Misinfo Belief Frames: A Case Study on Covid & Climate News

Saadia Gabriel[♠] **Skyler Hallinan**[♠] **Maarten Sap**[♠] **Pemi Nguyen**[♠]
Franziska Roesner[♠] **Eunsol Choi**[♣] **Yejin Choi**[♠♢]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[♣]The University of Texas at Austin, Austin, USA
[♢]Allen Institute for Artificial Intelligence, Seattle, USA
{skgabrie, msap, peming, franzi, yejin}@cs.washington.edu ,
{hallisky}@uw.edu, {eunsol}@cs.utexas.edu

Abstract

Prior beliefs of readers impact the way in which they project meaning onto news headlines. These beliefs can influence their perception of news reliability, as well as their reaction to news, and their likelihood of spreading the misinformation through social networks. However, most prior work focuses on fact-checking veracity of news or stylometry rather than measuring impact of misinformation.

We propose **Misinfo Belief Frames**, a formalism for understanding how readers perceive the reliability of news and the impact of misinformation. We also introduce the Misinfo Belief Frames (MBF) corpus, a dataset of 66k inferences over 23.5k headlines. Misinformation frames use commonsense reasoning to uncover implications of real and fake news headlines focused on global crises: the Covid-19 pandemic and climate change.

Our results using large-scale language modeling to predict misinformation frames show that machine-generated inferences can influence readers' trust in news headlines (readers' trust in news headlines was affected in 29.3% of cases). This demonstrates the potential effectiveness of using generated frames to counter misinformation.

1 Introduction

Understanding the impact of misinformation requires more than fact-checking its veracity. Prior beliefs of readers can bias them towards trusting misinformation (Britt et al., 2019). Also, readers are more likely to believe misinformation that is repeatedly circulated (Hasher et al., 1977). Misinformation that stokes fear during global crises (e.g. pandemics and climate disasters) can have major detrimental impacts on the emotional well-being of readers and increase uncertainty.

We propose **Misinfo Belief Frames**, a formalism for understanding perceived reliability of news. We also introduce the Misinfo Belief Frames (MBF) corpus consisting of 66k inferences over 23.5k headlines. Our formalism captures how readers view the implications of news events through the lens of pragmatic frames. Inspired by frame semantics (Fillmore, 1976), we provide a means of situating news content within a broader context. Misinfo Belief Frames go beyond linguistic meaning and focus on the way in which readers' prior

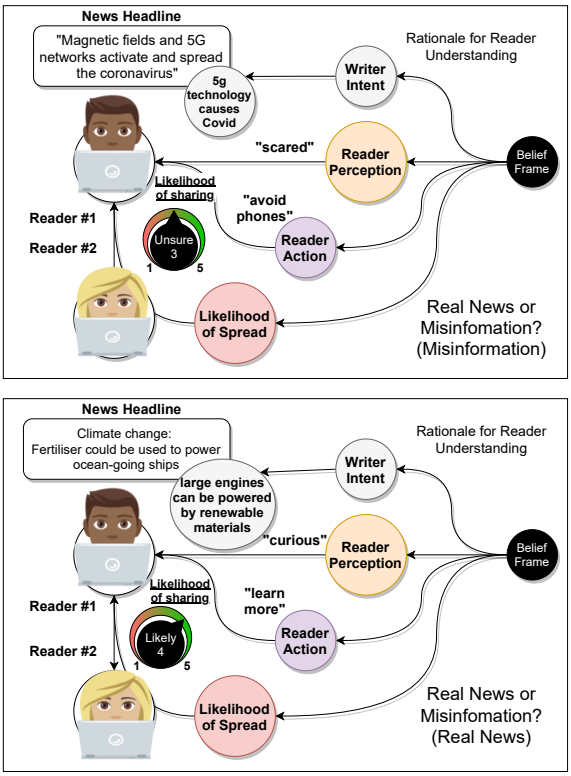

Figure 1: Understanding how a health or climate news article is interpreted as reliable or misinformation by readers requires pragmatic reasoning about not only linguistic features (e.g. emotions invoked by certain content words), but also domain-specific and social commonsense reasoning (e.g. *"What would someone do if they thought 5G networks were unsafe?"*, *"How would someone feel if they thought 5G spread viruses?"*). We propose Misinfo Belief Frames, pragmatic frames aimed at explaining implications from news headlines as well as its potential impact on readers.

beliefs project meaning onto news headlines. Our pragmatic frames use understanding of the role personal belief systems play in interpretation (Hall, 1973) and social commonsense (Apperly, 2010; Sap et al., 2019). As shown by Figure 1, belief frames consist of four dimensions related to implications of news events - *writer intent*, *reader perception*, *reader action* and *likelihood of spread*. Given the headline *"Magnetic fields and 5G networks activate and spread the coronavirus,"* we infer that a reader might *"feel scared"* and *"want to avoid phones."* We can also learn more about the perceived reliability of the news or overall appeal of the article given the likelihood the reader would share it within their network.

Prior work has shown the effectiveness of lexical techniques and large-scale language modeling for misinformation detection (Mihalcea and Strapparava, 2009; Rashkin et al., 2017; Karimi et al., 2018; Zellers et al., 2019). However, stylometry alone cannot reliably measure trustworthiness of news (Schuster et al., 2020). Reliable sources can use similar persuasive techniques and linguistic cues to misinformation. Prior binary labeling schemes also hide the impact and subliminal messaging captured by misinformation.

We use the pragmatic frames represented by MBF corpus inferences to analyze differences between misinformation and real news in the climate and Covid health domains. Notably, human evaluation shows that machine-generated MBF inferences affect readers' trust in headlines for 29.3% of cases. We also compare against prior formalisms for identifying rhetorical techniques in descriptions of news events using zero-shot prediction to show that emerging misinformation requires new understandings of how bad actors convey malicious intentions (rhetoric detection models achieve a F1 score of 34.59% at misinformation detection compared to a F1 score of 83.97% for task-specific models) . Finally, we contribute MBF benchmark results from large-scale models and show domain-specific pretraining improves misinformation detection over MBF headlines by 5%.

Our new formalism provides a framework for understanding *the potential impact* of news headlines by collecting crowdsourced annotations of reader perceptions, actions, and information spread. We introduce Misinfo Belief Frames and release the MBF corpus to aid in better design of methods to defend against misinformation, including generation of effective counter narratives to false beliefs spread by bad actors.

2 Misinfo Belief Frames

We introduce our belief frames for understanding misinformation. We first discuss related background work and then define our formalism.

2.1 Defining Reliability of News

Table 1 shows real and misinformation news examples from our dataset with headlines obtained from sources described in Section 3. A number of definitions have been proposed for labeling news articles based on reliability. To scope our task, we focus on false news that may be unintentionally spread (misinformation). This differs from disinformation, which assumes a malicious intent or desire to manipulate. In our framework, we focus on intent in terms of implications rather than questioning whether or not the writer's intentions were malicious given that it is unclear the extent to which original writers might have known article content was misleading. We summarize common definitions for news reliability in Appendix A.2 (See Table 11).

2.2 Rhetorical Aspects

Prior work on rhetorical framing (Nisbet and Scheufele, 2009; Card et al., 2015; Field et al., 2018) has noted the significant role *media frames* play in shaping public perception of social and political issues, as well as the potential for misleading representations of events in news media. However, past formalisms for rhetorical framing that rely on common writing or propaganda techniques (e.g. *appeal to fear* or *loaded language*, (Da San Martino et al., 2019)) may not represent emerging trends in misinformation, particularly as real news becomes more sensationalized (see section 6.5 for zero-shot analysis of propaganda techniques on news events in the MBF corpus). They also do not focus on the effectiveness of these techniques when used in practice. To that end, we propose a formalism focusing on readers' perception of the writers' intention, rather than specific well-known techniques.

2.3 Belief Frame Dimensions

We design pragmatic frames, described in free-text inferences, invoked by news event. Our formalism builds upon the encoder-decoder theory of media (Hall, 1973), which proposes that before an event

News Headline	Writer's Intent	Reader Reaction	Spread	Topic	Real / Misinfo Perceived	Real / Misinfo Gold
Covid-19 may strike more cats than believed.	Cats can get (and maybe transmit) covid	protect their cats	3.5	Covid	Real	Real
World health organization's report says not a single vegetarian has contracted COVID-19 so far.	Eating vegetables is a way to prevent getting covid	eat more vegetables	4	Covid	Misinfo	Misinfo
An "official" mask to combat the novel coronavirus was released.	Some masks are better than others	learn where to purchase mask	2	Covid	Real	Misinfo
Climate zealots have taken Canada hostage and the PM is missing in action.	People disagree with the choices made by the prime minister	feel angry	2	Climate	Misinfo	Misinfo
How to discuss "climate change" with a 'woke' teenager.	There are good ways to spark discussion	read the article	3	Climate	Real	Misinfo
Economists win Nobel for work on climate and growth.	A prize was won due to work on climate change	praise these people for working on solutions	4.5	Climate	Real	Real

Table 1: Example instances in MBF corpus showing belief frame annotations with writer intent and reader reaction (either a perception or action), as well as likelihood of the news event being shared. We also provide gold labels and the perceived labels obtained from annotators.

is communicated, a narrative discourse encoding the objectives of the writer is generated. We focus on the readers' interpretation of writer's intent and the impact on the readers. We use free-text inferences, following prior work on understanding of commonsense relations and social behavior (Speer and Havasi, 2012; Rashkin et al., 2018; Sap et al., 2020; Forbes et al., 2020). Each belief frame contains the following elements:

Event We describe each headline prompt presented to annotators as a news *event*, which summarizes the main message of an article. Section 3 explains this further. An example of a news event in our dataset is "*Covid-19 may strike more cats than believed*" (Table 1, row 1).

Writer Intent We ask annotators to reason about the *intentions* of the writer when describing a particular event. For example, given the headline "*An "official" mask to combat the novel coronavirus was released*" (Table 1, row 3), a reader might infer that the writer implied that "*some masks are better than others*." To provide a structure for annotators, we used a subset of the data (approx. 200 examples per news topic) to determine a list of 7 common important themes (e.g. technology or government entities) appearing in Covid and climate news. We provide a list of all the themes in Table 2, some themes are shared between news topics.[1]

We then asked annotators if each theme was relevant to the event. If a theme was relevant, we asked annotators to provide 1-3 inferences related to the chosen theme.

Reader Perception We ask annotators to describe how readers would *mentally* respond to a news event. For this, we ask annotators how readers would feel in reaction to a news event. These inferences include emotional reactions (e.g. "*feeling angry*") and observations (e.g. "*feeling that the described event X would trouble most people*"). For this dimension, we elicit 1-2 inferences and allow annotators to select from up to 7 themes.

Reader Action We ask annotators to describe how readers would *physically* respond to a news event. For this, we ask annotators what readers would want to do after seeing the news headline. These inferences describe physical actions a reader would want to take after reading about the news event (e.g. "*wanting to protect their cats*"). For this dimension, we elicit 1-2 inferences.

Likelihood of Spread To take into account variability in impact of misinformation due to low or

[1]Note that themes are not disjoint and a news article may capture aspects of multiple themes.

Theme	Climate	Covid
Climate Statistics	✓	
Natural Disasters	✓	
Entertainment	✓	
Ideology	✓	
Disease Transmission		✓
Disease Statistics		✓
Health Treatments		✓
Protective Gear		✓
Government Entities	✓	✓
Society	✓	✓
Technology	✓	✓

Table 2: Themes present in articles by each news topic. Some themes (e.g., society, technology) are covered by both climate and Covid domains, while others are domain specific.

Statistic	Train	Dev.	Test
Events	19,187	2,372	1,968
Unique Intents	37,225	4,751	4,083
Unique Perceptions	3,118	640	507
Unique Actions	14,239	2,051	1,534
Total Event/Inference Tuples	98,098	12,128	10,847

(a) Dataset-level breakdown of statistics for MBF corpus.

Statistic	Full Data
Avg. Intents per Headline (Climate)	2.09
Avg. Intents per Headline (Covid)	2.12
Avg. Perceptions per Headline (Climate)	1.67
Avg. Perceptions per Headline (Covid)	1.64
Avg. Actions per Headline (Climate)	1.35
Avg. Actions per Headline (Covid)	1.43

(b) Topic-level breakdown of statistics for MBF corpus.

Table 3: Dataset statistics.

high appeal to readers, we measure the likelihood of an article being shared. For this question, we ask annotators to rate each news event based on how likely it is that they would share the article given the event. We use a 1-5 Likert scale (Likert, 1932) with the following categories: {*Very Likely*, *Likely*, *Neutral*, *Unlikely*, *Very Unlikely*}.

Perceived Label Finally, we ask annotators whether or not they believe the news event is from a misinformation or real news article.

3 News Data Collection

We examined reliable and unreliable news events extracted from two domains with widespread misinformation: Covid-19 (Hossain et al., 2020) and climate change (Lett, 2017). We collect news from both misinformation sources and trustworthy outlets.

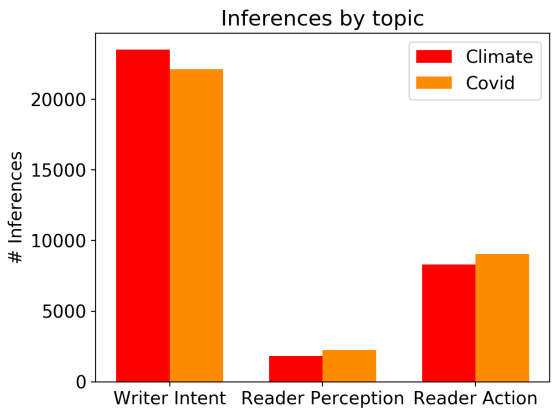

Figure 2: Breakdown of MBF corpus inferences by topic.

3.1 Covid-19 Dataset

For trustworthy news regarding Covid-19, we use the CoAID dataset (Cui and Lee, 2020) which contains 3,565 news headlines from reliable sources. These headlines contain Covid-19 specific keywords and are scraped from nine trustworthy outlets (e.g. the CDC, WHO, and NIH).

For unreliable news (misinformation), we use The CoronaVirusFacts/DatosCoronaVirus Alliance Database, a dataset of over 10,000 mostly false claims related to Covid-19.[2] These claims originate from social media posts, manipulated media, and news articles, that have been manually reviewed and summarized by fact-checkers.

3.2 Climate Change Dataset

We retrieved both trustworthy and misinformation headlines related to climate change from NELA-GT-2018-2020 (Gruppi et al., 2020; Norregaard et al., 2019), a dataset of news articles from 519 sources. Each source in this dataset is labeled with a 3-way trustworthy score (reliable / sometimes reliable / unreliable). We discard articles from "sometimes reliable" sources since the most appropriate label under a binary labeling scheme is unclear. To identify articles related to climate change, we used keyword filtering.[3]

4 MBF Corpus Annotation

We obtained 66,311 inferences from Covid and climate news (See section 3) by eliciting annotations for news events in 23,527 news articles (11,500 Covid related articles and 12,027 climate articles).

[2] https://www.poynter.org

[3] We kept any article headline that contained at least one of environment, climate, greenhouse gas, or carbon tax.

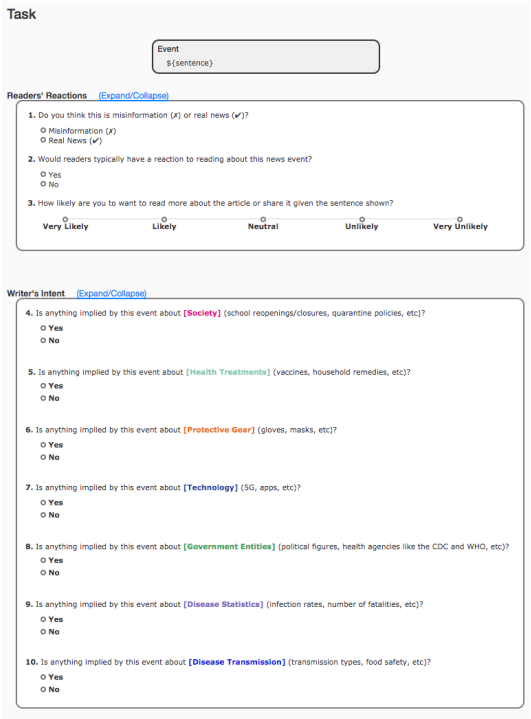

Figure 3: Layout of annotation task for collecting health-related commonsense inference data. See Figure 5 in the appendix for a larger version.

In this section we outline the structured annotation interface used to collect the dataset. Statistics for the full dataset are provided in Table 3. We provide a topic-level overview of inferences in Figure 2. The distribution of inferences is relatively even across topic categories.

4.1 Annotation Task Interface

Misinformation Belief Frames are annotated using the Amazon Mechanical Turk (MTurk) crowdsourcing platform.[4] The layout of our annotation task is given in Figure 3. For ease of readability, we present a *news event* summarizing the article to annotators, rather than the full text of the article.[5] We structure the annotation framework around the themes described in section 2.3.

4.2 Quality Control

We use a three-stage annotation process for ensuring quality control. In the initial pilot, we select a pool of pre-qualified workers by restricting to workers located in the US who have had at least 99% of their hits approved and have had at least 5000 hits approved. We paid workers at a rate of $0.4 per hit during these pilots. We approved workers who consistently submitted high-quality annotations for the second stage of our data annotation, in which we assessed the ability of workers to discern between misinformation and real news. For the second stage pilot and final task, we pay workers at a rate of $.6 per hit. We removed workers whose accuracy at predicting the label of news headlines fell below 70%. Our final pool consists of 79 workers[6] who submitted at least three annotations during the pilot tasks (on average workers in the second stage of quality control had submitted 170 annotations). We include all annotations from these workers in the pilots and final task as part of the dataset, discarding annotations from disqualified workers. We also removed headlines that received no annotations due to deformities in the original text (e.g. unexpected truncation) or vagueness. We achieve pairwise agreement of 78% on the label predicted by annotators during stage 3.

5 Modeling Belief Frames

We test the ability of large-scale language models to generate inferences for unseen news headlines using conditional generation (Sutskever et al., 2014; Rush et al., 2015). We use topic- and dimension-based special tokens to control generation of belief frames for T5 encoder-decoder (Raffel et al., 2020) and GPT-2 decoder-only models (Radford et al., 2019). Both models are based on a transformer architecture (Vaswani et al., 2017), consisting of transformer blocks with self-attention and feed-forward layers as well as layer normalization.

5.1 Training

Given a headline h of length T tokens, topic token $s_t \in \{[health],[climate]\}$ and dim token $s_d \in \{[writer\_intent],[effect\_on\_reader],[reader\_action]\}$, we pass the following input vector x to our language model:

$$x = h_1 \ldots h_T \,||\, s_d \,||\, s_t.$$

where $||$ represents concatenation.

For decoder-only models we also append the gold inference $y = g_1 \ldots g_N$, where N is the length of the inference, and take the loss over the full sequence. All our models are optimized using cross-entropy loss, where for a sequence t:

$$CE(t) = -\frac{1}{|t|}\sum_{i=1}^{|t|} logP_\theta(t_i|t_1, ..., t_{i-1}).$$

[4] https://www.mturk.com/

[5] These news events are either article headlines or claims.

[6] See Appendix A.1 for annotator statistics.

Here P_θ is the probability given a particular language model θ.

5.2 Inference

We predict each token of the output inference starting from the topic token s_t until the $[eos]$ token is generated. In the case of data with unknown topic labels, this allows us to jointly predict the topic label and inference. We decode using beam search.

6 Experiments

We first describe setup for experiments, as well as evaluation metrics for generation and classification experiments using our corpus (section 6.2). We then describe analysis over gold Misinfo Belief Frames (section 6.3). In sections 6.4 and 6.5, we show the performance of large-scale language models on the task of generating Misinfo Belief Frames and provide results for classification of news headlines.

6.1 Setup

We determine the test split according to date.[7] We use stratified sampling to determine the training and validation splits to ensure an even distribution of climate and Covid news. We use the HuggingFace Transformers library for all experiments (Wolf et al., 2020). Generation models are trained for a maximum of 8 epochs using early stopping based on dev. loss. We optimize using AdamW (Loshchilov and Hutter, 2019) and linear warmup. Hyperparameters are provided in Appendix A.3.

6.2 Evaluation Metrics

6.2.1 Automatic

We compare belief inference systems using common automatic metrics, including the BLEU (-1/2/3/4) ngram overlap metric (Papineni et al., 2002) and BERTScore (Zhang et al., 2020), a model-based metric for measuring semantic similarity between generated inferences and references. We additionally measuring diversity of generated inferences using Self-BLEU-2 (Zhu et al., 2018), and novelty of inferences.

6.2.2 Gold

We also evaluate aspects of gold inferences to consider potential differences between beliefs held for misinformation headlines compared to real news.

[7] We use news articles from 2021 and the last two months of 2020 for the test set.

Divergence For articles with more than one unique inference annotated along a particular belief frame dimension, we measure the divergence in beliefs or reactions invoked by the headline by measuring the average cosine distance between pairs of embedded gold inferences.

Sentiment We measure the sentiment of beliefs or reactions invoked by headlines by measuring the *valence* (degree of positivity), *arousal* (degree of emotionality), and *dominance* (degree of agency/control) of lexical content. For this evaluation we use the NRC-VAD lexicon (Mohammad, 2018).

6.2.3 Human Evaluation

For human evaluation, we assess generated inferences using the same pool of qualified workers who annotated the original data.

Overall Quality We ask the annotators to assess the overall quality of generated inferences on a 1-5 Likert scale (i.e. whether they are coherent and relevant to the headline without directly copying).

Influence on Trust We measure whether generated inferences impact readers' perception of news reliability. We ask annotators whether a given generated inference makes them perceive the news headline as more (+) or less (-) trustworthy.

Sociopolitical Acceptability We ask annotators to rate their perception of the beliefs invoked by an inference in terms of whether they represent a majority (mainstream) or minority (fringe) viewpoint.

6.3 Analysis of Gold Inferences

We conduct a series of analyses using gold inferences in the MBF corpus to better understand how readers perceive and act upon misinformation compared to real news.

6.3.1 Effect of Reader Perception on Article Sharing

Annotators tended to be cautious in reported sharing behavior (Figure 4). The average score for likelihood of sharing the article based on the news event was close to neutral (3). We found that annotators did have a higher likelihood of sharing real articles over misinformation articles (Table 5), and importantly generally claimed that they would not share articles that they thought were misinformation. For 2.3% of articles reported as misinformation annotators did provide a likelihood of

News Event (Spread)	Pred/Gold
COMMENTARY: We Can't Ignore the Harms of Social Distancing (4.0)	Misinfo/Real
NATO's Arctic War Exercise Unites Climate Change and WWIII (4.0)	Misinfo/Real
Eat Bugs! EU Pressing member States to Promote Climate Friendly Insect Protein Diets (4.0)	Misinfo/Misinfo
Coronavirus was created in Wuhan lab and released intentionally. (5.0)	Misinfo/Misinfo

Table 4: News events that were labeled as misinformation by annotators and also given a high aggregated likelihood of being shared (spread). We show the predicted and gold labels.

Label Type	Misinfo ↓	Real ↑	Effect size
Pred	2.529	3.214	0.763
Gold	2.203	3.225	1.107

Table 5: Likelihood of news events spreading, i.e. the annotators' rating for how likely it is they would share the article based on the shown news event. For "Pred", we ignore headlines where annotators were unsure about the label. For this and the following tables, arrows indicate the desired direction of the score. We use Cohen's d to compute effect size.

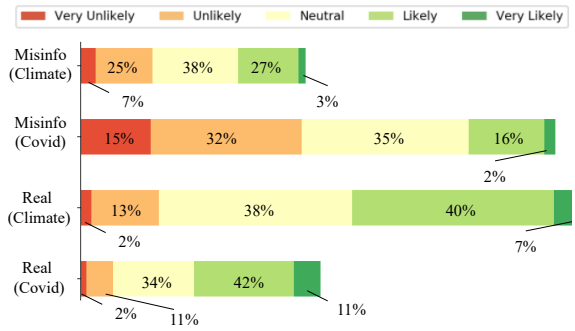

Figure 4: Distribution of spread (likelihood of sharing) scores in the training set. Aggregated scores are rounded.

sharing ≥ 4. We show examples of these articles in Table 4. While the reasoning for this is unclear, the annotators' belief frame predictions for reader perceptions and actions may provide insight. For example, annotators were skeptical of the misinformation news event "*Coronavirus was created in Wuhan lab and released intentionally.*," but said they would share it anyway and provided "*readers would feel curious*" and "*readers would want to know if the wild claim has any truth to it*" as related inferences. Concerningly, this indicates even very obvious misinformation may still be shared by generally knowledgeable readers when it contains content they deem particularly interesting or they want to corroborate the article content with others.

Overall, however, we found that annotators' perception of an article as being more reliable played a positive role in their decision to share it.

6.3.2 Divergence of Opinions

In Figure 6 (see the Appendix) we visualize the t-SNE representations of writer intention inferences associated with 100 news events. We use the BERT-large model to embed inferences. The mappings show that while writer intentions associated with the same news event tend to be clustered together, there are a number of outliers where the same news event leads to writer intentions with differing contextual representations. When we randomly sample 2000 misinformation events and 2000 real news events, we find a statistically significant difference between average cosine distance (divergence) of embeddings for writer intentions. Misinformation news events lead to more divergent writer intention annotations ($div = 0.596$) than real news events ($div = 0.542$).

6.3.3 Sentiment of Events

We find that there are distinctions between sentiment conveyed by misinformation compared to real news events (Table 8). In particular, misinformation news events scored lower on valence, arousal and dominance. This implies that misinformation tends to have a slightly more negative effect on readers (low valence), and use language conveying low agency or lack of control (low dominance). While the effect of news reliability is smaller on arousal (emotionality), we still find that real news has higher arousal. Headlines that annotators *believed* were misinformation also had slightly higher negative sentiment.

6.4 Generating Belief Frames

The automatic evaluation results of our generation task are provided in Table 7. Results are mixed, with GPT-2 performing better on ngram metrics while T5 generations had higher scores on our contextual BERTScore metric, diversity and novelty. For human evaluation, we restricted to headlines where the four model baseline variants generated different inferences, then randomly sampled 58 model-generated "writer's intent" inferences from the dev. set. We elicited 3 unique judgements per

Model	Quality (1-5)	Influence on Trust (%) +Trust	Neutral	-Trust	Socially Acceptable? (%) Yes	No
T5-base	2.91	11.11	72.52	16.37	83.66	16.34
T5-large	2.84	10.34	75.86	13.79	82.58	17.42
GPT-2 (small)	2.88	4.02	75.87	**20.11**	**87.59**	**12.41**
GPT-2 (large)	**2.98**	**12.07**	**70.69**	17.24	82.89	17.11

Table 6: Results of human evaluation (generation task).

	Model	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	BERTScore ↑	Diversity ↓	Novelty ↑
dev.	T5-base	96.962	75.252	49.103	35.674	**58.200**	**97.083**	88.646
	T5-large	97.002	72.288	48.353	35.137	56.771	97.157	**88.874**
	GPT-2 (small)	**98.823**	**76.057**	**53.010**	40.167	55.329	98.438	86.724
	GPT-2 (large)	98.567	75.940	52.904	**40.500**	55.775	98.281	83.705
test	GPT-2 (large)	98.025	81.284	60.048	45.085	57.113	98.671	82.892

Table 7: Automatic baseline results (generation task).

Dimension	Category	Pred ↑	Gold ↑
Valence	Misinfo	0.205	0.210
	Real	0.257	0.264
Arousal	Misinfo	0.176	0.178
	Real	0.214	0.221
Dominance	Misinfo	0.201	0.208
	Real	0.253	0.260

Table 8: VAD scores for news events (training set). The "Pred" column uses labels assigned by annotators, while the "Gold" column uses the ground-truth labels. Statistically significant results for $p < .001$ are underlined. For "Pred", we ignore headlines where annotators were unsure about the label.

headline for a total of 696 judgements. Annotators were not told whether or not inferences were machine-generated, and we advised annotators to mark inferences that were copies of the headlines as low quality. Inferences were also templated in the form *"The writer is implying that [inference]"* for ease of readability.

While there were not significant differences between model variants, we found that the GPT-2 large model was rated as having slightly higher quality generations than the other model variants (Table 6). Most model generations were rated as being *"socially acceptable"*, however in as many as 17.42% of judgements, generations were found to be not acceptable. This may be due in part to outlandish claims in headlines. Interestingly, all models were rated capable of persuading readers to trust or distrust headlines - In particular, readers rated GPT-2 large as influencing their opinion in 29.31% of judgements. This is an indicator machine-generated belief frame-based interpretations of headlines may serve as useful aids in countering misinformation.

6.5 Classification Results

To test the limits of using stylometry to identify misinformation in our dataset, we predict the presence of rhetorical techniques commonly associated with propaganda in news event descriptions. For this, we use four pre-trained BERT propaganda classification models (Da San Martino et al., 2019) which we denote here as Prop-BERT. These models can be used to predict if any of 18 known rhetorical techniques are used to describe a news event.[8] For our zero-shot misinformation detection setting, we classify a news event as real if it is not associated with any rhetorical techniques and misinformation otherwise. As shown by table 10, F1 results are worse than a majority baseline when we classify based only on predicted rhetorical features. This is likely due to the fact both real and misinformation news uses these techniques (See Table 13 for examples).

Neural misinformation detection models are able to outperform humans at identifying misinformation (achieving a max F1 of 83.97 compared to human performance F1 of 73.08[9]), but this is still a nontrivial task for large-scale models. When we use Covid-BERT (Müller et al., 2020), a variant of BERT pretrained on Covid-related tweets, we see an improvement of 5.02% over BERT without domain-specific pretraining (Table 9). This indicates greater access to domain-specific data helps

[8] See the paper for the full list.

[9] We count disagreements as being labeled misinformation here, discarding disagreements leads to F1 of 72.68.

Model	F1
Majority Baseline	40.490
Prop-BERT (zero-shot)	34.585
BERT-large (supervised)	78.957
Covid-BERT-large (supervised)	83.974

Table 9: Automatic baseline results (test, classification task).

Model	F1
Majority Baseline	40.490
Prop-BERT (Base)	33.082
Prop-BERT (Granu)	33.931
Prop-BERT (Joint)	**34.585**
Prop-BERT (Mgn)	32.542

Table 10: Zero-shot baseline results (test).

in misinformation detection, even if the veracity of claims stated in the data is unknown.

7 Related Work

Prior work on detection of deceptive writing and misinformation has mostly focused on linguistic features (Ott et al., 2011; Rubin et al., 2016; Rashkin et al., 2017; Wang, 2017; Hou et al., 2019), as well as social network user interactions (Volkova et al., 2017). There has also been work on integration of knowledge graphs (Pan et al., 2018) and framing detection as a NLI task (Yang et al., 2019). Zellers et al. (2019) show the effectiveness of using large-scale neural language modeling to detect machine-generated misinformation. In contrast, we focus on the impact of readers' prior beliefs on *perception* of news reliability. This is related to stance detection (Ghanem et al., 2018; Hardalov et al., 2021), however our pragmatic frames go beyond understanding the stance of a reader and explicitly capture how reader beliefs affect their actions.

Recent work has also highlighted the importance of understanding the impact from misinformation, particularly in health domains (Dharawat et al., 2020; Ghenai and Mejova, 2018), but still focus on traditional methods rather than directly modeling readers' behavior in reaction to news events.

8 Conclusion

In this work, we introduced Misinfo Belief Frames, a pragmatic formalism for understanding reader perception of news reliability. We use these belief frames to construct a corpus of inferences over news headlines that explain potential interpretation and reaction of readers. We show that machine-generated belief frames can used to change perceptions of readers, and while large-scale language models are able to discern between real news and misinformation, there is still room for future work on detection.

Ethical considerations. There is a risk of frame-based machine-generated reader interpretations being misused to produce more persuasive misinformation. However, understanding the way in which readers perceive and react to news is critical in determining what kinds of misinformation pose the greatest threat and how to counteract against its effects.

Broader impact. The rapid dissemination of information online has led to an increasing problem of falsified or misleading news spread on social media like Twitter, Reddit and Facebook (Vosoughi et al., 2018; Geeng et al., 2020). This misinformation can reinforce sociopolitical divisions (Vidgen et al., 2020; Abilov et al., 2021), pose risks to public health (Ghenai and Mejova, 2018), and undermine efforts to educate the public about global crises (Ding et al., 2011). New methods like Misinfo Belief Frames aimed at understanding not only whether readers recognize misinformation but also how they react to it can help in mitigating spread.

Acknowledgements

This research is supported in part by NSF (IIS-1714566), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), DARPA SemaFor program, and Allen Institute for AI.

References

Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. Voterfraud2020: a multimodal dataset of election fraud claims on twitter. In *Proceedings of AAAI 2021*.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of theory of mind*. Psychology Press.

M. Britt, J. Rouet, Dylan Blaum, and K. Millis. 2019. A reasoned approach to dealing with fake news. *Pol-*

icy Insights from the Behavioral and Brain Sciences, 6:101 – 94.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *ArXiv*, abs/2006.00885.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation.

Ding Ding, Edward W. Maibach, Xiaoquan Zhao, Connie Roser-Renouf, and Anthony Leiserowitz. 2011. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

C. J. Fillmore. 1976. Frame semantics and the nature of language *. *Annals of the New York Academy of Sciences*, 280.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2018. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Amira Ghenai and Yelena Mejova. 2018. Fake cures: User-centric modeling of health misinformation in social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA. Association for Computing Machinery.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2020. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles.

S. Hall. 1973. *Encoding and Decoding in the Television Discourse*. Media series: 1972. Centre for Cultural Studies, University of Birmingham.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis- and disinformation identification. *ArXiv*, abs/2103.00242.

Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1):107–112.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ruihong Hou, Verónica Pérez-Rosas, S. Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. *ArXiv*, abs/1909.01543.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake

news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Res Lett. 2017. Fake news threatens a climate literate world. *Nature Communications*, 8(15460):1.

R. Likert. 1932. A technique for the measurement of attitude scales. In *Archives of Psychology, 22 140, 55.*

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore. Association for Computational Linguistics.

Clyde R. Miller. 1939. The techniques of propaganda. *How to Detect and Analyze Propaganda*.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

M. Müller, M. Salathé, and P. Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *ArXiv*, abs/2005.07503.

Matthew C. Nisbet and Dietram A. Scheufele. 2009. What's next for science communication? promising directions and lingering distractions. *American Journal of Botany*, 96(10):1767–1778.

Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web – ISWC 2018 - 17th International Semantic Web Conference, 2018, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 669–683, Germany. Springer Verlag. 17th International Semantic Web Conference, ISWC 2018 ; Conference date: 08-10-2018 Through 12-10-2018.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Unpublished manuscript*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3679–3686, Istanbul, Turkey. European Languages Resources Association (ELRA).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.

Bertie Vidgen, Austin Botelho, David A. Broniatowski, E. Guest, M. Hall, H. Margetts, Rebekah Tromble, Zeerak Waseem, and Scott A. Hale. 2020. Detecting east asian prejudice on social media. *ArXiv*, abs/2005.03909.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *ArXiv*, abs/1907.07347.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Yaoming Zhu, S. Lu, L. Zheng, Jiaxian Guo, W. Zhang, J. Wang, and Y. Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

A

A.1 Annotator Statistics

We provided an optional demographic survey to MTurk workers during annotation. Of the 69 annotators who reported ethnicity, 84% identified as White, 9% as Asian/Pacific Islander, 6% as Hispanic/Latino and 1% as Black/African-American. For self-identified gender, 51% were male and 49% were female. Annotators were generally well-educated, with 70% reporting having a professional degree, college-level degree or higher. Most annotators were between the ages of 25 and 54 (87%). We also asked annotators for news preferences. Twitter, Reddit, CNN, Reuters and the New York Times were reported as the 5 most common news sources. The main task questions presented to annotators are given in Figure 5.

A.2 Reliability of News Definitions

We provide definitions for reliability of news in Table 11.

A.3 Model Hyperparameters

A.3.1 Classification

Supervised classification models are finetuned on our corpus. The BERT model is trained with a learning rate of 1.5e-5, while Covid-BERT is trained with a learning rate of 8e-6. Propaganda detection models are trained using the settings given in (Da San Martino et al., 2019). Examples of predictions made by propaganda detection models are given in Table 13, showing that while these models are accurate at predicting rhetorical techniques, such techniques appear in both real and misinformation news headlines.

A.3.2 Generation

For GPT-2, models are finetuned with a learning rate of 2e-5. We use a learning rate of 5e-5 for T5. For GPT-2 small, we use a batch size of 4 and for GPT-2 large we use a batch size of 16. All T5 models are trained with a batch size of 32. We use beam search with a beam size of 3 for the generation task. Examples generations are provided in Table 12.

Type	Description	Covered by MBF
Misinformation	Misinformation is an umbrella term for news that is false or misleading.	✓
Disinformation	Unlike misinformation, disinformation assumes a malicious intent or desire to manipulate. In our framework, we focus on intent in terms of implications rather than questioning whether or not the writer's intentions were malicious given that it is unclear the extent to which original writers might have known article content was misleading.	Potentially
Fake News	As defined by (Allcott and Gentzkow, 2017), fake news refers to *"news articles that are intentionally and verifiably false, and could mislead readers."* (Golbeck et al., 2018) notes that fake news is a form of hoax, where the content is factually incorrect and the purpose is to mislead. This also overlaps with the definition of disinformation.	Potentially
Propaganda	Propaganda is widely held to be news that is *"an expression of opinion or action by individuals or groups, deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends"* (Miller, 1939). Propaganda is therefore wholly defined in terms of the intent of a writer or group of writers, and may contain factually correct content.	✓
Satire	We refer to articles written with a humorous or ironic intent as "satire." We do not explicitly cover satire in MBF, but it is possible that some misinformation articles began as satire and were misconstrued as real news.	Potentially
Real (Trusted)	We consider this to be news that is factually correct with an intent to inform. We note that while real news is distinct from most of the article types shown here, it can also function as propaganda.	✓

Table 11: Article types based on intention and perceived reliability.

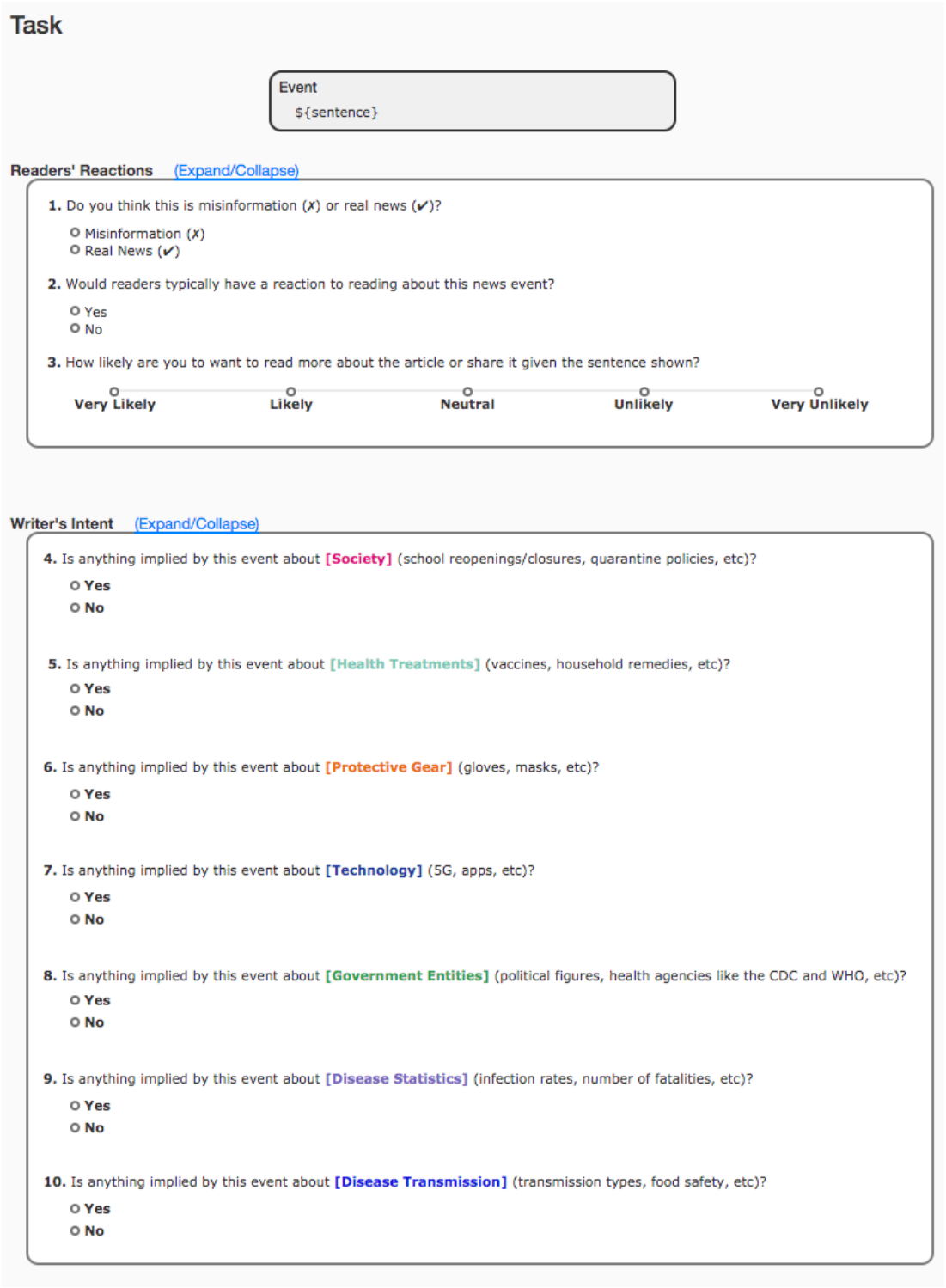

Figure 5: Layout of annotation task for collecting Covid-related commonsense inference data.

News Headline	GPT-2 (small)	GPT-2 (large)	T5-base	T5-large
Clorox wipes shortage will last until 2021.	The writer is implying that covid-19 will last until 2021.	The writer is implying that clorox wipes are out of stock.	The writer is implying that clorox wipes shortage will last until 2021.	The writer is implying that clorox shortage will last until 2021.
Climate' activists seek to ditch polar bear pics as growing populations defy 'emergency' narrative.	The writer is implying that climate change activists want to abandon polar bear pictures.	The writer is implying that climate change is not real.	The writer is implying that climate activists want to ditch polar bear pics as growing populations defy emergency narrative.	The writer is implying that polar bears don't care about climate change.
Eggs and warm meals help avoiding COVID-19.	The writer is implying that there is a cure for covid.	The writer is implying that people should eat more eggs.	The writer is implying that people should eat more warm meals.	The writer is implying that eggs and warm meals help avoid covid-19.
Climate litigation: a government grab for cash.	The writer is implying that the government is trying to manipulate the courts.	unknown intent	The writer is implying that climate litigation is a government grab for cash.	The writer is implying that the government is using climate litigation as a way to make money.

Table 12: Example belief frame writer intents generated from four models in human evaluation.

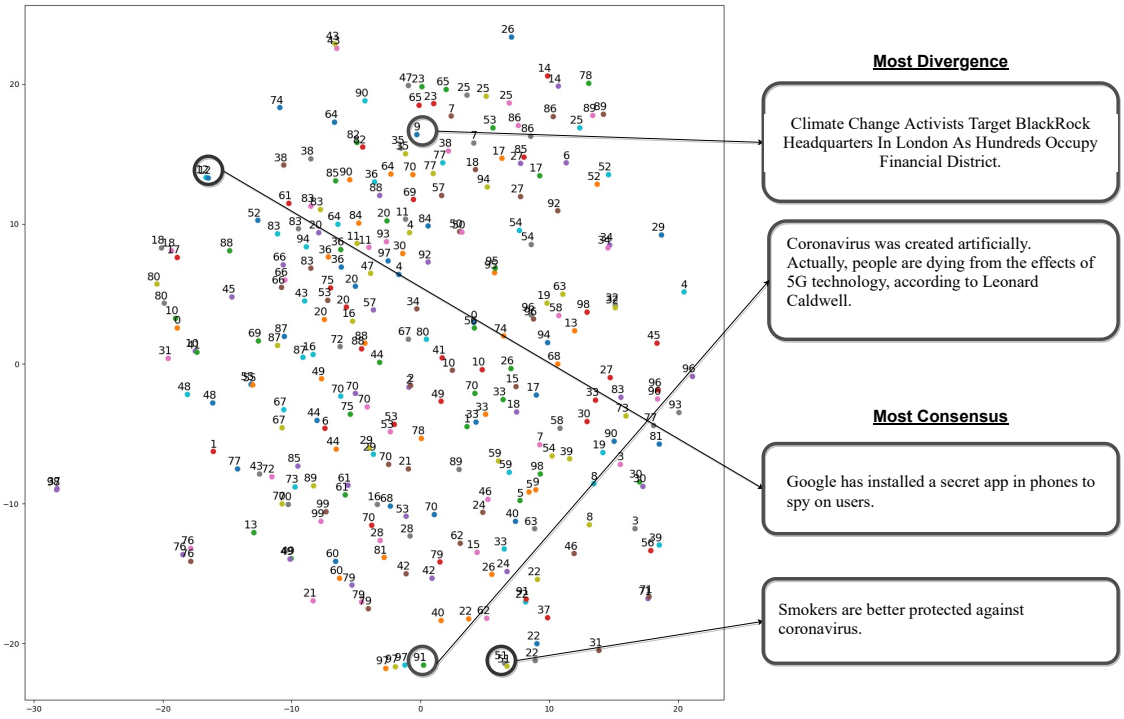

Figure 6: We map BERT embeddings of writer intent inferences associated with 100 news events in the MBF training set using t-SNE dimensionality reduction. Labels show the index of the associated news event. We highlight the top 2 news events that caused the most divergence and consensus respectively in annotated intentions across events.

News Headline	Label
‘Happy corals’ (loaded language) : climate crisis sanctuary teeming with (loaded language) life found off east Africa	Real
Pandemic is a fraud (doubt) and RT-PCR tests aren’t trustworthy (doubt)	Misinfo
International shipping is killing the climate (appeal to fear)	Real
Triple threat of Covid, climate change and conflict (loaded language) has plunged (loaded language) millions into need	Real
Indonesia’s top officials “dancing without obeying health protocols” (loaded language) during the pandemic	Misinfo

Table 13: Examples of real and misinformation news headlines from our dataset. We also provide rhetorical techniques predicted by a pretrained BERT propaganda detection model (Da San Martino et al., 2019) and manually annotated spans that justify labels. As shown here, both real and misinformation news use emotive language and other persuasive techniques to influence readers.