Large Scale Narrative Messaging around Climate Change: A Cross-Cultural Comparison

Haiqi Zhou[♣] **David G Hobson**[♢] **Derek Ruths**[♢] **Andrew Piper**[♣]

[♣]Department of Languages, Literatures and Cultures
[♢]School of Computer Science
McGill University

Abstract

In this study, we explore the use of Large Language Models (LLMs) such as GPT-4 to extract and analyze the latent narrative messaging in climate change-related news articles from North American and Chinese media. By defining "narrative messaging" as the intrinsic moral or lesson of a story, we apply our model to a dataset of approximately 15,000 news articles in English and Mandarin, categorized by climate-related topics and ideological groupings. Our findings reveal distinct differences in the narrative values emphasized by different cultural and ideological contexts, with North American sources often focusing on individualistic and crisis-driven themes, while Chinese sources emphasize developmental and cooperative narratives. This work demonstrates the potential of LLMs in understanding and influencing climate communication, offering new insights into the collective belief systems that shape public discourse on climate change across different cultures.

1 Introduction

Understanding the stories we tell is a key priority for those engaged in climate discourse. Stories serve as a fundamental method through which people exchange information and forge shared understandings about cause and effect in our world, essentially explaining "why things occur" (Todorov, 1981; Herman, 2009). Studies indicate that storytelling can be an effective means of overcoming resistance to new ideas and changing people's intentions to act (Shen et al., 2015; Braddock and Dillard, 2016; Ratcliff and Sun, 2020). Consequently, there is significant interest among climate advocates in leveraging the power and influence of narrative to shift public perspectives (Fløttum and Gjerstad, 2017).

In this paper, we propose a method for surfacing the latent *narrative messaging* of news stories related to climate change. Narrative messaging refers to an overarching, higher-level message that a given story conveys to its readers, one that may be more or less explicit in the body of the story. Narrative messaging is thus akin to broader narratological concepts such as "schemas" (Brewer and Lichtenstein, 1980; Russell and Van Den Broek, 1992), "archetypes" (Campbell, 2008; Frye, 2020), "frames" (Entman, 1993), and "meta-narratives" (White, 2014). Despite addressing narratives at varying levels of abstraction, these models converge on a fundamental premise: stories inherently share common elements, and their selection is orchestrated by higher-level schemas or messages that shape the narrative's construction and interpretation.

For our purposes here, we define narrative messaging as consisting of a story's "moral" or "lesson," i.e. an intrinsic message that readers are intended to take away that transcends the specific details of the story. For example, in a fable such as "The Lion and the Mouse," the moral / message of the story is "a kindness is never wasted." In a news article about climate change that focuses on policy disputes, the moral / message might be "Political compromise is important for finding solutions." While we typically think of story morals as reserved only for short didactic fiction (such as fables), narratologists have long argued that all narratives have implicit value-driven schemas that govern how they are told (Booth, 1998).

As a narrative message, a story moral focuses on the *values* and *intentions* of the storyteller. The moral or lesson of a story is in some sense an answer to the question, "Why is this person telling me this story?" Rather than focus on the specific content of the story, attention to story morals focuses on a more general lesson to be learned ("economic interests and energy security concerns often hinder global consensus on phasing out fossil fuels"). Surfacing such latent narrative values at large scale can facilitate the process of un-

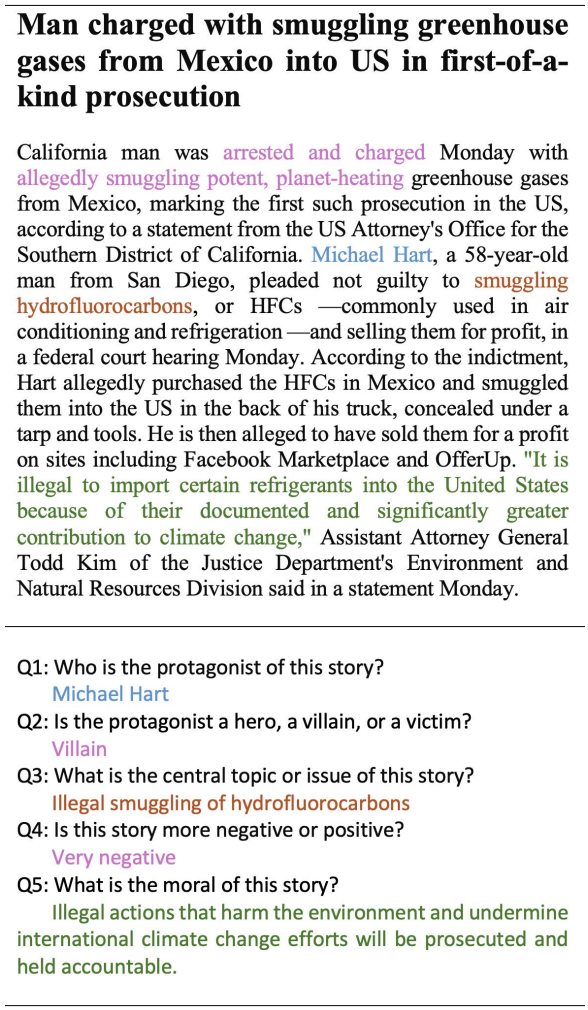


Figure 1: Excerpted version of our prompts on a sample news article.

derstanding collective beliefs around particular societal concerns, their differences across cultures, and any meaningful changes over time.

In our work, we leverage the affordances of large language models (LLMs) like GPT-4 to extract a series of narrative features associated with a given news story up to and including the story's central "moral" or "lesson" (Fig. 1). Such features include the identification of the main agent of the story, the central topic of the story, any antagonist or negative agent, the story's overall valence, along with a free-form moral that is also rendered as a single keyword or phrase.

LLMs present a promising avenue for automating the labeling of story morals. Despite ongoing challenges with hallucinations in LLMs (Xu et al., 2024), their ability to infer underlying meanings—akin to deriving morals from narratives that are often implicit rather than explicitly stated—aligns well with their capabilities. Furthermore, the ubiquity of narratives and narrative-like moral statements on the internet suggests that these concepts are likely well-represented in LLM training datasets. However, this also necessitates a critical evaluation of potential cultural biases embedded within these models.

We proceed first with a review of prior work related to our topic, particularly with respect to prior work on climate communication. We then introduce and validate our method using a combination of automated metrics and human annotation. Finally, we apply our method to the study of a collection of ca. 15,000 news articles written in English and Mandarin that are subsetted by different climate-related topics and different ideological groupings (state / offshore, left / right, see Table 4). We explore techniques of aggregating our story morals to identify salient differences in the larger narrative messaging surrounding issues related to climate change across North American and Chinese-language media.

2 Prior Work

In the field of "environmental communication," the notion of "story morals," as investigated in this study, closely aligns with the concept of "framing," which has been extensively examined in prior research. Framing, sometimes also known as schemas, involves highlighting certain aspects of perceived reality in communicative texts to promote specific problem definitions, causal interpretations, moral evaluations, and recommended solutions (Entman, 1993). Framing's significance is particularly evident in discussions on environmental issues, where climate change represents a "super wicked problem" characterized by the urgent need for action, yet hindered by delayed impacts and insufficient institutional efforts (Levin et al., 2009; Lazarus, 2009; Rodrigo-Alsina, 2019). Consequently, effective framing is essential to bridge the gap between awareness and action in environmental protection (Pan and Kosicki, 1993; Lakoff, 2010; Bushell et al., 2017; Fløttum and Gjerstad, 2017).

A substantial body of literature has explored various frames in climate discourse, such as "social progress," "scientific uncertainty," and "conflict" (Nisbet, 2009; Tong, 2014; Bolsen and Shapiro, 2018), covering different periods and geographic regions (Anderson, 2009). Some studies compare

climate issue framing in various nations (Brossard et al., 2004; Boykoff, 2007; Xie, 2015); while others perform temporal analyses correlating media coverage with significant climatic and political events, such as the COP and Kyoto Protocol (McComas and Shanahan, 1999; Young and Dugas, 2011; Keller et al., 2020; Pan et al., 2021).

Traditionally, these studies have relied on manual coding methodologies, where coders are trained to identify specific elements in articles, such as scientific controversies, typically resulting in a dataset comprising a few hundred articles. Recently, however, automated methods like topic modeling have been adopted in climate framing research to enhance data analysis efficiency (Keller et al., 2020; Rabitz et al., 2021). Our work can be seen as a further extension of such automated approaches, but with a novel focus on the values of narrative messaging by leveraging the affordances of LLMs.

Within the NLP community, the analysis of narrative understanding has gained significant interest (Ranade et al., 2022; Clark et al., 2022). This research encompasses the aim of narrative detection and understanding, across varied contexts, such as literature, social media, and health-care related communication (Ganti et al., 2023; Antoniak et al., 2023). Recent research in this area has also begun exploring the idea of "collective narratives" which involves synthesizing smaller narrative elements (such as tweets, blog entries, or news articles) into overarching narrative frameworks (Zhao et al., 2023; Shahsavari et al., 2020). The research collaborative *Climate Change AI* has been an important early mover in bringing together the ML and NLP fields with climate change concerns (Rolnick et al., 2022).

Our work builds off this prior work by bringing together the approaches and theories of computational narrative understanding towards the goal of studying climate-related communication in different cultural contexts. Instead of applying manually labeled codes to smaller collections, we show how GPT-4 can help surface intrinsic narrative schemas related to the implicit values driving large volumes of news production and that those schemas align with human judgments. In addition to scaling up our understanding of climate-related communication, our approach also shifts the focus from content-related questions (i.e. "what happened") towards more value-driven questions (i.e. "why is this being told?"). Doing so, we argue, can help surface important insights into the collective and often latent belief systems that govern what stories get told and how.

3 Story Morals

3.1 Model

We define a "story moral" as *a general lesson that the narrator wishes to impart to the audience about the world*. While the idea of the "moral" is often associated with a particular ancient narrative tradition,[1] all stories are theoretically governed by a higher-order message that the storyteller wishes to convey, consciously or unconsciously, to guide or reinforce the audience around some belief or a goal. Such messaging is an implicit component of the narrative "schema" that shapes how a story is told and what aspects of the world the narrator chooses to focus on.

While some prefer to use the concept of "framing" to capture these latent narrative schemas around media communication, we prefer the concept of the "story moral" because of the way it draws attention to the behavioral values associated with any given story. The moral of the story is something we can use to guide future actions and thus is explicitly related to behavioral effects (whether it achieves those is a different question).

In order to surface the "story moral" for a given news article, we employ the prompting sequence as described in Table 1. We first extract a summary to help the model focus on key narrative elements. We then identify principal agents, such as the protagonist and antagonist, the central topic of the story, a free-form moral and moral keywords that assume positive and negative valence.

We experiment with two prompt flow frameworks: a *full-context* pipeline, where all prompts are given cumulatively (including the summary and the original text) so that each prior prompt and its answer are included in the subsequent prompt. Alternatively, we experiment with a *simplified framework* with only the summary as the context of each question to reduce cost and compute resources. Fig. 1 illustrates an example output for a sample news story. All prompting exercises were done using GPT-4 (specifically, 0125-

[1]While Aesop's Fables are the best-known genre associated with story morals in the West, similar types of tales exist in both Hindu (Panchatantra) and Buddhist (Jatakas) traditions that date back to around the fifth century BCE indicating the genre's trans-cultural significance.

Category	Prompt
Summary	Can you summarize this story? State your answer as a single paragraph.
Agent	Who is the protagonist of this story? State your answer as a single name.
Agent	Is the protagonist a hero or a villain (i.e., are they portrayed positively or negatively), or are they a victim? You may choose more than one. If none, say none.
Agent	Who is the antagonist of this story? State your answer as a single name. If there is none, say none.
Topic	What is the central topic or issue of this story? State your answer as a single keyword or phrase.
Valence	Is this story more negative or positive? State your answer as a single number between 1 and 5 where 5 = very positive, 1 = very negative, 3 = neutral.
Moral	What is the moral of this story? State your answer as a single sentence.
Moral Keyword Positive	What is the moral of this story? State your answer as a single word or phrase followed by "is a good behavior".
Moral Keyword Negative	What is the moral of this story? State your answer as a single word or phrase followed by "is a bad behavior".

Table 1: Story moral prompts used in this study

preview) through OpenAI's API and using a temperature of zero to minimize output randomness.

3.2 Validation

For the purposes of validation, we use a combination of human assessment and automated metrics. In order to understand GPT's performance across different cultural settings, we use a test dataset of 64 news articles drawn from political news spanning CNN, Al-Jazeera English and four sources of Chinese-language news (described in Table 4). The mean length of documents is 987 words with a minimum of 250 and a maximum of 2,200.

To compare to reference answers, we employed a group of undergraduate students to provide answers to the prompts in Table 1 for each passage (with the summarization question omitted). Six native English-speaking and four native Mandarin-speaking student annotators were hired. Annotators were provided with a codebook of category definitions and examples, and underwent at least one round of practice annotations to affirm consistency of interpretations to the definitions. All human responses were open-responses made in English, and were made independently of each other and from GPT-4.

3.2.1 Human Evaluation

For the more deterministic categories (protagonist and antagonist), we measured direct agreement between GPT and majority / any human annotations. Multiple GPT responses (equal to the number of human annotators) were collected to enhance robustness. Averaged over the sets of GPT responses, we found an average of 49% / 61% (protagonist) and 71% / 97% (antagonist) for the majority / any agreement conditions. Table 5 in the Appendix has the full details.

For each of the more open-ended categories of Moral, Positive Moral Keyword, Negative Moral Keyword, and Topic, we used the following approach involving Amazon's Mechanical Turk platform (AMT) to determine applicability and preferences for human- vs. machine-generated answers. Crowd workers were presented with three options, one from GPT-4 and two that were randomly selected from among the human annotators. The crowd workers were tasked with choosing the "most applicable" and "least applicable" options for each category given the passage text. Crowd workers were given no explicit instructions about what constituted a good or bad option and were given the freedom to select based on their own preferences, so as to avoid any selection bias.

	Agreement (%)			Fleiss κ	GPT Majority (%)	χ^2
	1	2	3			
Most applicable						
Moral	9.38	71.88	18.75	0.05	59.38	$p < 10^{-5}$
Positive Moral	25.00	43.75	31.25	0.16	37.50	$p = 0.14$
Negative Moral	21.88	65.62	12.50	0	34.38	$p = 0.28$
Central Topic	12.50	68.75	18.75	0.07	53.12	$p < 10^{-5}$
Least applicable						
Moral	28.12	65.62	6.25	-0.11	9.38	$p = 0.03$
Positive Moral	6.25	75.00	18.75	0.15	34.38	$p = 0.28$
Negative Moral	6.25	87.50	6.25	0.01	18.75	$p = 0.35$
Central Topic	18.75	71.88	9.38	-0.03	12.50	$p = 0.08$

Table 2: Inter-annotator agreement and GPT selection rate among AMT workers during the human evaluation of the simplified prompt framework. The first 3 columns give the breakdowns of agreement among the annotators; that is, how often 1, 2, or 3 annotators agreed on an option as a percentage of the total number of passages. The fourth column gives the Fleiss κ coefficients for inter-annotator agreement. The fifth column gives the observed rate at which GPT was selected by the majority of AMT workers. The final column gives the p-value for the χ^2 goodness of fit test under the null hypothesis that GPT responses were only selected at random ($p = 1/3$), and therefore had an expected probability of $7/27$ ($\approx 26\%$) of being selected in the majority $\big(P(X \geq 2) = 7/27$ for a binomial random variable X with $n = 3$ (the number of AMT responses) and $p = 1/3\big)$.

To ensure quality responses, we required workers to have a lifetime success rate of more than 95%, and workers had to correctly answer a passage comprehension question to be considered. To partially address new concerns among researchers of crowd workers using ChatGPT to answer the questions, all passages were provided as images.

For passages in Mandarin, English translations of the text were provided to the workers to ensure we were drawing on the same pool of annotators. While the use of translations may have modified the essence of some of the articles, manual inspection of the translations deemed them to be accurate and of high quality. Nevertheless, we acknowledge that the use of translations may introduce potential cultural disparities as the morals are not being rated by individuals from the same cultural background as those who produced the morals. Future work should seek to expand the crowdsourcing validation to more closely study the differences between cultural and linguistic groups. All the same, the results shown below show no significant differences in the overall preferences between the two languages.

In total, responses from three AMT workers were collected for each passage. For evaluating the full-context prompt framework, the full dataset was used, and for the simplified prompt framework, a subset of 32 articles, with an equal split among all news sources, was used.

As seen in Table 2 for the simplified prompt framework, for each category we achieved majority agreement in 75-90% of cases. Inter-rater agreement was extremely low, however, because while two annotators may have chosen a human moral they may have chosen different ones. Nevertheless, we observe that the GPT morals and central topics were selected well above a random baseline, and the positive and negative morals were no worse than random, as indicated by a χ^2 goodness of fit test. In no case did crowd-workers preferentially choose GPT answers as "least applicable" in a statistically significant way.

Table 7 in the Appendix shows comparable, albeit slightly better, results for the full-context framework. This can likely be attributed to the fact that the exclusive use of summaries in the simplified prompt framework occasionally omits elements that are pertinent to constructing a good moral. This notwithstanding however, these morals from GPT were still favored over the human morals by the AMT workers. As neither prompt workflow showed a negative preference for GPT responses, we elected to employ the simplified framework for our full analysis.

	human-human	human-GPT	GPT-GPT	U-test
Moral				
Rouge-1	0	8.00	58.62	$p < 10^{-4}$
Rouge-L	0	7.41	51.61	$p < 10^{-4}$
GloVe	55.01	64.74	91.03	$p < 10^{-4}$
STSb-MPNet	25.89	38.83	85.11	$p < 10^{-4}$
NLI-MPNet	33.17	46.63	89.58	$p < 10^{-4}$

Table 3: Median similarity (out of 100) of pairwise morals between the different groups of annotators in the validation dataset. P-values reflect a Mann-Whitney U-test (rank-sum test) with a null hypothesis that the human-human and human-GPT distributions are the same. GPT morals are from the full-context prompt framework.

3.2.2 Automated Validation

For the automated validation of morals and central topics, we used a group of semantic textual similarity (STS) scores relevant to our human annotations These included ROUGE-1 and ROUGE-L (Lin, 2004), and cosine similarity using pretrained embedding models from the SentenceTransformers library (Reimers and Gurevych, 2019). ROUGE-based metrics were implemented using the HuggingFace library and all embedding models were implemented using the SentenceTransformers library (Reimers and Gurevych, 2019). For these latter models, the specific models included averaged GloVe word embeddings (6b-300d) (Pennington et al., 2014), and the stsb-mpnet-base-v2 and nli-mpnet-base-v2 models (Song et al., 2020).

Table 3 shows a complimentary picture using the automated evaluation metrics to compare the human responses to those from GPT on a single example of our categories (e.g. moral) (see Table 8 for full details). All pairs of responses were compared for a given story which were then combined to create a single distribution of pairwise similarity between annotations. The human-human column indicates the median of the distribution comparing only human to human responses, the human-GPT column indicates the median in only comparing human responses to GPT responses, and the GPT-GPT column compares GPT responses to other GPT responses for replicability. As noted above, the number of GPT responses was chosen to be the same as the number of human annotations.

Overall, we find that the semantic variation between human responses is higher (i.e. exhibits lower similarity) than that between GPT and human responses across all metrics suggesting that GPT is decently approximating an aggregate human point of view. As Table 8 in the Appendix indicates, positive and negative morals are exceptions with respect to some metrics, however the differences are small even when statistically significant. We also note that GPT exhibits very high similarity scores to itself on multiple runs for the same text though these still exhibit some variation. Central Topic is an exception where GPT always repeated its answer verbatim.

Finally for valence, the average standard deviation between human-only responses was 0.68 across all passages, compared to 0.66 when introducing GPT responses, thus showing good compatibility between the responses. Details can be found in Table 6 in the Appendix.

4 Analysis

4.1 Data

Our dataset for this study comprises approximately 15,000 news articles sourced from Dow Jones Factiva, segmented by language into Mandarin and English (see Table 4). The articles are filtered by length to fall between 250 and 2,500 words and published during the calendar year 2023. They were selected using five key terms associated with environmental issues: "climate change" (气候变化/氣候變遷 for Taiwanese sources), "pollution" (污染), "carbon emissions" (碳排放), "renewable energy" (再生能源), and "sustainability." For "sustainability," we utilized the Chinese term huan bao (环保), which connotes "environment-friendly." This term was chosen because it is prevalent in Chinese environmental discourse, whereas its direct translations occur less often in English contexts. This strategic choice of keywords ensures that our dataset robustly represents significant environmental discussions within each language's media landscape.

Source Name	Region	Class	No. Articles
People's Daily (人民日报)	Mainland China	State	3748
Global Times (环球时报)	Mainland China	State	1167
Ming Pao (明報)	Hong Kong	Offshore	1315
Liberty Times (自由時報)	Taiwan	Offshore	2028
CNN	U.S.	Liberal	1730
The New York Times	U.S.	Liberal	2111
The Wall Street Journal	U.S.	Conservative	1499
The Globe and Mail	Canada	Conservative	1499

Table 4: Summary of news sources and their characteristics

4.2 Valence

A linear regression analysis reveals that both region (North American vs. Chinese sources) and ideology (Liberal vs. Conservative, state vs. offshore) significantly affect average valence. Specifically, being in North America is associated with a lower valence (M=2.87) compared to Chinese-speaking regions (M=3.45). Mainland Chinese sources are in turn more positive (M=3.56) than offshore sources (M=3.29) most likely due to censorship and the decline of critical voices in state media (Guo et al., 2023). While statistically significant differences exist between Conservative (M=2.93) and Liberal (M=2.83) sources, they are the most similar.

4.3 Distinctive Values

Next we use the Fightin' Words model for lexical feature selection (Monroe et al., 2009) to identify salient differences in moral keywords between our cultural subsets. As can be seen in Fig. 2 (top), Chinese and North American sources exhibit different moral focuses: Chinese sources emphasize promoting "international cooperation" and "sustainable development," while North American sources concentrate on "addressing" and "adapting" to climate change, reflecting a potentially stronger sense of crisis, but also individualism. China's call for international cooperation corresponds to findings from previous studies that climate change has transitioned from a concern primarily addressed by developed nations to a global issue where China is actively engaged and takes a proactive stance (Pan et al., 2021). The Chinese emphasis on "promoting" solutions contrasts with a North American emphasis on reacting to and accepting consequences. The developmental framing of Chinese-language news as a whole suggests a far more proactive stance than the North American one.

Within North America (Fig. 2 (middle)), Conservative outlets place a much stronger emphasis on markets, investments, and economic and financial issues. This reflects the 2019 Pew Research Center survey that shows Conservative Republicans being skeptical towards climate policies — a majority (62%) of this group says these policies hurt the economy (Hefferon, 2019). We also note that Conservative media view climate change through the "adapting" lens, while Liberal media emphasizes the more pro-active "addressing" lens. Conservative messaging focuses on climate change as something to be lived with and accommodated, whereas Liberal media tends to view the problem more holistically as impacting human health and the natural environment.

In Chinese-language outlets, state media centers its discourse on development, underpinned by values of sustainability and environmental consciousness. Whereas offshore (Hong Kong and Taiwan) media value "sustainability" as an end in itself, state media focus on "sustainable" (and "green") as a modifier of development. This also aligns with findings from previous studies that climate change was no longer viewed solely as an environmental obstacle to socio-economic development in China, but rather as a manageable challenge that drives and creates opportunities for economic growth (Pan et al., 2021). In contrast, offshore Chinese media tend to focus more on the individual level, emphasizing everyday life with keywords such as "community," "recycling," and "local," which can partly be attributed to the smaller size of Hong Kong and Taiwan.

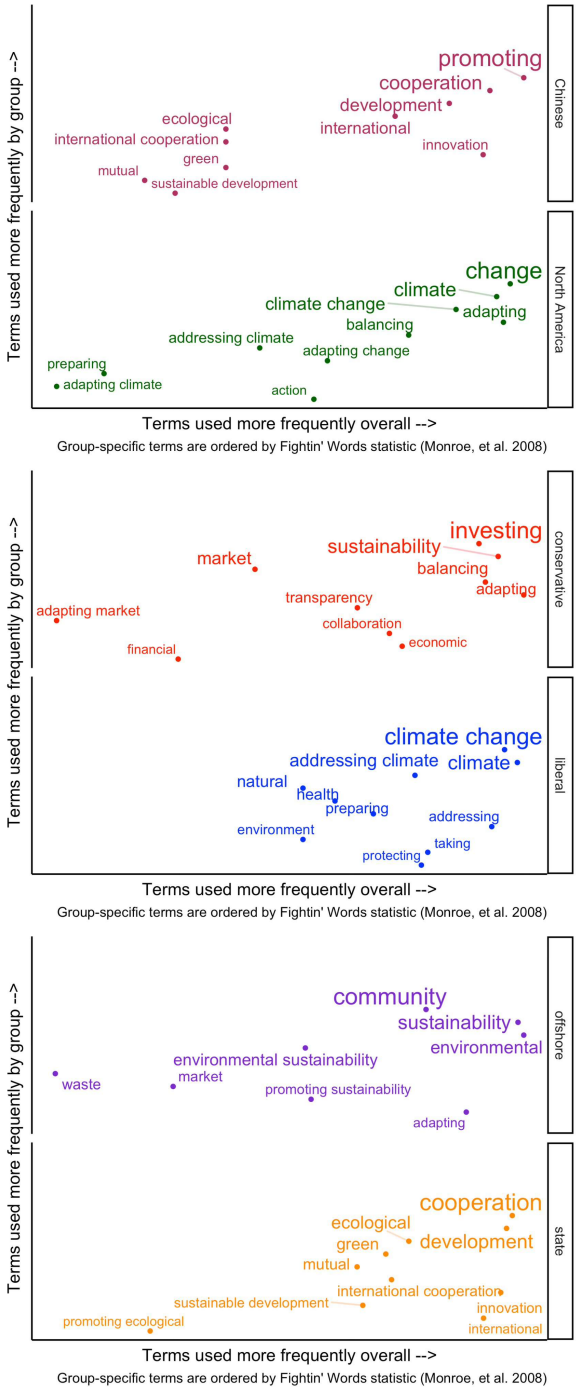
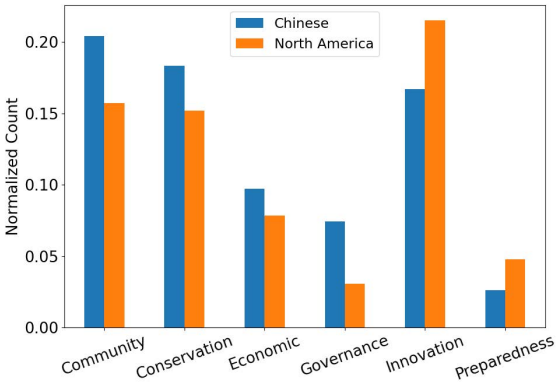
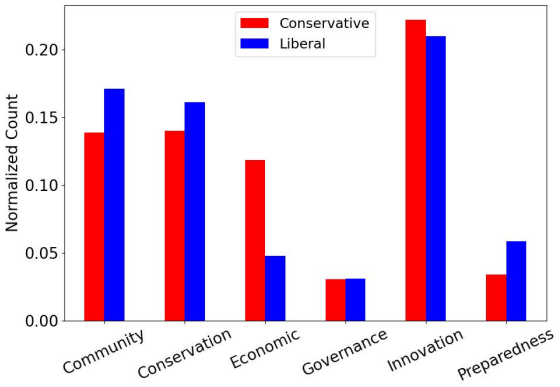


Figure 2: Fightin' Words illustration of distinctive positive moral keywords between North American and Chinese-language news sources.



(a) Chinese-language sources vs. North-American sources



(b) Liberal sources vs. Conservative sources

Figure 3: Open-coding exercise using GPT to identify six salient issues under which positive morals are grouped.

4.4 GPT-Assisted Open Coding

One of the challenges of the Fightin' Words approach is generalizing about larger trends to which individual morals align. To address this limitation, we engaged ChatGPT-4 in the process of open coding (Strauss and Corbin, 2004). We first gave GPT truncated lists of the most distinctive moral keywords based on the Fightin Words method and asked it to devise 5-6 categories that best represented the terms. These categories were reviewed for appropriateness by the authors, headings were adjusted for brevity, and then a list of the 200 most distinctive keywords was inputted with the request of assigning them to their respective categories (with no overlap). Not all words were assigned and not all assignations were agreed upon by the authors and so a round of manual adjustments were undertaken.

The final rubric consisted of six categories: Community & Justice, Conservation, Economy, Preparedness, Governance, and Innovation. Each

category has a unique set of keywords, for which we extracted the normalized counts and compared North American and Chinese sources as well as Conservative and Liberal ones. The results, shown in Fig. 3, align with our findings from Fightin' Words. Specifically, Chinese media shows a significantly stronger focus on issues related to governance and community participation, while North American sources emphasize the importance of facing risks and promoting innovation.

Fig. 3b reveals distinct differences in keyword emphasis between Conservative and Liberal sources. Liberal sources have higher normalized counts for Community, Conservation, and Preparedness, suggesting a strong focus on social justice issues and environmental protection. In contrast, Conservative sources emphasize Economic and Innovation categories more, indicating a prioritization of economic growth and technological advancement.

Previous studies have found significant differences in frames employed by Conservative and Liberal media. The Wall Street Journal, for example, has been shown to use more frames emphasizing negative economic consequences, suggesting that proposed solutions are unlikely to be effective, and highlighting political conflict (Feldman et al., 2017). However, based on our keyword analysis and valence comparison, we can speculate that recent articles from Conservative media continue to prioritize the economy while addressing environmental topics, albeit with a more positive frame.

5 Conclusion

Our study illustrates a workflow that can be applied to understand the narrative messaging of values around climate change across different cultural contexts. We show the robust validity of large language models like GPT-4 to derive high-level conceptual information about narratives in strongly different cultural and linguistic contexts. In particular, we surface key "values" associated with climate-related news, with Chinese media focusing on a more "developmental" approach compared to a more "adaptive" approach on the North American side.

Such workflows will be an important dimension towards scaling up our understanding of climate communication. While we focus on surfacing implicit narrative values around climate change from the bottom-up, researchers can also use our method to test more specific hypotheses and content-related "frames" determined in advance. As we discuss in the limitations section, more work is necessary to better understand the biases or norms implicit in LLMs. Nevertheless, we believe LLMs are going to be an important tool in understanding, interpreting, and influencing climate communication moving forward.

Limitations

Understanding climate communication at a large scale poses a number of research challenges. While we look at eight different news outlets across two different national and ideological contexts, wider sampling and including more cultures will be an essential next step as we scale-up this work. Our sample is also limited by the keyword filtering such that future work might explore other ways of identifying a fuller sample of climate-related communication.

While we observe strong levels of human-judged validity in terms of the appropriateness of GPT-generated morals, more work can be done to understand intercultural differences surrounding the perception of narrative messaging. Additionally, it is important to note that while GPT-generated content is marked by high levels of semantic relatedness across multiple runs of the same queries there is still some observed variability even when the temperature is set to 0 making exact replication unlikely.

Another important limitation here is the dependence on GPT-4 as the primary LLM. Future work will want to explore the behavior of other large frontier models as well as the ability to employ smaller, specialized models to avoid the large carbon footprint of the bigger models.

Finally, our results also offer numerous avenues for further exploration beyond the methods presented in this paper. Future research can employ different clustering methods, compare these methods with existing methodologies like topic modeling, and apply automated coding by LLMs to answer questions that were previously human-coded. This will enable a more direct investigation into the nuances of surrounding narrative messaging as it relates to climate change.

Acknowledgements

We would like to thank the reviewers for their constructive feedback and suggestions. This research

was supported by the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council of Canada.

References

Alison Anderson. 2009. Media, Politics and Climate Change: Towards a New Research Agenda. *Blackwell Publishing Ltd Sociology Compass*, 32:166–1821751.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2023. Where do people tell stories online? story detection across online communities. *arXiv preprint arXiv:2311.09675*.

Toby Bolsen and Matthew A. Shapiro. 2018. The US News Media, Polarization on Climate Change, and Pathways to Effective Communication. *Environmental Communication*, 12(2):149–163.

Wayne C Booth. 1998. Why ethical criticism can never be simple. *Style*, pages 351–364.

Maxwell T. Boykoff. 2007. Flogging a Dead Norm? Newspaper Coverage of Anthropogenic Climate Change in the United States and United Kingdom from 2003 to 2006. *Area*, 39(4):470–481.

Kurt Braddock and James Price Dillard. 2016. Meta-analytic evidence for the persuasive effect of narratives on beliefs, attitudes, intentions, and behaviors. *Communication monographs*, 83(4):446–467.

William F Brewer and Edward H Lichtenstein. 1980. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*.

Dominique Brossard, James Shanahan, and Katherine McComas. 2004. Are Issue-Cycles Culturally Constructed? A Comparison of French and American Coverage of Global Climate Change. *Mass Communication and Society*, 7(3):359–377.

Simon Bushell, Géraldine Satre Buisson, Mark Workman, and Thomas Colley. 2017. Strategic narratives in climate change: Towards a unifying narrative to address the action gap on climate change. *Energy Research & Social Science*, 28:39–49.

Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

Elizabeth Clark, Faeze Brahman, and Mohit Iyyer. 2022. Proceedings of the 4th workshop of narrative understanding (wnu2022). In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Lauren Feldman, P. Sol Hart, and Tijana Milosevic. 2017. Polarizing news? representations of threat and efficacy in leading us newspapers' coverage of climate change. *Public Understanding of Science*, 26(4):481–497. PMID: 26229010.

Kjersti Fløttum and Øyvind Gjerstad. 2017. Narratives in climate change discourse. *WIREs Climate Change*, 8(1):e429.

Northrop Frye. 2020. *Anatomy of criticism: Four essays*, volume 69. Princeton University Press.

Achyutarama Ganti, Eslam Ali Hassan Hussein, Steven Wilson, Zexin Ma, and Xinyan Zhao. 2023. Narrative style and the spread of health misinformation on twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4266–4282.

Jing Guo, Xiaoyun Huang, and Kecheng Fang. 2023. Authoritarian environmentalism as reflected in the journalistic sourcing of climate change reporting in china. *Environmental Communication*, 17(5):502–517.

Cary Funk and Meg Hefferon. 2019. U.S. Public Views on Climate and Energy.

David Herman. 2009. *Basic elements of narrative*. John Wiley & Sons.

Tobias R. Keller, Valerie Hase, Jagadish Thaker, Daniela Mahl, and Mike S. Schäfer. 2020. News Media Coverage of Climate Change in India 1997–2016: Using Automated Content Analysis to Assess Themes and Topics. *Environmental Communication*, 14(2):219–235.

George Lakoff. 2010. Why it Matters How We Frame the Environment. *Environmental Communication*, 4(1):70–81.

Richard Lazarus. 2009. Super Wicked Problems and Climate Change: Restraining the Present to Liberate the Future. *Georgetown Law Faculty Publications and Other Works*.

K. Levin, B. Cashore, Steven Bernstein, and G. Auld. 2009. Playing it forward: Path dependency, progressive incrementalism, and the "Super Wicked" problem of global climate change. *IOP Conference Series: Earth and Environmental Science*, 6(50):502002.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Katherine McComas and James Shanahan. 1999. Telling Stories About Global Climate Change: Measuring the Impact of Narratives on Issue Cycles. *Communication Research*, 26(1):30–57.

Burt Monroe, Michael Colaresi, and Kevin Quinn. 2009. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16.

Matthew C. Nisbet. 2009. Communicating Climate Change: Why Frames Matter for Public Engagement. *Environment: Science and Policy for Sustainable Development*, 51(2):12–23.

Yeheng Pan, Michaël Opgenhaffen, and Baldwin Van Gorp. 2021. China' s Pathway to Climate Sustainability: A Diachronic Framing Analysis of People' s Daily' s Coverage of Climate Change (1995–2018). *Environmental Communication*, 15(2):189–202.

Zhongdang Pan and Gerald M. Kosicki. 1993. Framing analysis: An approach to news discourse. *Political Communication*, 10(1):55–75.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Florian Rabitz, Audron Teleien, and Eimant Zolubien. 2021. Topic modelling the news media representation of climate change. *Environmental Sociology*, 7(3):214–224.

Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. 2022. Computational understanding of narratives: A survey. *IEEE Access*, 10:101575–101594.

Chelsea L Ratcliff and Ye Sun. 2020. Overcoming resistance through narratives: Findings from a meta-analytic review. *Human Communication Research*, 46(4):412–443.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Miquel Rodrigo-Alsina. 2019. Talking about climate change: The power of narratives. In *Climate Change Denial and Public Relations*. Routledge.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96.

Robert L Russell and Paul Van Den Broek. 1992. Changing narrative schemas in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 29(3):344.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Fuyuan Shen, Vivian C Sheer, and Ruobing Li. 2015. Impact of narratives on persuasion in health communication: A meta-analysis. *Journal of advertising*, 44(2):105–113.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Anselm L Strauss and Juliet Corbin. 2004. Open coding. *Social research methods: A reader*, pages 303–306.

Tzvetan Todorov. 1981. *Introduction to poetics*, volume 1. U of Minnesota Press.

Jingrong Tong. 2014. Environmental Risks in Newspaper Coverage: A Framing Analysis of Investigative Reports on Environmental Problems in 10 Chinese Newspapers. *Environmental Communication*, 8(3):345–367.

Hayden White. 2014. *Metahistory: The historical imagination in nineteenth-century Europe*. JHU Press.

Lei Xie. 2015. The Story of Two Big Chimneys: A Frame Analysis of Climate Change in US and Chinese Newspapers. *Journal of Intercultural Communication Research*, 44(2):151–177.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Nathan Young and Eric Dugas. 2011. Representations of Climate Change in Canadian National Print Media: The Banalization of Global Warming. *Canadian Review of Sociology/Revue canadienne de sociologie*, 48(1):1–22.

Wanying Zhao, Fiona Guo, Kristina Lerman, and Yong-Yeol Ahn. 2023. Discovering collective narratives shifts in online discussions. *arXiv preprint arXiv:2307.08541*.

Appendix

	Agree with majority	Any agreement	No agreement	Average human popular vote
Protagonist	49.22	61.33	38.67	62.89
Antagonist	71.88	96.88	3.12	69.99

Table 5: Percent agreement of GPT responses with human responses for protagonists and antagonists. Values represent the average agreement over all GPT responses collected (equal to the number of human annotators).

	human-human	human-GPT	GPT-GPT
Valence (average standard deviation)	0.68	0.66	0.08
Protagonist Type (average Jaccard index)	44.05	42.46	89.06

Table 6: Average standard deviations in valence responses and Jaccard index (Jaccard is out of 100) in protagonist type between the different distributions of responses. The human-human column compares all pairs of responses (to the same passage) among the human annotators, the human-GPT group compares all pairs of responses between human and GPT responses, and the GPT-GPT column compares all responses between GPT responses. The number of GPT responses was always chosen to be equal to the number of human annotators.

	Agreement (%)			Fleiss κ	GPT Majority (%)	χ^2
	1	2	3			
Most applicable						
Moral	14.1	50.0	35.9	0.03	73.44	$p < 10^{-5}$
Positive Moral	17.2	64.1	18.8	-0.01	57.81	$p < 10^{-5}$
Negative Moral	18.8	67.2	14.1	0	51.56	$p < 10^{-5}$
Central Topic	9.4	48.4	42.2	0.15	73.44	$p < 10^{-5}$
Least applicable						
Moral	15.6	60.9	23.4	0.06	7.81	$p < 10^{-3}$
Positive Moral	18.8	65.6	15.6	0.05	12.50	$p = 0.01$
Negative Moral	26.6	57.8	15.6	0.01	17.19	$p = 0.11$
Central Topic	6.3	62.5	31.3	0.22	7.81	$p < 10^{-3}$

Table 7: Inter-annotator agreement and GPT selection rate among AMT workers during the human evaluation of the full-context prompt framework. The first 3 columns give the breakdowns of agreement among the annotators; that is, how often 1, 2, or 3 annotators agreed on an option as a percentage of the total number of passages. The fourth column gives the Fleiss κ coefficients for inter-annotator agreement. The fifth column gives the observed rate at which GPT was selected by the majority of AMT workers. The final column gives the p-value for the χ^2 goodness of fit test under the null hypothesis that GPT responses were only selected at random ($p = 1/3$), and therefore had an expected probability of $7/27$ ($\approx 26\%$) of being selected in the majority $\big(P(X \geq 2) = 7/27$ for a binomial random variable X with $n = 3$ (the number of AMT responses) and $p = 1/3\big)$.

	human-human	human-GPT	GPT-GPT	U-test
Moral				
Rouge-1	0	8.00	58.62	$p < 10^{-4}$
Rouge-L	0	7.41	51.61	$p < 10^{-4}$
GloVe	55.01	64.74	91.03	$p < 10^{-4}$
STSb-MPNet	25.89	38.83	85.11	$p < 10^{-4}$
NLI-MPNet	33.17	46.63	89.58	$p < 10^{-4}$
Positive Moral				
Rouge-1	0	0	57.14	$p < 10^{-3}$
Rouge-L	0	0	57.14	$p < 10^{-3}$
GloVe	31.07	40.73	81.84	$p < 10^{-4}$
STSb-MPNet	27.12	26.66	76.73	$p = 0.62$
NLI-MPNet	38.58	35.57	81.70	$p = 0.01$
Negative Moral				
Rouge-1	0	0	66.67	$p = 0.03$
Rouge-L	0	0	66.67	$p = 0.03$
GloVe	24.69	29.45	84.28	$p = 0.15$
STSb-MPNet	20.87	18.85	86.10	$p < 10^{-3}$
NLI-MPNet	30.84	26.14	86.54	$p < 10^{-4}$
Central Topic				
Rouge-1	0	11.11	100	$p < 10^{-4}$
Rouge-L	0	11.11	100	$p < 10^{-4}$
GloVe	44.17	55.79	100	$p < 10^{-4}$
STSb-MPNet	33.78	41.45	100	$p < 10^{-3}$
NLI-MPNet	39.63	48.61	100	$p < 10^{-4}$

Table 8: Median similarity (out of 100) between the different groups of annotators in the validation dataset. The human-human column compares all pairs of responses (to the same passage) among the human annotators; the human-GPT group compares all pairs of responses between human and GPT responses; and the GPT-GPT column compares all responses between GPT responses. The p-values are calculated using a Mann-Whitney U-test (rank-sum test) with a null hypothesis that the human-human and human-GPT distributions are the same.