GW news articles, published from Jan. 1, 2000 to April 12, 2020 by 63 U.S. news sources. We collected these articles using the MediaCloud API<sup>4</sup> and SerpAPI.<sup>5</sup> The keywords we used for API requests were: {climate change, global warming, fossil fuels, carbon dioxide, methane, co2}. We note that some of the articles in our dataset come from newswires (N=1.3K), but as we show later, including wire articles does not affect our studies’ conclusions. Moreover, since it is ultimately up to media outlets to decide which wire articles to publish, we believe that instances of

<sup>4</sup><https://cyber.harvard.edu/research/mediacloud>

<sup>5</sup><https://serpapi.com/search-api>

Left-leaning outlets		Right-leaning outlets	
NYT	6K	Breitbart	2.7K
Moth. Jones	3.2K	Fox	2.6K
WaPo	2K	Forbes	2K
CS Monitor	1.9K	Wash. Times	1.4K
The Nation	1.4K	Daily Caller	1.2K
Vox	1.4K	Newsmax	1.2K
Dem. Now	1K	Wash. Exam.	1K
<i>Total</i>	20K	<i>Total</i>	36K

Table 1. Number of unique articles from the top 7 left-leaning and right-leaning media outlets in our dataset (LL and RL), by volume of articles contributed. We categorize political leaning using the Media Bias/Fact Check project.

opinion-framing from wire articles are still reflective of what an outlet endorses (despite not originating from the outlet). We also include op-ed articles in our dataset, as their exclusion is made challenging by idiosyncrasy in their coding across outlets. Future work might exclude op-ed articles for model training and analysis. Please refer to Appendix A for details on our filtering and de-duplication steps. Tab. 1 and Fig. 2 summarize the distribution of articles by source.

To identify the rhetorical components of relevant sentences, we make use of syntactic dependency parsing to extract embedded OPINION spans (e.g., *Scientists believe that [climate change requires immediate action]*) from a given article, as well as spans for SOURCE (who or what the OPINION is ascribed to) and PREDICATE (the verb that syntactically embeds the OPINION). Note that we exclude OPINIONS under the scope of negation or modals.

Our pipeline consists of first passing each article through the spaCy pre-processing pipeline with a neural coreference resolution add-on,<sup>6</sup> then extracting and annotating instances of SOURCE, PREDICATE and OPINION using a rule-based algorithm (please refer to Appendix B). To validate our algorithm, we manually annotated 25 articles and compared results. We found that a dependency parsing-based approach has a high recall, identifying all clausal complements including some false positives such as indirect questions and subjunctive clauses. We therefore used several lexical resources to filter the extracted clauses to indicative

<sup>6</sup><https://github.com/huggingface/neuralcoref>, which implements the model from Clark and Manning (2015).

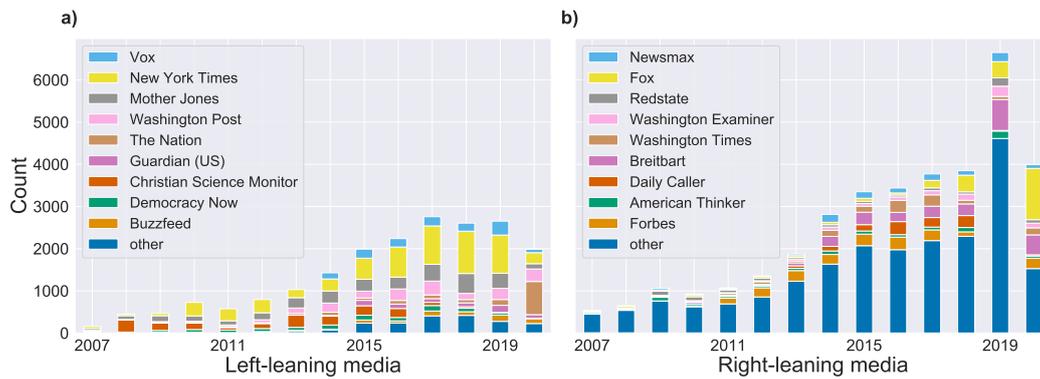


Figure 2. Number of GW articles in our dataset from 2007 to 2020 in a) Left-, b) Right-leaning media.

statements.

Finally, since many of the OPINIONS that we extracted are not explicitly on the topic of GW, we only keep the OPINION spans that contain a stem from a set of 73 manually curated keywords (e.g., *climat, environ, temperatur*).

### 3.2 Crowd-sourcing labels for the dataset

We used AMT to label a subset of 2,050 OPINION spans containing high-precision keywords (see Appendix C). The set of 2,050 spans was constructed iteratively by randomly sampling, then manually filtering spans containing potentially upsetting material (e.g., mocking Greta Thunberg’s disability) or that were off-topic (e.g., used “climate” in the sense of a workplace environment). For each OPINION, we collected judgements as to whether it expresses the target opinion: “Climate change/global warming is a serious concern,” with the potential labels being “agree,” “neutral,” or “disagree.”

Following 4 pilot studies, we decided to collect 8 judgements per item (to enable robust analysis of demographic variation in annotator judgements), for a total of 16,400 annotations, paying the California minimum wage of \$12USD per hour. Using typical exclusion criteria, we recruited a set of 398 qualified annotators over 5 rounds and had them rate 30-50 items. We also asked for basic demographic information and their personal opinions on a series of questions related to GW (see Appendix D for details and an example).

Although stance datasets are typically created with the notion of a “true” label for each item, we note that there is some degree of inherent ambiguity in this task due to the complex nature of the GW debate as well as the items’ being taken out of context. The average inter-annotator agreement (IAA) measured as Krippendorff’s alpha ranged

from 0.54 to 0.64 over the 5 rounds of annotation, though the vast majority of disagreements were between adjacent labels. Some items with high disagreement are shown in Tab. 2, showing the possibility of genuine ambiguity in GW stance.

### 3.3 Demographic effects on annotation

Given that GW has become a polarized issue in the US, we test whether we observe any bias related to party affiliation in stance annotation. Past work has called attention to the importance of considering demographic biases in annotation (Cowan and Khatchadourian, 2003; Sap et al., 2019). Intuitively, we might expect that those skeptical of GW would be more likely to perceive a sentence as exaggerating its threat, and therefore more likely to classify the sentence as one that suggests that GW is a serious concern (even though they themselves may disagree).

In order to test for the presence of demographic bias, we make use of Bayesian hierarchical ordinal regression models to estimate the effect of various annotator characteristics, such as party affiliation (Gelman and Hill, 2007), which we fit using Stan (Carpenter et al., 2017). Because we have 8 annotations per item and 30-50 annotations from each annotator, we model variation in both items and worker biases, with the latter drawn from a hierarchical prior incorporating annotator characteristics (please see Appendix E for details).

As expected, we do find clear evidence of a slight bias along party lines. For a typical OPINION, (self-identified) Republicans are approximately 1.05 ( $\pm 0.016$  s.d.) times more likely to label an item as “agree” compared to non-Republicans, and similarly less likely to respond with “disagree.” We see the opposite trend for Democrats, though the effect of the latter is mitigated by the inclusion of addi-

- 
1. Global warming is inevitably going to be, at best, managed.
  2. Global warning will be over-ridden by this effect, giving humankind and the Earth 30 years to sort out our pollution.
  3. The global warming debate is over.
  5. Global warming would open stretches of the Arctic Ocean to shipping and drilling.
- 

Table 2. Examples of items eliciting the highest disagreement among annotators (measured as entropy over labels). Each of these items was annotated with all 3 labels – “agree,” “neutral,” and “disagree.” The stance of these items seems to depend not only on the linguistic content present but also on who the speaker might be, or what the statement is said in response to, making them difficult to label.

tional covariates. More surprisingly, we also find a slight gender bias, with those who self-identify as female being 1.04 times more likely to respond with “agree” ( $\pm 0.011$  s.d.). This effect is robust to the inclusion of other variables, but should be interpreted with caution, as women were somewhat underrepresented in our study (see Tab. 7 in Appendix E for full modeling results). Regardless, this reinforces the importance of taking potential annotator biases into account (Cowan and Khatchadourian, 2003; Sap et al., 2019) and is suggestive for further research.

### 3.4 Aggregating annotations

Because some workers are more reliable than others, we again make use of Bayesian modeling to aggregate the annotations for each item. Drawing inspiration from MACE (Hovy et al., 2013), we fit a model which includes a distribution over labels associated with each item (i.e., agree, neutral, disagree), corresponding biases for each annotator, and a parameter indicating the degree to which they are influenced by their own biases. Whereas MACE assumes that annotators sometimes choose labels at random on individual instances, but otherwise identify the true label, we assume that annotators are always somewhat influenced by their biases, but to differing degrees. This model allows us to simultaneously infer a distribution over labels for each instance (i.e., the probability of each label being chosen by a typical worker), as well as bias and vigilance terms for each annotator. (Please see Appendix F for full model details). Based on this model, we assign the highest probability label to each OPINION, as summarized in Table 3.

## 4 A model for GW stance classification

In order to classify stance in  $Op_{full}$ , the full dataset of 500K OPINIONS, we train a model using the set of 2K annotated examples. The goal of this task is to predict the stance of a sentence  $S$  toward the

Label	Count
neutral	873
agree	777
disagree	400

Table 3. Distribution of labels in **GWSD**, as aggregated by our model when the label with highest inferred probability is selected.

target opinion  $T$  (“Climate change/global warming is a serious concern”). To evaluate performance, we first select a random test set of 200 annotated instances (stratified by label and political leaning of the source media outlet) and use 5-fold cross validation to train on the remaining 1850 examples.

Here, we report on variations on a BERT classifier (Devlin et al., 2019), as well as a linear baseline, in order to provide a sense of relative performance in comparison to past work. To ensure comparison against a strong baseline, we perform a grid search over hyperparameters for both approaches, and choose the best model from each according to validation accuracy, evaluating only the best model of each type on the held-out test set.

For our neural model, we use the general-purpose BERT<sub>base</sub> architecture, trained by minimizing cross-entropy loss. We use the Transformers library<sup>7</sup> as the basis for the models that we develop and compare. As potential augmentations, we experiment with a) fine-tuning the base model as a language model to unlabeled data; b) including the text of the target opinion as an input to the model; and c) using label weights as opposed to simply using the most probable label. For the weighted version, we include a copy of each training instance with each label, along with an instance weight corresponding to the label probability estimated by our label aggregation model above. (Full details of hyperparameter tuning in Appendix H).

<sup>7</sup><https://huggingface.co/transformers/>

The test-set performances of best models we obtain are shown in Table 4, along with majority class and human performance (see Appendix F). The best performing BERT model used weighted data and incorporated the target opinion as an input, but was not fine-tuned as a language model. The accuracy of this model is competitive with human performance (estimated using leave-one-out subsets of 10% of annotators), and mis-classifications of “agree” as “disagree” or vice versa occurred in less than 9% of test examples.

Further inspection of the validation results reveals that training on the weighted data offers a statistically significant improvement on validation accuracy, but the expected performance is statistically indistinguishable with respect to fine-tuning and/or incorporating the target opinion as an input. The best linear model was a simple  $l_2$ -weighted logistic regression classifier using unigrams and bigrams (details in Appendix H).

	acc	$F_A$	$F_N$	$F_D$	$F_{avg}$
<i>Majority class</i>	0.43	0.0	0.52	0.0	0.17
<i>Linear</i>	0.62	0.55	0.66	0.56	0.60
<i>BERT</i>	<b>0.75</b>	<b>0.68</b>	<b>0.76</b>	<b>0.75</b>	<b>0.73</b>
<i>Human</i>	0.71				

Table 4. Test-set performance, reported as accuracy, and macro-F1 score for each label (agrees, neutral, disagrees) and on average, of the best model of each type, trained using hyperparameters values corresponding to the model with the best cross-fold validation performance, with the overall best performing model shown in bold. See Appendix F for further details on how human performance was estimated.

## 5 Analyses

In this section, we first describe the lexicons of framing devices we use for our analyses (§5.1). We then present analyses that address our two research questions.

In §5.2, we find that *qualitatively*-speaking, both sides leverage similar linguistic framing devices for self-affirmation and opponent-doubt, but *quantitatively*-speaking, GW-skeptical media engages in more opponent-doubt. In §5.3, we find that both sides use opinion-framing to ascribe OPINIONS expressing their *own* stance to SOURCES known to publicly endorse the *opposing* view, thereby depicting such SOURCES as hypocritical.

### 5.1 Linguistic framing devices

Since GW opinion is closely connected to one’s attitude toward scientific evidence, we focus on framing devices with epistemic and evidential connotations in creating lexicons of affirming and doubting framing devices. We draw from work on factuality, commitment, and persuasion, as well as our own lexical semantic analysis, to create seed word sets; these seed sets are then augmented using WordNet to become our final lexicons.

**Affirming devices** We include factive and semi-factive predicates (*point out, understand* (N=20)), studied extensively in de Marneffe et al. (2011), Saurí and Pustejovsky (2012), Rudinger et al. (2018), Jiang and de Marneffe (2019), Ross and Pavlick (2019), among others. We add verbs with connotations of factivity and/or high subject commitment (*confirm, attest, certify, validate* (N=7)). We also add high commitment adjectives (*proven, settled* (N=4)) and adjectives of “hyping” from Lerchenmueller et al. (2019) (*breakthrough, expert* (N=38)). To complement these adjectives that affirm the *quality* of evidence, we add modifiers that affirm the *quantity* of evidence and index consensus (*many, numerous, dozens of* (N=11)).

**Doubting devices** We include words from semantic fields largely antonymous to those represented in the affirming seed words: neg-factive verbs (Saurí and Pustejovsky, 2009) such as *claim, pretend* (N=5), low commitment verbs (*doubt, dispute* (N=3)), low commitment adjectives (*dubious, so-called* (N=7)), adjectives of undermining (*flawed, debunked* (N=47)) and adjectives indexing lack of consensus (*few, contentious* (N=6)). We additionally include verbs with argumentative connotations (*argue, insist* (N=11)), as these can reinforce frames of debate and controversy.

We hope that our full lexicons (see Appendix I) will be useful for future work that looks at opinion-framing, especially in the context of other scientific debates (e.g., the COVID-19 pandemic).

### 5.2 Study 1 results

We apply our stance classification model to  $Op_{full}$ <sup>8</sup> to get a stance label for all embedded OPINIONS. We restrict our analysis to OPINIONS receiving a

<sup>8</sup>Because OPINION spans from certain media outlets are over-represented in  $Op_{full}$ , we repeat all analyses in Studies 1 and 2 while excluding data points from the top 5 LL and RL outlets (10 total) and obtain largely similar results (see Appendix L) to those presented in the main paper.

non-neutral label, as we can better guarantee having few mis-classifications of GW-agree (the sentence agrees with the target that GW is a serious concern) as GW-disagree (the sentence disagrees with the target that GW is a serious concern), and vice versa. We use political leaning as categorized by the Media Bias/Fact Check project<sup>9</sup> as a proxy for stance toward GW, with left-leaning and right-leaning outlets (LL and RL) corresponding to GW-accepting and GW-skeptical media, respectively. To find instances of self-affirmation in GW-accepting media, we retrieve GW-agree OPINIONS occurring with a PREDICATE or SOURCE modifier from the group of affirming devices (e.g., *show*, *peer-reviewed*); to find instances of opponent-doubt, we retrieve GW-disagree OPINIONS occurring with PREDICATES or SOURCE modifiers from the set of doubting devices (e.g., *claim*, *misleading*). This is repeated for GW-skeptical media, with OPINION stances swapped.

The resulting distribution over coverage types is shown in Fig. 3, indicating that the two sides are *not* symmetric in terms of their quantities of each coverage type: though both sides engage in more self-affirmation overall, GW-skeptical media (i.e., RL) shows a greater amount of opponent-doubt. This pattern corroborates prior work documenting the use of doubt by opponents of GW to dilute the scientific consensus (Oreskes and Conway, 2011).

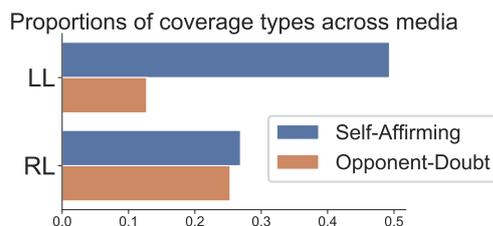


Figure 3. Proportions (among non-neutral OPINIONS) of self-affirming vs. opponent-doubting coverage in LL and RL, showing that LL primarily exhibits discourse where a GW-agree OPINION occurs with an affirming device, whereas RL exhibits more balanced amounts of self-affirmation and opponent-doubt. Most of the remaining OPINIONS are framed by words beyond those in our lexicons.

Turning to qualitative aspects of self-affirming and opponent-doubting discourse, we find that the two sides show symmetry in the framing devices used: devices that LL tends to use to frame GW-agree OPINIONS (e.g., *understand*, *recall*, *discover*;

<sup>9</sup><https://mediabiasfactcheck.com/>

*important*, *peer review*) tend to be used by RL for GW-disagree OPINIONS, and devices that RL uses to frame GW-agree OPINIONS (e.g., *pretend*, *claim*; *inaccurate*, *alleged*) tend to be used in LL for GW-disagree OPINIONS (see Fig. 4). We measure the tendency for a framing device to occur with a given OPINION stance as a log-odds-ratio between the number of times it frames OPINIONS of each stance, excluding words that occur under 20 times (see Appendix J for details). Broken down by the individual framing device (Figs. 5-6), we also see that, with some exceptions, the use of framing devices across LL and RL displays some symmetry. Notably, there seems to be a lack of affirming modifiers framing GW-agree OPINIONS in RL, suggesting that RL uses different modifiers to qualify SOURCES as convincing.<sup>10</sup>

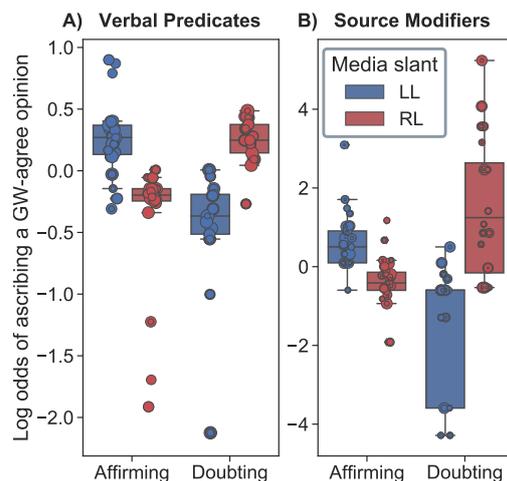


Figure 4. Distribution of the (log) odds of ascribing a GW-agree OPINION in LL and RL for affirming and doubting **a)** PREDICATES; **b)** SOURCE modifiers, showing that LL tends to ascribe GW-agree OPINIONS using affirming devices over doubting devices, whereas RL tends to ascribe GW-agree OPINIONS using doubting over affirming devices. Each point represents one framing device, and the size corresponds to its frequency in  $Op_{full}$ .

<sup>10</sup>As a robustness check, we repeat the same log-odds computation for the subset of data that excludes articles from newswires and find that the results are highly correlated with the full dataset (Pearson's  $r = 0.90$ ,  $p < 0.0001$  for verbs, Pearson's  $r = 0.82$ ,  $p < 0.0001$  for modifiers).

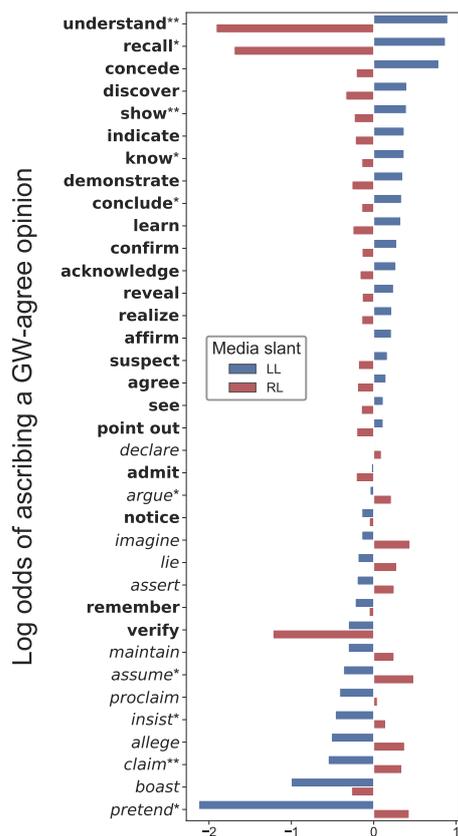


Figure 5. Log odds of ascribing a GW-agree OPINION for **affirming** and *doubting* predicates present in LL and RL, showing an overall symmetry in the devices LL and RL use for self-affirmation and opponent-doubt. A double asterisk (\*\*) indicates a significant bias for GW-agree OPINIONS in both LL and RL; (\*) indicates significance in one side. Significance ( $p < 0.05$ ) is determined via a chi-squared test and applying Benjamini-Hochberg correction with a false discovery rate of 0.1. Word order is given in descending value of log odds, as measured in LL.

### 5.3 Study 2 results

How faithfully does the media ascribe OPINIONS to SOURCES? We use Wikipedia lists<sup>11</sup> for GW-activist and GW-skeptic entities (Greta Thunberg, The Sierra Club; William Happer, The Heartland Institute) to label the stance of SOURCES that are named entities, after using fuzzy matching to resolve SOURCES to a canonical form. We define an OPINION as *faithfully* ascribed if the stance of the

<sup>11</sup> Activist lists: [https://en.wikipedia.org/wiki/Category:Climate\\_activists](https://en.wikipedia.org/wiki/Category:Climate_activists), [https://en.wikipedia.org/wiki/Category:Climate\\_change\\_environmentalists](https://en.wikipedia.org/wiki/Category:Climate_change_environmentalists). Unfortunately, the lists we used for climate change skeptics and deniers have since been deleted by Wikipedia. We manually removed entries that are neither people nor organizations, e.g., “Environmental Activism of Al Gore.”

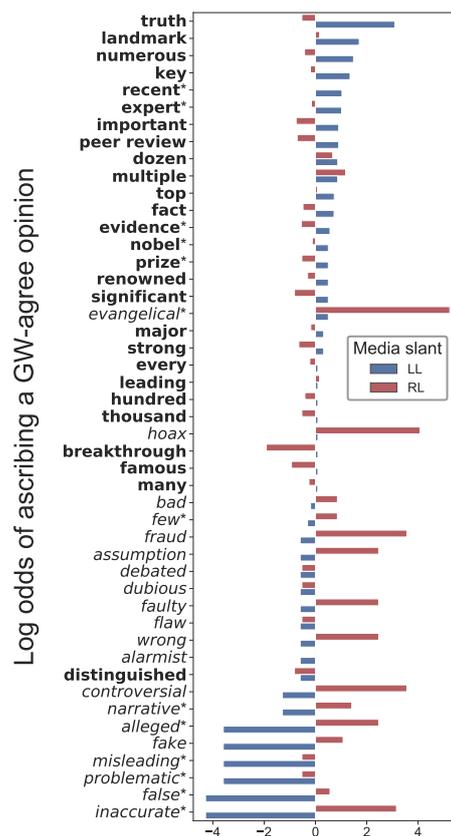


Figure 6. Log odds of ascribing a GW-agree OPINION for the **affirming** and *doubting* modifiers present in LL and RL. An asterisk (\*) indicates a significant bias for GW-agree OPINIONS in either LL or RL. Significance ( $p < 0.05$ ) is determined via a chi-squared test and applying Benjamini-Hochberg correction with a false discovery rate of 0.1. Word order is given in descending value of log odds, as measured in LL.

OPINION matches the stance of the SOURCE, e.g., a GW-agree OPINION is ascribed to a GW-activist.

Surprisingly, among the 4.3K OPINIONS ascribed to a named entity from the Wikipedia lists, we find that 37% and 38% are unfaithfully ascribed in LL and RL, respectively, suggesting that both sides frequently attribute OPINIONS to entities that differ from the well-established public positions of those entities. (See Appendix Tab. 10 for examples of unfaithfully ascribed OPINIONS.)

When we examine the unfaithful instances from LL more closely, we notice that the most frequent SOURCES have ties to the fossil fuel industry (e.g., *Exxon knew that the result of burning fossil fuels would create a climate crisis*), emphasizing the narrative of hypocritical oil companies that have long known about the harmful effects of greenhouse gases. In RL, by contrast, the unfaithful

Left-leaning media	Right-leaning media
understand, <b>concede</b> , <b>recall</b> , demonstrate, know, acknowledge, agree	realize, <b>recall</b> , learn, see admit, <b>concede</b> reveal

Table 5. PREDICATES biased toward hypocritical opinion attribution, i.e., attributing an own-side OPINION to an opposing-side SOURCE, in LL and RL. **Bolded** PREDICATES are used for hypocritical attribution in both LL and RL.

instances quote from a wide-range of activists and scientific bodies, but similarly emphasize these entities’ hypocrisy: *Gore admits that carbon dioxide is only responsible for about 40 % of the warming; NASA concedes that its temperature data are less than reliable*).

Finally, we ask whether certain PREDICATES are favored for ascribing OPINIONS unfaithfully. We might expect verbs like *admit* and *acknowledge*, which have connotations of reluctance, to be used for this purpose, and for verbs like *declare* and *insist* to be disfavored—it would be counter-intuitive for a reader of *The New York Times* to see the sentence, *Exxon insists that fossil fuels cause global warming*, for example.

To answer this question empirically, we measure each PREDICATE’S tendency to ascribe an OPINION to a SOURCE with an activist vs. skeptic stance, similar to how we measured PREDICATES’ tendency to embed an OPINION with a given stance. We retrieve in Tab. 5 the PREDICATES that are biased under this measure toward ascribing GW-agree OPINIONS to GW-skeptic SOURCES, and vice versa.

Interestingly, in addition to verbs we expected (*acknowledge*, *admit*, *concede*), we also find verbs like *understand*, *agree*, *realize*, *know*. One tendency among these verbs seems to be that they denote non-spoken acts of belief. Intuitively, it would be incompatible with real world events to describe Exxon as vocally denouncing fossil fuels or Al Gore as vocally criticizing climate science, but it is possible to describe such entities as silently holding contradictory beliefs (and in doing so, highlight their hypocrisy). However, we also see exceptions (*demonstrate* in LL, *reveal* in RL), suggesting that more complex interactions are involved.

## 6 Discussion and future work

In this work, we introduced **GWSD**, a novel dataset of 2K sentences from news media for studying GW stance. Using our dataset, we trained a weighted BERT model competitive with human performance to predict the stance of 500K opinions in news articles. Our initial analyses showed that both sides of the GW debate make use of framing devices in largely symmetric ways, though GW-skeptic media exhibits more opponent-doubt, in line with prior work on the propagation of GW skepticism (Oreskes and Conway, 2011). We also found that both sides exhibit considerable amounts of unfaithful opinion attribution, in particular to portray figures as hypocritical. Future work could take a more fine-grained approach to our analyses, such as disaggregating op-ed articles from non-op-eds or adopting labels for outlet stance beyond the binary “right-” vs. “left-leaning.” We also categorized named entities as either activists or skeptics, which obscures distinctions between, e.g., corporations with economic incentives for GW skepticism vs. individuals that may be ideologically motivated.

Our methodology may also be useful for work in argument mining: the main object of our inquiry—asccribed OPINIONS and the linguistic devices of SOURCE and PREDICATE used as syntactic markers of the attributive act—represents a novel dimension along which to analyze how premises are used to support claims (Stab and Gurevych, 2017).

Our work also highlights challenges inherent to studying stance: we found that many items can be ambiguous at the sentence-level, without a single “true” stance, and that demographic attributes like party affiliation and gender can affect how people respond. At the same time, we showed how Bayesian modeling can be used to account for this variation. Such findings reinforce the idea that NLP should be conscious of who the training data comes from, and how a model might be biased as a result. We hope that future research can benefit from and extend the current work to study argumentation inclusive of the many subjective and demographically-diverse attitudes in our society.

### Acknowledgements

We thank the reviewers and the Stanford NLP Group for helpful feedback, and Adina Abeles for feedback on the demographics portion of the MTurk task.

## References

- Dee Alexander, WJ Kunz, and Fred Walter Householder. 1964. *Some classes of verbs in English*, volume 1. Linguistics Research Project, Indiana University.
- Toby Bolsen, Risa Palm, and Justin T Kingsland. 2019. The impact of message source on the effectiveness of communications about climate change. *Science Communication*, 41(4):464–487.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Lorraine I Bridgeman and Fred Walter Householder. 1965. *More classes of verbs in English*. Indiana University Linguistics Club.
- Carmen Rosa Caldas-Coulthard. 2002. On reporting reporting: The representation of speech in factual and factional narratives. In *Advances in written text analysis*, pages 309–322. Routledge.
- Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. [Stan: A probabilistic programming language](#). *Journal of Statistical Software*, 76.
- Ignacio Cases, Clemens Rosenbaum, Matthew Riemer, Atticus Geiger, Tim Klinger, Alex Tamkin, Olivia Li, Sandhini Agarwal, Joshua D. Greene, Dan Jurafsky, Christopher Potts, and Lauri Karttunen. 2019. [Recursive routing networks: Learning to compose modules for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3631–3648, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.
- Robert B Cialdini. 1993. *Influence: The psychology of persuasion*. Harper Collins.
- Kevin Clark and Christopher D. Manning. 2015. [Entity-centric coreference resolution with model stacking](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.
- Gloria Cowan and Désiré Khatchadourian. 2003. Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly*, 27.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [Show your work: Improved reporting of experimental results](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Robert M. Entman. 2006. [Framing: Toward Clarification of a Fractured Paradigm](#). *Journal of Communication*, 43(4):51–58.
- Sarah R Esposito, Matthew J Hornsey, and Jennifer R Spoor. 2013. Shooting the messenger: Outsiders critical of your group are rejected regardless of argument quality. *British Journal of Social Psychology*, 52(2):386–395.
- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. [Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies](#). In *Proceedings of EMNLP*.
- Kelly S Fielding, Matthew J Hornsey, Ha Anh Thai, and Li Li Toh. 2020. Using ingroup messengers and ingroup values to promote climate change policy. *Climatic Change*, 158(2):181–199.
- Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.
- Elisabeth Gidengil and Joanna Everitt. 2003. Talking tough: Gender and reported speech in campaign news coverage. *Political communication*, 20(3):209–232.
- Ivan Habernal and Iryna Gurevych. 2016. [Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. [Learning whom to trust with MACE](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association*

- for *Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.
- Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. [Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.
- George Lakoff and Sam Ferguson. 2006. The framing of immigration. Retrieved from *The Rockridge Institute*. URL: <https://escholarship.org/uc/item/0j89f85g>.
- Marc J Lerchenmueller, Olav Sorenson, and Anupam B Jena. 2019. Gender differences in how scientists present the importance of their research: Observational study. *BMJ*, 367.
- Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2011. Veridicality and utterance understanding. In *2011 IEEE Fifth International Conference on Semantic Computing*.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [SemEval-2016 task 6: Detecting stance in tweets.](#) In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Jessica Gall Myrick and Suzannah Evans Comfort. 2020. The pope may not be enough: How emotions, populist beliefs, and perceptions of an elite messenger interact to influence responses to climate change messaging. *Mass Communication and Society*, 23(1):1–21.
- Naomi Oreskes and Erik M Conway. 2011. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. New York: Bloomsbury.
- Rita Orji, Regan L Mandryk, and Julita Vassileva. 2015. Gender, age, and responsiveness to Cialdini’s persuasion strategies. In *International Conference on Persuasive Technology*, pages 147–159. Springer.
- Alexander Michael Petersen, Emmanuel M Vincent, and Anthony LeRoy Westerling. 2019. Discrepancy in scientific authority and media visibility of climate change scientists and contrarians. *Nature communications*, 10(1):1–14.
- Pew Research Center. 2020. [As economic concerns recede, environmental protection rises on the public’s policy agenda.](#)
- James Powell. 2017. Scientists reach 100% consensus on anthropogenic global warming. *Bulletin of Science, Technology & Society*, 37(4):183–184.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 480–489.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. [Linguistic models for analyzing and detecting biased language.](#) In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.
- Ellen Riloff and Janyce Wiebe. 2003. [Learning extraction patterns for subjective expressions.](#) In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.
- Alexis Ross and Ellie Pavlick. 2019. [How well do NLI models capture verb veridicality?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.
- Joel Ross, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 2010. Who are the Turkers? Worker demographics in Amazon Mechanical Turk. In *Proceedings of CHI*.
- Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. *arXiv preprint arXiv:2009.09609*.
- Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. [Neural models of factuality.](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.
- Roser Saurí and James Pustejovsky. 2012. [Are you sure that this happened? Assessing the factuality degree of events in text.](#) *Computational Linguistics*, 38(2):261–299.
- Edwin Simpson and Iryna Gurevych. 2018. [Finding convincing arguments using scalable Bayesian preference learning.](#) *Transactions of the Association for Computational Linguistics*, 6:357–371.

- Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. [Modeling factuality judgments in social media text](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland. Association for Computational Linguistics.
- Christian Stab and Iryna Gurevych. 2017. [Parsing argumentation structures in persuasive essays](#). *Computational Linguistics*, 43(3):619–659.
- Oren Tsur, Dan Calacci, and David Lazer. 2015. [A frame of mind: Using statistical models for detection of framing and agenda setting campaigns](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.
- Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. [Committed belief tagging on the Factbank and LU corpora: A comparative study](#). In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Diyi Yang and Robert E. Kraut. 2017. [Persuading teammates to give: Systematic versus heuristic cues for soliciting loans](#). *Proc. ACM Hum.-Comput. Interact.*, 1.
- Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. [Conversational flow in Oxford-style debates](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

## Appendices

### A Data collection details

**URL filters** We filtered out articles that may be irrelevant on the basis of containing one of the following URL tags:

/automobiles/, /autoreviews/, /autoshow/, /business/, /campaign-stops/, /crosswords/, /booming/, /giving/, /gmcvb/, /jobs/, /lens/, /letters/, /newyork-today/, /nutrition/, /sept-11-reckoning/, /smallbusiness/, /sunday-review/, /garden/, /arts/, /theater/, /sports/, /dining/, /books/, /weekinreview/, /your-money/, /movies/, /fashion/, /technology/, /pageoneplus/, /travel/, /nytnow/, /public-editor/, /education/, /learning/, /podcasts/, /style/, /t-magazine/, /reader-center/, /awardsseason/, /briefing/, /deal-book/, /es/, /greathomesanddestinations/, /interactive/, /media/, /mutfund/, /obituaries/, /personal-tech/, /realestate/, /smarter-living/, /todaypaper/, /your-money/, /yourtaxes/, /slideshow/, /interactive/, /tag/, /author/, /clips/, /podcasts/, /subject/, /authors/, /category/, /person/, /category/, /shows/, /video/, /topic/, /topics/, /de/, /tags/, /slideshow/, /interactive/, /transcripts/, /headlines/

**Article deduplication details** We deduplicated articles by normalized URLs. In addition, we noticed that the same article corresponded in some cases to multiple different normalized URLs in our dataset, due to hyperlinking from different sections of a news site (e.g., blog section, RSS feed, front page). We de-duplicated these articles by comparing the article titles using a criterion adapted from Petersen et al. (2019): for two titles  $T_j, T_k$  with Damerau-Levenshtein edit distance of  $D_{jk}$ , if

$$D_{jk} \leq 0.2 \cdot \text{Min}(|T_j|, |T_k|),$$

then we consider the two titles, and hence the corresponding URLs, to index the same article.

### B SOURCE, PREDICATE, OPINION extraction algorithm

1. Find complement clause(s) in the dependency parse of a sentence, i.e., sub-tree(s) whose root has the dependency label “ccomp” (= OPINION);
2. Get head(s) of the complement clause(s), which correspond to the main verb that syntactically embeds the comp. clause (= PREDICATE); get children of the PREDICATE with

the dep. label “prt” (particle) in cases of multi-token verbs, e.g. *point out*;

3. To find the SOURCE, first check if the PREDICATE token is a participle (e.g., “a researcher, **warning** that [...]”—if yes, then find the head noun, otherwise, look within all children of PREDICATE and find the syntactic subject (token with the “nsubj\*” dependency label). In some cases, the head noun/syntactic subject may have the dependency label “relcl”, indicating that it’s inside a relative clause (e.g., “[...], **who** warns that”)—in this case, the true SOURCE is the antecedent of the relative pronoun, which we fetch by getting the head of the relative pronoun;
4. Get additional modifiers of SOURCE, PREDICATE and OPINION by recursively retrieving their children.

### C Lexical filters

We use the following lexical resources to filter extracted complement clauses to true indirect statements on the topic of GW.

**The Indiana Lists** Our algorithm returns subjunctive clausal complements, e.g., “Politicians require that [oil companies pay a carbon tax]”, which are nearly identical to embedded opinions, e.g., “Politicians claim that [oil companies pay a carbon tax]”. The Indiana Lists (Alexander et al., 1964; Bridgeman and Householder, 1965) categorize predicates according to whether they syntactically embed a subjunctive or indicative complement clause. We keep extracted (SOURCE, PREDICATE, OPINION) tuples only if the PREDICATE lemma is one of 418 indicative-clause-embedding verbs in these lists. This filter also effectively excludes extracted instances such as “We watch [oil companies pay a carbon tax]”.

**Implicatives** In addition to separating (S,P,O) tuples with overt negations (*The researchers did not say [that the effects of global warming are clear]*, *No studies find that [...]*), we also need to separate tuples that are implicitly negated (*The studies fail to find that [...]*, *Researchers refuse to say [...]* in order to accurately study how opinions are attributed. Since the dependency parser only recognizes explicit cases of negation, we use a list of 92 implicative constructions from Cases et al. (2019) to exclude tuples where the PREDICATE is in the scope of such an implicitly negating expression.

**Indirect questions** We exclude complement clauses that represent indirect questions (*Scientists ask what the future of nuclear looks like*) by excluding tuples that have a question word from the set {*who, what, when, where, how, whether, which*} as the complementizer.

**Topic keywords** climat, climact, global, warm, carbon, fossil, oil, energi, environ, co2, green, ice, glacier, glacial, melt, sea, temperatur, heat, hot, methan, greenhous, arctic, antarct, celsiu, fahrenheit, ecosystem, pole, environ, coal, natur, human, economi, electr, futur, health, scienc, econom, air, pollut, fire, wildfir, ipcc, epa, market, scientist, earth, planet, wind, solar, record, fuel, ocean, nuclear, scientif, pipelin, emit, emiss, concensu, renew, accord, forest, pruit, drought, hurrican, atmospher, activist, coast, agricultur, water, plant, weather, polar

## D AMT task details

To choose the subset of items for annotations from our full set of extracted OPINION spans, we filter to items that contain a smaller set of keywords (“climate”, “warming”, “carbon”, “co2”, or “fossil fuels”) and make a manual selection for each round of annotation such that the final sample is roughly balanced across different outlets.

We settle on a task design as follows: annotators are told that we are collecting their judgments of GW stance for a series of sentences; we then show an instructions page and guide them through 6 practice trials. They then annotate the main trial items for *agreeing*, *disagreeing*, or being *neutral* with respect to the target opinion, “Climate change/global warming is a serious concern.” Additional help text for each label is adapted from the setup described in Mohammad et al. (2016). The main trial items consist of 5 screen sentences and 30-50 sentences that have been transformed from the extracted OPINION using basic operations such as cleaning whitespace, capitalizing the first word, adding clause-final punctuation, matching for tense, and substituting abbreviations of named entities with the non-abbreviated form.

We divide the annotation into 5 rounds and recruit 8 annotators to annotate each item. Other than one worker who did the task on 3 different rounds, all other annotations come from unique annotators. We also restrict to annotators whose IP address is in the US, who have a minimum HIT approval rating of 98%, and at least 1,000 HITs

Target opinion: Climate change/global warming is a serious concern.

We can't afford to wait until everyone is feeling the pain of the climate emergency before we do something about it.

AGREES DISAGREES NEUTRAL

1. AGREES (the writer probably agrees with the target opinion)

This could be for any of reasons shown below:  
 - the writer mentions **explicit support** for the target  
 - the writer mentions **support for something/someone aligned with the target**,  
 - the writer mentions **opposition to something/someone opposing the target**

2. DISAGREES (the writer probably disagrees with the target opinion)

This could be for any of the following:  
 - the writer mentions **explicit opposition** to the target  
 - the writer mentions **opposition to someone/something aligned with the target**,  
 - the writer mentions **support for someone/something opposing the target**

3. NEUTRAL (it is unclear whether or not the writer supports the target opinion)

The writer could be either in support of or against the target; the sentence does not reveal the writer's stance.

approved. We collect annotator age, gender, level of education, political affiliation, state of residence, as well as measures of their own stance towards GW borrowed from the American Public Opinion on Global Warming project.<sup>12</sup> There is some demographic imbalance in our total sample of annotators (see Tab. 6 in E) but the distribution is similar to the estimated demographics of the AMT population located in the US as a whole (Ross et al., 2010). The price per item was set to ensure that workers were paid the California minimum wage of \$12 USD per hour.

## E Demographic and linguistic effects on annotations

The marginal statistics for annotator demographics are given in Table 6, and show a relatively representative sample in terms of age, gender, education, and political affiliation, though women are distinctly under-represented.<sup>13</sup>

In order to measure the bias associated with various characteristics of annotator demographics, we make use of the hierarchical ordinal logistic model given below. In this model,  $Y_{ij}$  is the response of annotator  $j$  to instance  $i$  (taking a value in  $\{1, 2, 3\}$ , corresponding to “disagree”, “neutral”, “agree”). In addition,  $q_i$  is the unnormalized stance associated with instance  $i$  (on a spectrum from “disagree” to “agree”),  $w_j$  is the bias associated with worker  $j$ ,

<sup>12</sup><https://pprggw.wordpress.com/>

<sup>13</sup>For political affiliation by age and gender in the US, see <http://pewrsr.ch/2FVWtww>

Answer	% of annotators
Age over 34	48.3 %
Female	37.3 %
Male	62.5 %
College degree or higher	66.5%
Democrat	46.0 %
Republican	21.2 %
Independent	28.8 %
Other political affiliation	4.0 %

Table 6. Demographic information on the 400 Mechanical Turk annotators who participated in our study.

$X_j$  is a vector of covariates associated with worker  $j$ ,  $\sigma_q^2$  and  $\sigma_w^2$  are learned variance parameters, and  $c_1$  and  $c_2$  are learned thresholds. We model the probability of each response according to:

$$p(Y_{ij} = k) = \begin{cases} 1 - g(\eta_{ij} - c_1) & \text{if } k = 1 \\ g(\eta_{ij} - c_1) \\ \quad - g(\eta_{ij} - c_2) & \text{if } 1 < k < K \\ g(\eta_{ij} - c_2) & \text{if } k = K \end{cases} \quad (1)$$

where

$$\eta_{ij} = q_i + w_j \quad (2)$$

$$q_i \sim \mathcal{N}(0, \sigma_q^2) \quad (3)$$

$$w_j \sim \mathcal{N}(\beta^T X_j, \sigma_w^2) \quad (4)$$

To complete the model, we place weakly informative half-normal priors on  $\sigma_q^2$  and  $\sigma_w^2$ , and weakly informative normal priors on  $\beta$ .

Using the above specification, we fit a series of models in which  $X_j$  represents, in turn, each of the covariates individually, followed by a series of combined models. We fit these models in Stan using 5 chains with 2000 samples, the first half thrown away as burn in.

Table 7 shows the estimated effects from each model on the propensity to respond with “agree” relative to “neutral”. Those with 95% credible intervals which exclude 1.0 are marked in bold. The results on the propensity to respond with “neutral” relative “disagree” are not shown, but are broadly similar.

In addition, we test whether the political leaning of the source outlet has any effect on the annotations received by the items drawn from the source. We find, unsurprisingly, that items from left-leaning media (LL) are significantly more likely to receive ratings of “agree”, but we do not

find a significant difference in level of annotator agreement for items drawn from LL vs. RL. Finally, we find that item length (no. of words) is slightly correlated with IAA (measured as entropy over labels; Spearman’s  $\rho = 0.06$ ,  $p = 0.016$ ).

## F Estimating Stance Distributions

To aggregate all ratings and obtain estimates of the stance distribution for each instance, we use a variant of the above model which allows inferring a distribution for each instance and each worker, along with a parameter representing the degree to which an annotator is failing to pay attention to the instance being annotated. Although the “disagree”, “neutral”, and “agree” categories can be treated as ordered (as above), here we treat them as unordered nominal categories, so as to allow for the possibility, for example, that an instance evokes both “agree” and “disagree”, but not “neutral” (i.e. it is ambiguous, but clearly not neutral).

Let  $Y_{ij}$  be the response from worker  $j$  to item  $i$ , let  $q_{ik}$  be the degree to which label  $k$  applies to instance  $i$ , and let  $w_{jk}$  be the bias of worker  $j$  towards label  $k$ . Finally, let  $v_j$  be the vigilance of worker  $j$  (i.e. the degree to which they pay attention to the prompt). We assume the following model

$$Y_{ij} \sim \text{Multinomial}(\text{Softmax}_k(\eta_{ij})) \quad (5)$$

$$\eta_{ijk} = v_j \cdot q_{ik} + (1 - v_j) \cdot w_{jk} \quad (6)$$

$$q_{ik} \sim \mathcal{N}(\mu_k, \sigma_q^2) \quad (7)$$

$$w_{jk} \sim \mathcal{N}(0, \sigma_w^2) \quad (8)$$

and fit it in Stan, placing weakly informative priors on  $\sigma_q^2$ ,  $\sigma_w^2$ , and a uniform prior on  $v_j \in [0, 1]$ . In order to help stabilize the model we set the mean parameter of the prior on  $q_{ik}$  to be  $\mu_k = \log p_k$ , where  $p_k$  is the overall proportion of the corresponding response in the data.

In order to estimate human performance for the purpose of comparison, we fit this model multiple times, but each time leave out a random 10% of the annotators. As can be seen in Figure 8, there is great variation in the degree to which annotators agree with the label inferred from the remaining 90% of annotators. To characterize the distribution of work accuracies, we fit a mixture of two normal distributions, and report the mean of the distribution corresponding to the high-accuracy annotators in the main paper (0.71).

Covariate	M1	M2	M3	M4	M5	M6	M7	M8
Age over 34	0.98							0.98
Female		<b>1.04</b>					<b>1.04</b>	<b>1.04</b>
College degree or higher			1.0					1.0
Democrat				<b>0.96</b>		0.97	0.98	0.98
Republican					<b>1.06</b>	<b>1.05</b>	<b>1.04</b>	<b>1.05</b>

Table 7. Effects of annotator demographics on the propensity to respond with “agree” rather than “neutral”. Coefficients in bold have 90% credible intervals which exclude 1.0.

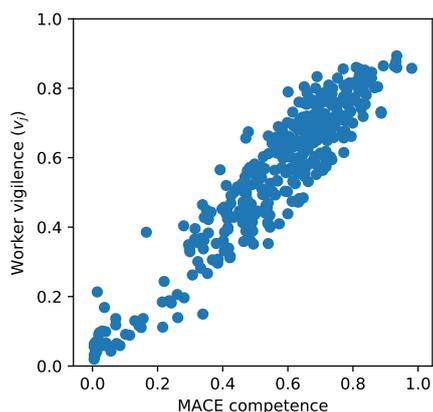


Figure 7. Showing the correlation between worker competence (estimated using MACE) and worker vigilance (estimated using our model) for the 400 annotators who participated in our data collection.

MACE / Ours	disagree	neutral	agree
disagree	386	6	0
neutral	12	852	19
agree	2	15	785

Table 8. Confusion matrix of (dis)agreements between MACE and our model

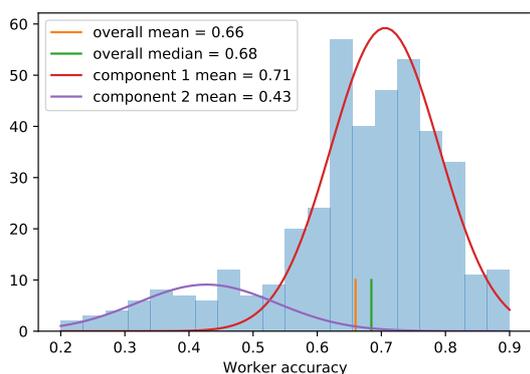


Figure 8. To estimate human performance, we refit the label aggregation model multiple times, each time leaving out 10% of annotators, and then comparing their annotations against the inferred label for the corresponding items. This plot shows that human performance appears to be a mixture of two distributions representing low and high agreement annotators.

## G Additional BERT experiments

Leading up to our hyperparameter search and baseline comparison, we experiment with a variety of training set-ups. Due to class imbalance in our training data (see Table 3), we try downsampling the majority classes as well as upsampling the minority class by adding back translations thereof. However, we did not obtain performance gains from either strategy in preliminary experiments. We further experiment with additional features in the form of the pre- and postceding  $n$  sentences ( $n = [1, 2]$ ) surrounding a training example, but did not obtain performance gains. We are also limited in the kinds of additional features (e.g., the political leaning of the outlet that a sentence comes from, the source entity that a sentence is attributed to) that we can use, since our goal is to analyze how the stance of the embedded statement is correlated with precisely these variables. We also try fine-tuning BERT first on a language modeling task (using our raw news data) and a natural language inference task (using the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets), respectively, prior to fine-tuning for sequence classification, but obtain no performance gains. Finally, we experiment with using tweets from known GW activists/skeptics as well as GW article headlines taken from extreme liberal/conservative news sources as additional training data, inferring labels based on the stance of the Twitter user or news source. However, we find that adding these examples yields a lower performance compared to using only the human-annotated data. This is not too surprising, given that embedded statements in the news tend to be longer and more complex than tweets/headlines.

## H Hyperparameter Tuning

All experiments took less than 7 days on one GPU (16 cores, 2.6GhZ, 128GB mem). For the BERT-based models (110M parameters), we use a max-

Hyperparameter	$\Delta$ accuracy	$p$ -value
Label weights	0.020	< 0.001
LM fine-tuning	0.004	0.11
Target opinion	-0.002	0.48
LR 2e-5 vs 1e-5	0.009	0.03
LR 4e-5 vs 1e-5	0.002	0.35

Table 9. Estimated effects of various hyperparameter choices on the average validation performance of the BERT<sub>base</sub> model.  $p$ -values are obtained using a Wilcoxon signed-rank test on the paired results from grid search.

imum sequence length of 256, a batch size of 16, and train for 7 epochs, saving a checkpoint after each epoch. In addition, we perform a grid search over the following hyperparameters:

- Label weights: [True, False]
- Language model fine-tuning: [True, False]
- Target opinion as second input: [True, False]
- Learning rate: [1e-5, 2e-5, 4e-5]

We train models for each combination of settings using five random seeds, and ultimately choose the hyperparameter configuration (including number of training epochs and random seed) that has the best validation performance, averaged over five folds, for a total of 600 configuration tested (including seeds and folds). We then retrain a model using those hyperparameter values on all non-test data.

Because we are using grid search, we can conveniently compare the effects of various hyperparameter choices. The overall average validation performance was 0.71, with a standard deviation of 0.04, and a 95% interval of [0.64, 0.77]. Table 9 shows the average increase in accuracy associated with each hyperparameter choice, along with a  $p$ -value computed using a Wilcoxon signed-rank test. As can be seen, using label weights leads to a significant increase in accuracy, as does using a learning rate of 2e-5 in comparison to 1e-5.

For the linear models (91504 params), we consider both logistic regression and SVM models, again using grid search and choosing the best-performing model on average validation performance, as described above. For the SVM, we search over all combinations of the following hyperparameters:

- Label weights: [True, False]

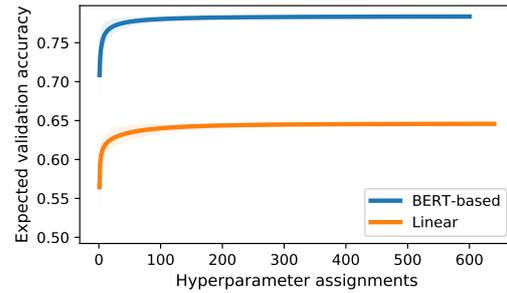


Figure 9. Expected validation performance of both types of models using validation accuracy scores from the hyperparameter grid search.

- n-gram order: [1, 2]
- kernel [rbf, linear, polynomial]
- gamma [scale, auto]
- Stopword removal: [True, False]
- Convert digits: [True, False]
- Regularization strength {0.01, ..., 1000}

For the logistic regression model, we search over all combinations of the following hyperparameters:

- Label weights: [True, False]
- n-gram order: [1, 2]
- Stopword removal: [True, False]
- Convert digits: [True, False]
- Regularization type: [ $l_1$ ,  $l_2$ ]
- Regularization strength {0.01, ..., 1000}

The mean validation accuracy among a total of 640 linear models tested is 0.56, with a standard deviation of 0.06, and a (0.41-0.62) 95% confidence interval. The linear model which performed best on validation data was a logistic regression bigram model using label weights, trained with  $l_2$  regularization, no stopword removal, no digit conversion, and regularization strength of 1.0.

Figure 9 compares these results directly, showing that the expected validation performance (Dodge et al., 2019) of the BERT-based models is uniformly better than that of the linear models, at least in terms of number of hyperparameter assignments.

## I Framing devices

### Affirming devices

- **Factive and semi-factive verbs:** uncover, realize, know, understand, learn, concede, remember, recall, discover, show, reveal, see, forget, find, point out, indicate, acknowledge, admit, realize, notice
- **High-commitment verbs:** certify, verify, corroborate, affirm, confirm, agree, conclude
- **High commitment adjectives:** proven, settled, conclusive, definitive
- **Hyping adjectives:** famed, unequivocal, skillful, notable, strong, famous, Nobel, skillful, Nobelist, Nobel Laureate, Nobel prize winner, Nobel prize winning, prize winning, award winning, distinguished, well-grounded, esteemed, proficient, key, evidence, noted, top, preeminent, breakthrough, significant, intelligent, of import, celebrated, novel, recent, major, landmark, important, distinguished, renowned, peer-reviewed, expert, leading
- **Consensus of evidence adjectives:** thousand, 1000, hundred, 100, unanimous, diverse, substantial, many, multiple, dozen, numerous

### Doubting devices

- **Neg-factive verbs:** pretend, lie, claim, allege, assume
- **Low commitment verbs:** doubt, dispute, debate
- **Argumentative verbs:** boast, declare, argue, maintain, contend, insist, proclaim, assert, brag, tout, convince
- **Low commitment modifiers:** narrative, evangelical, hoax, dubious, alleged, in question, so-called
- **Undermining adjectives:** discredited, debunked, distorted, misleading, inaccurate, corrupted, sketchy, faulty, erroneous, deficient, wrong, flawed, imprecise, incomplete, insufficient, invalid, unreliable, adulterated, false, mistaken, cherry-picked, defective, presumptive, non-peer-reviewed, exaggerated, overdone, overstated, delusive, awry, fake, bad,

misguided, substandard, fictive, fictitious, uncomplete, blemished, uncompleted, shoddy, dubitable, lacking, moot, untrue, problematic, faux, incorrect, inferior

- **Lack of consensus adjectives:** controversial, contentious, debated, few, debatable, contested

## J Quantifying bias toward framing OPINIONS with a GW-agree stance

We measure,  $B_{f,L}$ , the tendency for a framing device,  $f$ , within media with leaning  $L$ , to frame a GW-agree OPINION as:

$$B_{f,L} = \log \left( \frac{a_f}{A - a_f} \right) - \log \left( \frac{d_f}{D - d_f} \right),$$

where  $a_f$  is the number of times  $f$  occurs with a GW-agree OPINION,  $A$  is the total number of GW-agree OPINIONS,  $d_f$  is the number of times  $f$  occurs with a GW-disagree OPINION, and  $D$  is the total number of GW-disagree OPINIONS, all within  $L$ .

## K Named entity fuzzy matching

We use FuzzyWuzzy (<https://github.com/seattgeek/fuzzywuzzy>) to retrieve fuzzy matches of named entity SOURCES, setting the limit of matches to  $N = 100$ . We then manually filter out incorrect matches.

Tillerson	acknowledged that	climate change has ‘real’ and ‘serious’ risks but has previously downplayed climate change effects.
Exxon	knows that	fossil fuels caused global warming in the 1970s.
Exxon	knew that	the result of burning fossil fuels would create a climate crisis.
Gore	admits that	carbon dioxide is only responsible for about 40 percent of the warming.
Even the IPCC	acknowledges that	their previous estimates of “climate sensitivity” to greenhouse gases their reported in 2007 were significantly exaggerated.
NASA	concedes that	its temperature data are less than reliable.

Table 10. Examples of unfaithfulness in opinion attribution. **Top:** Examples of LL attributing **GW-agree OPINIONS** to **GW-skeptic SOURCES**. **Bottom:** Examples of RL attributing **GW-disagree OPINIONS** to **GW-activist SOURCES**. The IPCC refers to the U.N.’s Intergovernmental Panel on Climate Change.

## L Results on non-top-5 LL and RL media

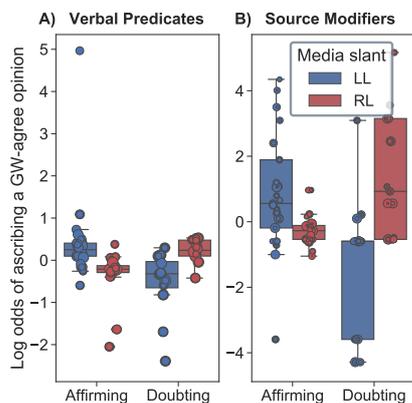


Figure 10. Log odds of ascribing a GW-agree OPINION in LL and RL for affirming and doubting PREDICATES (left panel) and SOURCE modifiers (right panel).

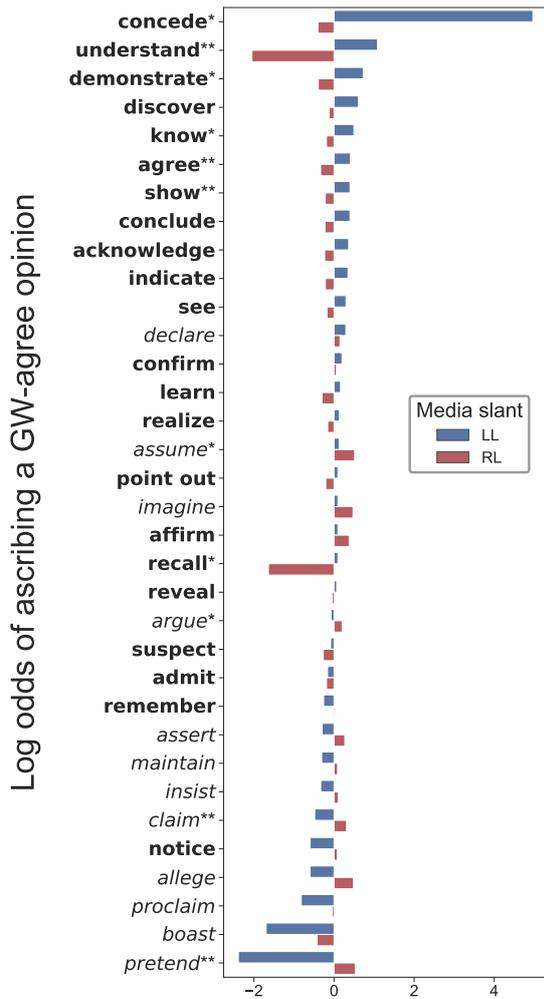


Figure 11. Log odds of ascribing a GW-agree OPINION in RL and LL for different PREDICATES, excluding the top 5 outlets by number of articles in each. A double asterisk (\*\*) indicates a significant bias for GW-agree OPINIONS in both LL and RL; (\*) indicates significance in one side. Significance ( $p < 0.05$ ) is determined via a chi-squared test and applying Benjamini-Hochberg correction with a false discovery rate of 0.1. Word order is given in descending value of log odds, as measured in LL.

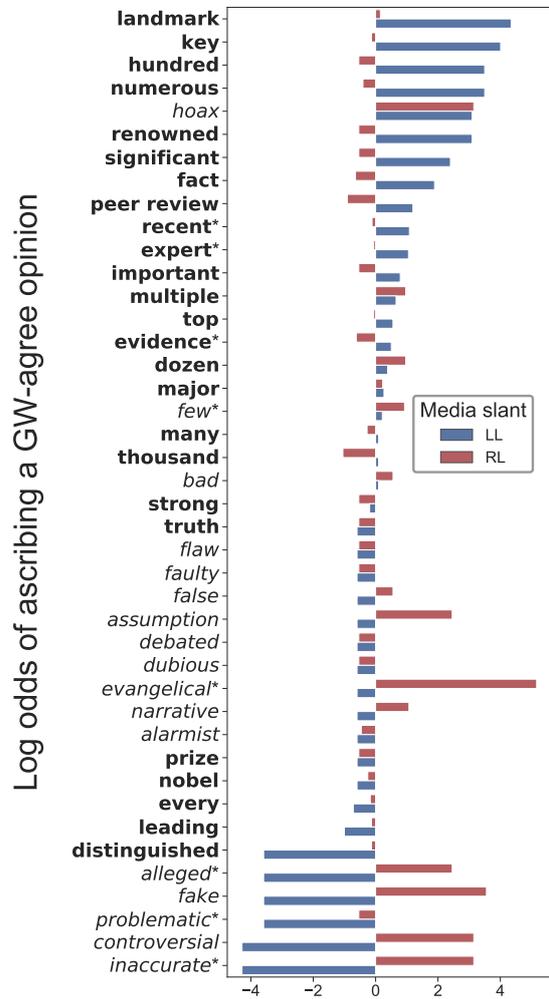


Figure 12. Log odds of ascribing a GW-agree OPINION in RL and LL for different SOURCE modifiers, excluding the top 5 outlets by number of articles in each. A single asterisk (\*) indicates a significant bias for GW-agree OPINIONS in either LL or RL. Significance ( $p < 0.05$ ) is determined via a chi-squared test and applying Benjamini-Hochberg correction with a false discovery rate of 0.1. Word order is given in descending value of log odds, as measured in LL.