Applying the Character-Role Narrative Framework with LLMs to Investigate Environmental Narratives in Scientific Editorials and Tweets

Francesca Grasso
University of Turin
Dep. of Computer Science
Turin, Italy
fr.grasso@unito.it

Stefano Locci
University of Turin
Dep. of Computer Science
Turin, Italy
stefano.locci@unito.it

Manfred Stede
University of Potsdam
Department of Linguistics
Potsdam, Germany
stede@uni-potsdam.de

Abstract

Communication aiming to persuade an audience uses strategies to frame certain entities in 'character roles' such as hero, villain, victim, or beneficiary, and to build narratives around these ascriptions. The *Character-Role Framework* is an approach to model these narrative strategies, which has been used extensively in the Social Sciences and is just beginning to get attention in Natural Language Processing (NLP). This work extends the framework to scientific editorials and social media texts within the domains of ecology and climate change. We identify characters' roles across expanded categories (human, natural, instrumental) at the entity level, and present two annotated datasets: 1,559 tweets from the *Ecoverse* dataset and 2,150 editorial paragraphs from *Nature & Science*. Using manually annotated test sets, we evaluate four state-of-the-art Large Language Models (LLMs) (GPT-4o, GPT-4, GPT-4-turbo, LLaMA-3.1-8B) for character-role detection and categorization, with GPT-4 achieving the highest agreement with human annotators. We then apply the best-performing model to automatically annotate the full datasets, introducing a novel entity-level resource for character-role analysis in the environmental domain.

1 Introduction

There is a long history in the literature demonstrating how stories are central to how humans understand and communicate about the world, with language playing a key role in constructing and delivering specific messages (Armstrong and Ferguson, 2010; Polkinghorne, 1988). This is particularly important when examining linguistic representations of climate change (CC) and environmental issues (Wolters et al., 2021; Stibbe, 2015, 2021; Jones et al., 2022). Studies from various disciplines, such as ecolinguistics (Fill and Muhlhausler, 2006; Alexander and Stibbe, 2014), political and social sciences (Nerlich et al., 2010; Grundmann and Krishnamurthy, 2010), have demonstrated how environmental and CC narratives are crucial in understanding how individuals and entities like governments and media interpret and relate to ecological issues and the natural world and as a consequence, how they behave towards them (Fløttum and Gjerstad, 2017). The linguistic construction of entities as social actors in these 'stories' can reveal the author's *framing* choices and communicative intent (Hulme, 2015). For instance, "Expanding oil drilling operations will boost economic growth and create thousands of jobs in struggling communities" frames oil drilling positively, emphasizing economic benefits while downplaying environmental concerns. The *Character-Role Framework* – introduced by Gehring and Grigoletto (2023) and drawing on the Narrative Policy Framework (NPF) (Jones and McBeth, 2010; Jones, 2018) – is based on the premise that framing entities in specific roles (hero, villain, victim, beneficiary) is key to understanding a narrative's intent and potential effects. While prior work using this framework has primarily appeared in social and political sciences (Bergstrand and Jasper, 2018; Wolters et al., 2021), its adaptation to Natural Language Processing (NLP) tasks remains limited. Existing studies have either focused on policy narratives (Gehring and Grigoletto, 2023) or explored related tasks like character-role extraction (Stammbach et al., 2022) focusing on a higher-level analyses (e.g. at the paragraph-level). Moreover, Frermann et al. (2023) extended framing analysis to the document-level by integrating narrative media framing with entity roles. In this paper, we extend and adapt the Character-Role Framework to investigate CC and environmental narratives. Specifically, we focus on identifying characters across three categories (human, instrumental, natural) and four roles (hero, villain, victim, beneficiary) at the *entity level*, applying the framework to both scientific editorials and social media. To achieve this, we first created two

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 49–67
July 31, 2025 ©2025 Association for Computational Linguistics

manually annotated test sets: (i) 150 CC-related scientific editorial paragraphs from *Nature* and *Science* (Stede et al. (2023)) and (ii) 300 tweets from the *Ecoverse* dataset (Grasso et al., 2024a), covering a wide range of environmental topics. Characters were identified and roles assigned based on linguistic cues, subsequently categorized as human, instrumental, or natural - a category newly introduced in this work. We evaluated four Large Language Models (LLMs) (GPT-4o, GPT-4, GPT-4-turbo, and Llama-3.1-8B) as additional annotators to measure alignment with human annotations. We subsequently applied the best-performing models to larger datasets, resulting in **1,559** tweets and **2,150** editorial paragraphs annotated for further analysis. Our contributions are fivefold: (i) We offer a novel approach for analyzing CC and environmental texts using the Character-Role Framework across social media and scientific editorials. (ii) We extend the framework by adopting a bottom-up entity-level approach and introducing the "natural" character category. (iii) We release two new annotated datasets for narrative analysis in this domain. (iv) We evaluate LLMs on the entity-level character-role detection and categorization tasks, marking a significant advancement for this framework in NLP[1]. (v) We conduct a qualitative error analysis of model misclassifications and provide preliminary insights into emerging narrative patterns.

2 Theoretical Background

The Character-Role Framework, introduced by Gehring and Grigoletto (2023), builds on foundational work in climate change (CC) narratives, such as those by Fløttum and Gjerstad (2017, 2013)[2]. Their theoretical and methodological framework applies the concept of policy narrative from the Narrative Policy Framework (NPF) (Jones, 2018) to CC discourse, recognizing the role of 'stories' used to communicate and discuss CC issues in shaping opinions and behaviors. This aligns with traditions in ecolinguistics and ecocriticism, which emphasize the importance of studying how language shapes perceptions, behaviors, and actions regarding environmental issues (Stibbe, 2015, 2021; Fill and Muhlhausler, 2006). The NPF adopts a structuralist approach to narrative and posits that they can be generalizable and have an identifiable structure and measurable elements (e.g., characters, setting) (Jones et al., 2022). Among the narrative components, "characters" play a prominent role in that they determine and are determined by the "plot". Key character roles include: (i) *victims*, who are harmed or at risk of harm; (ii) *villains*, responsible for causing harm; (iii) *heroes*, who work to resolve the harm; and (iv) *beneficiaries*, who gain from the events described. Gehring and Grigoletto (2023) further distinguish characters as either human (individuals or entities made up of people) or instrumental (abstract entities like policies or laws). Narratives are categorized as either simple (involving a single character) or complex (involving multiple characters). This framework theory and the assumption on which it is based can be easily extended and applied to other communicative units, such as social media and scientific communication, as each communicative act entails a rather specific (more or less overt) communicative intention. In environmental narratives, any entity can take on the role of a character (Gehring and Grigoletto, 2023): framing institutions, natural entities, or even concrete objects in specific roles can influence perceptions, preferences, and actions towards these entities or events. For example, Kuha (2017) highlights the crucial role of linguistic cues in shaping how language users represent both themselves and other social actors, especially in terms of agency and responsibility.

3 Related Work

Character Roles, Narratives, and Related Tasks The study of environmental narratives has traditionally been rooted in the Humanities, for instance in fields such as Ecolinguistics and Ecocriticism (Alexander and Stibbe, 2014; Stibbe, 2015, 2021). Much research specifically on climate change (CC) narratives has been situated in the Political and Social Sciences, sometimes using the Narrative Policy Framework (NPF) to analyze topics like the political discourse on environment (Peterson, 2021), COVID-19 narratives (Peterson et al., 2021) or economy reports (Goldberg-Miller and Skaggs, 2022). There is only little work in the NLP field on ecology-oriented corpora so far (Grasso et al., 2024a; Bosco et al., 2023), but CC-related topics

[1] The code, the complete set of prompts and the anonymized datasets are available in this repository: https://github.com/stefanolocci/Character-Role-Narrative-Framework-LLMs.

[2] We are aware of the huge body of literary-science-oriented research on narratology, in the tradition of Propp and Bakhtin, but for reasons of space we do not make a comparison here but limit ourselves to the more social-science-related view.

have recently gained traction within the NLP community (Stede and Patz, 2021; Grasso et al., 2024b; Stammbach et al., 2024). Recent work has explored character-role extraction in NLP, such as identifying *villain* roles using rule-based approaches and BERT (Klenner et al., 2021) or extracting character roles via zero-shot question-answering (Stammbach et al., 2022). In the context of CC narratives, Gehring and Grigoletto (2023) applied the character-role framework to US policy discourse on Twitter, focusing on economic narratives and a narrow character set at the tweet level. Beyond social-media and short-text analysis, Zhou et al. (2024) used LLMs to analyze CC narratives extracting latent moral messaging from North American and Chinese news. Our work extends this line of research by introducing entity-level annotations across multiple roles and categories, thus offering a broader and more fine-grained analysis of environmental narratives. The character-role task bears similarities to the field of entity-level sentiment detection (Rønningstad et al., 2023), where linguistic indicators like polarity-inducing verbs or modifiers are used to determine whether a certain entity is being portrayed positively or negatively.

LLMs and the CC/Environment Domain In the wider field of applying NLP to the CC and environment domain, notable contributions include Bulian et al. (2023) and Zhu and Tiwari (2023), who propose evaluation frameworks for analyzing LLM responses to CC topics. Koldunov and Jung (2024) developed a prototype tool using LLMs to provide localized climate-related data, while Leippold et al. (2024) created an AI tool for fact-checking CC claims utilizing LLMs. Thulke et al. (2024) introduced a family of domain-specific LLMs designed to synthesize interdisciplinary research on CC. Grasso and Locci (2024) assessed the performance and self-evaluation capabilities of different LLMs in classification tasks within the CC and environmental domain, while Grasso et al. (2025) proposed a novel framework for assessing anthropocentric bias in LLM-generated texts. Fore et al. (2024) experimented with LLMs for CC topics, finding that, while effective with fine-tuning, to ensure accuracy they require safeguards against misinformation.

4 Dataset Creation and Annotation

This section outlines how we applied, adapted, and extended the Character-Role Framework by adopting a novel bottom-up approach. We focus on building two manually annotated datasets for character-role detection within CC and environmental narratives. These datasets will serve as test sets for evaluating the performance of the LLMs before expanding to create the final, larger datasets.

4.1 NatSciEdCC and Ecoverse

We did not aim to restrict the scope of our investigation to a specific (sub)domain or a limited set of possible entities (which we refer to as a top-down approach), as we believe that relying on keywords or predefined lists could limit the diversity of characters discovered in a broader environmental domain. Instead, our goal was to capture a more heterogeneous and comprehensive set of character-roles, even if it might increase the complexity of the task and pose challenges to both human annotation and models' performance. To still ensure domain consistency while maintaining diversity, we used two datasets that cover various subtopics and discussions related to CC and environmental issues:

(i) The *NatSciEdCC* corpus (Stede et al., 2023) consists of 490 plain text files from *Nature* and *Science* editorials related to climate change. The texts are segmented into single paragraphs of varying lengths and annotated with multiple dimensions, including topicality (CC relevance) and frame coding.

(ii) *Ecoverse* (Grasso et al., 2024a) is a dataset of 3,023 tweets covering various environmental topics, including CC. It is annotated for eco-relevance, environmental impact, and the author's stance toward environmental causes (supportive, neutral, or skeptical/opposing). The dataset is openly available under a CC BY-SA 4.0 license.

4.2 Data Cleaning and Dataset Creation

Our goal was to create two datasets that contain rich and diverse narratives, focusing on texts with well-defined narrative elements while minimizing overly vague or noisy content. Given the heterogeneous nature of the language in *Ecoverse* tweets, ranging from formal news sources to informal user posts, and the structured language of scientific editorials, we applied tailored filtering steps to ensure meaningful and balanced content for analysis.

Filtering *Ecoverse* To maximize narrative diversity and reduce noise, we applied the following filtering steps: (i) Tweets unrelated to environmental or climate change (CC) topics were excluded, based on pre-existing eco-relevance annotations. (ii) To minimize the inclusion of overly hashtag-heavy tweets, which tend to lack substantial content, we removed tweets containing more than three hashtags. From the resulting set, we selected 300 tweets for manual annotation. To ensure diversity, we randomly sampled: (i) 180 tweets from environmental publications and news outlets (e.g., National Geographic, New York Times); (ii) 120 tweets from individual users, equally divided into 60 supportive tweets and 60 skeptical/opposing tweets. After combining these selections, we shuffled them to create a diverse dataset for manual annotation.

Filtering *NatSciEdCC* Similarly, for scientific editorials, we applied the following steps to ensure meaningful and balanced content: (i) We excluded extremely short paragraphs (fewer than 24 tokens) to focus on texts with sufficient narrative structure for analysis. (ii) We selected paragraphs with the highest topicality scores related to CC, based on existing annotations. (iii) To capture a broad spectrum of narrative tones, we performed sentiment analysis (Hutto and Gilbert, 2014) on the paragraphs and selected the 50 most positive, 50 most negative paragraphs, and 50 with mid-range sentiment. This ensured a balanced dataset of 150 paragraphs. Given that paragraphs are significantly longer than tweets—often two to four times the length—we determined that a dataset of 150 paragraphs would be sufficient for manual annotation and analysis.

4.3 Dataset Annotation

4.3.1 Character Definition

As Gehring and Grigoletto (2023) note, any entity can be a character in a narrative, making it useful to distinguish broader categories. To adapt our analysis to different text types (social media and editorials) and a wider range of ecological topics beyond climate change policy discussions, we expanded previous definitions of characters. In addition to "human" and "instrumental" characters, we introduced a novel third category—*natural characters*. This decision is informed by ecolinguistics (Stibbe, 2015, 2021; Fill and Muhlhausler, 2006), which shows how natural elements are often personified or attributed with agency in everyday language.

Language use frequently constructs natural entities as sentient or volitional, as seen in expressions in our datasets such as "nature destroys" "the forest heals", "the land is threatened", "rivers are stressed". These strategies also hide the true human agents behind these processes (Kuha, 2017). Recognizing this, the inclusion of natural characters enriches our analysis and opens possibilities for future ecocritical discourse analysis, where the use of such verbs and agency can be further examined.

Thus, our final character categories include:

• *Human Characters*: Individuals or groups (e.g., corporations, governments, organizations) whose actions, inactions, or beliefs significantly influence the narrative.

• *Instrumental Characters*: More abstract entities (e.g., policies, laws, technologies) or human-driven processes (e.g., "urbanization", "deforestation") that play key roles in the narrative and are produced or initiated by human characters.

• *Natural Characters*: Non-human entities (e.g., soil, oceans, animals) and natural phenomena (e.g., "climate change", "pandemics") when they are portrayed as playing an active or passive role in the narrative.

4.3.2 Task Description and Guidelines

The task and guidelines were consistent across both tweets and editorial paragraphs, with slight adjustments made during the annotation process to accommodate the differences in text types. Unlike Gehring and Grigoletto (2023), who annotated at the tweet level, we opted for a finer granularity by annotating at the entity level.

Guidelines Annotators received detailed guidelines[3], which mirrored the prompts later used for LLMs. These were based on the character-role definitions from Gehring and Grigoletto (2023) but tailored to the different text types. Annotators were tasked with identifying prominent characters by assigning them one of four roles: **Hero**, **Villain**, **Victim**, or **Beneficiary**. Linguistic indicators such as polarized or action-driven words — modifiers, verbs — (e.g., "heal", "save") helped determine roles. For editorials, where language can be more subtle, annotators also considered the 'overall sentiment' towards an entity when linguistic cues were not directly adjacent. Annotation proceeded sentence by sentence, and annotators paid attention to role shifts, where a character's role could evolve within a sentence or paragraph. They were instructed to focus on the author's communicative intent and avoid assumptions based on external world knowledge. Only nouns or noun phrases were eligible for labeling, including pronouns like "we",

[3] Guidelines of both character-role and character categorization tasks are provided in Appendix A.3.

which often reflect social actors portraying themselves. This was particularly relevant in tweets, where first-person pronouns are frequently used to express personal ideas. Below are examples of tweets (1)-(2) and editorials (3) with expected annotations, following the guidelines, with villains in burgundy red, heroes in green, victims in orange, and beneficiaries in blue.

(1) *Humboldt penguins face existential threats from climate change and overfishing—but also from habitat theft, as the penguins use guano for nesting while humans covet it for fertilizer.*

(2) *With #climate change impacting agriculture, Genetically Modified crops offer promising solutions, including reducing greenhouse gas emissions. However, the focus on profit over #sustainability risks farmer livelihoods & the #environment.*

(3) *US President Donald Trump is promoting a retrograde energy agenda and has vowed to pull the United States out of the Paris agreement. Still, despite their efforts, Trump's allies have been unsuccessful in stopping the rise of renewable energy companies, while local communities are benefiting from this.*

4.3.3 Annotation Process

The annotation process consisted of two consecutive phases: (i) character-role annotation for the datasets of **300** tweets and **150** editorial paragraphs, and (ii) character categorization for a subset of these annotations. In the second phase, 50 tweets and 50 editorial paragraphs were selected for categorizing entities into three categories: human, instrumental, and natural. Both annotation tasks were carried out using the Label Studio open-source data labeling tool[4], with a NER template and a tailored labeling setup.

Character-Role Annotation The primary character-role annotation task was conducted by two annotators, both part of the same research team. One is an author of this paper. One annotator self-identified as a male social scientist, and the other as a female linguist. To ensure consistency and a shared understanding of the guidelines, an iterative two-step training process was undertaken. Initially, the two annotators performed a pilot annotation on a secondary dataset of 20 tweets and 15 editorial paragraphs (with the same distribution as in the main datasets). After completing the annotation, the annotators compared their results, discussed disagreements and differing interpretations, and worked together to refine the guidelines where necessary. This process was repeated until both annotators were confident in applying the guidelines consistently. Following the pilot phase, the annotators began work on the main datasets, annotating two initial batches of 50 tweets and 20 paragraphs respectively. After these batches, we monitored the Inter-Annotator Agreement (IAA) to measure consistency. Annotators discussed any problematic cases and further refined the guidelines accordingly. Once most issues were addressed, the annotation of the remaining tweets and paragraphs continued without the need for further discussion sessions.

Categorization Task The second annotation task involved assigning one of the three character types - human, instrumental, natural - to previously labeled entities. A third annotator, a student member of our research group who self-identified as non-binary, was provided with task-specific guidelines. This annotator worked on a subset of 50 tweets and 50 paragraphs previously annotated for character/roles, containing respectively 101 and 373 labeled entities. This subset was chosen based on agreement between the two annotators in the first task to ensure higher consistency. The decision to use only a subset was made because, despite the reduced number of texts, each paragraph and tweet contained a significant number of annotated entities. Moreover, this task was deemed relatively objective, so only one annotator was used. Table 8 in Appendix A.1 reports the label distribution for this task. Both character-role annotated datasets and character categorization subsets were then used to instruct the LLMs for automatic character-role detection and categorization, as discussed in Section 5.

4.4 IAA and Datasets Statistics

To measure the agreement between the two annotators, we treated the character-role task as a Named Entity Recognition (NER) task, taking into account two elements: the text spans of each annotated entity within the text unit (either a paragraph or a tweet) and the label assigned to that text span. Agreement was achieved if the two annotators annotated the same entity with either an identical text span or overlapping text spans (e.g., "the President" vs. "President"). We used Precision, Recall, and F1-score to calculate the agreement, as these metrics account for both span overlap and label consistency (Brandsen et al., 2020; Hripcsak and Rothschild, 2005).

[4] https://labelstud.io

- IAA Tweets (300): Precision = 0.80, Recall = 0.73, F1-score = 0.76.
- IAA Editorial Paragraphs (150): Precision: 0.81, Recall: 0.87, F1-score: 0.84

Detailed comments and insights on disagreements are reported in Section 6. Table 1 reports the label distribution among annotators throughout the datasets. Table 7 in Appendix A.1 report datasets statistics.

Label	Tweets			Editorials		
	A1	A2	GPT4	A1	A2	GPT4
Hero	187	216	273	311	285	256
Beneficiary	199	175	177	163	73	55
Villain	112	166	210	285	290	280
Victim	82	87	179	135	143	172
Total labels	580	644	839	894	791	763

Table 1: Label Distribution for Tweets and Editorials test sets for A1, A2 and best model (A1: Annotator 1, A2: Annotator 2).

5 Experiments with LLMs: Methodology

5.1 Motivation and Models Employed

We aimed to evaluate the performance of large language models (LLMs) on the previously unexplored tasks of entity-level character-role detection and character categorization. The main advantage of using LLMs is their ability to perform well with less task-specific training data, as they are pretrained on vast amounts of text. Additionally, LLMs have performed well in similar tasks such as NER (Wang et al., 2023; Villena et al., 2024). However, they are also susceptible to hallucinations, where they generate outputs not grounded in the input data (Mittal et al., 2024). Therefore, we employed careful prompt engineering and iterative testing to optimize their performance in this highly subjective and complex task. We experimented with the following LLMs, covering both closed and open models: **GPT-4o**[5]; **GPT-4-turbo**; **GPT-4** (et al., 2024); **Llama-3.1-8B**[6].

Our methodology for the models' classification experiments proceeded in three phases: (i) an *Exploratory Phase* to refine prompt design and model setup, (ii) an *Exploitation Phase* to assess model performance against human annotators, and (iii) *Classification Phase* to apply the best-performing model and optimal setting across larger datasets of tweets and editorials.

5.2 Exploratory Phase: Prompt Design

The clarity of prompts is crucial for generating accurate outputs in classification tasks (Deldjoo, 2023). Given that small adjustments in prompts or model settings (e.g., temperature) can significantly impact results, we conducted exploratory experiments to refine both the prompts and model setup. **Character-role Task** We tested a zero-shot setting on a small sample of 5-7 textual units for each text type (paragraphs and tweets), iterating through different prompt strategies across models (GPT-4, GPT-4o, GPT-4-turbo, and Llama-3.1-8B-Instruct)[7]. This allowed us to monitor output variations and refine the prompts accordingly. The prompts closely mirrored the guidelines provided to human annotators, and we also experimented with different output formats. By the end of this phase, we finalized the best prompt format for each model. To ensure consistency with human annotations, we used an in-line tag annotation format.

Character Categorization Task Given the simpler nature of the character categorization task, we directly applied a few-shot setting, leveraging insights from the main task's exploratory phase. The output format remained consistent with the character-role detection task, using in-line tagging for the category names.

Some examples of the prompts used are provided in Appendix A.4, while the complete set of prompts, including all versions for all the LLMs, can be found in the GitHub repository linked earlier.

5.3 Experimental Setup

We conducted the experiments on the Paperspace platform[8], utilizing a configuration that includes an NVIDIA P6000 24GB GPU, 30GB of RAM, and a 8-core CPU. We employed the Huggingface Pipeline abstraction[9] to load the Meta-Llama-3.1-8B-Instruct. For the GPT models, we utilized OpenAI's APIs[10]. After conducting a qualitative manual analysis of responses generated at various temperature settings (ranging from 0.1 to 0.9), we

[5] https://openai.com/index/hello-gpt-4o/

[6] https://llama.meta.com/

[7] We also experimented with Llama-2-13B, but it frequently hallucinated, so we excluded it from further testing.

[8] https://www.paperspace.com/

[9] https://huggingface.co/docs/transformers/main_classes/pipelines

[10] https://openai.com/blog/openai-api

determined that a value of 0.2 provided the best balance between coherence and adherence to the prompts.

5.4 Exploitation Phase

After the prompts and parameters optimization, we conducted the main experiment to compare the performance of the LLMs against human annotators.
Character-role task We tested three prompt settings: (i) zero-shot, (ii) one-shot, and (iii) few-shot. Each model's performance was evaluated on the 300-tweet and 150-editorial test sets, measuring agreement with the two human annotators using Precision, Recall, and F1-score. The best F1-scores were achieved by GPT-4 with the few-shot setting for tweets (0.65) and the one-shot setting for editorials (0.70). Full agreement results for all settings and models are presented in Tables 2 and 3.

Tables 1 and 7 provide a comparison of GPT-4 label distribution and statistics against human annotations. Section 6 discusses the results and offers insights into the models' classification errors.
Character Categorization Task We tested the models in a few-shot setting on 50 tweets and 50 paragraphs manually annotated by Annotator 3. The task was to predict the correct character category (human, instrumental, or natural) for each previously labeled entity. The best-performing model was GPT-4o for tweets (F1: 0.88), and GPT-4o and GPT-4-turbo for paragraphs (F1: 0.78). Full performance metrics for all models (Precision, Recall, and F1-score) can be found in Table 4.

5.5 Classification Phase

In the final phase, we applied the best-performing model, GPT-4, in a few-shot configuration for character-role classification to automatically label two larger datasets: the 1,259 eco-related tweets from the *Ecoverse* dataset and an additional 2,000 editorial paragraphs from *NatSciEdCC*. The selection of these 2k paragraphs followed the same criteria used for the creation of the 150-paragraph test set, as detailed in Section 4.2. The aim was to scale the character-role detection process and create two fully annotated datasets. After merging the test sets with these new annotations, we obtained: (i) a **1,559**-tweet dataset and (ii) a **2,150**-editorial paragraph dataset. Finally, we applied the best-performing model for character categorization, GPT-4o, to label all characters in both datasets, assigning them one of three categories: human, instrumental, or natural. Table 5 shows the role distribution among these categories, and Table 6 provides the dataset statistics.

6 Results and Discussion

As shown in Tables 2 and 3, GPT-4 generally outperformed the other models, while Llama-3.1-8B showed the lowest agreement, particularly struggling with longer paragraphs, where it tended to hallucinate. GPT-4o delivered strong results for editorial paragraphs. For tweets, the few-shot setting yielded the best results (F1: 0.60 for Annotator 1 and F1: 0.65 for Annotator 2), while editorials saw the highest score in the one-shot setting (F1: 0.70). GPT-4o also showed high precision in zero-shot settings for editorials. Interestingly, in GPT-4o experiments, the one-shot setup performed slightly better the few-shot one in both editorial and tweet tasks (F1: 0.70 vs. F1: 0.66 for editorials, F1: 0.59 vs. F1: 0.52 for tweets). This may indicate that providing fewer examples helps the model generalize better, avoiding overfitting and inconsistencies. Models consistently aligned more with Annotator 2 for tweets, with minimal differences in editorials. Interestingly, GPT-4o aligned more closely with Annotator 1 in editorials. Overall, models performed better in recognizing character-roles in editorials, likely due to their structured and homogeneous nature.

6.1 Disagreements Analysis

To better understand the areas of disagreement, we have to distinguish two levels: (i) disagreement on text spans, where one annotator recognized a character and the other did not (i.e., presence/absence), and (ii) disagreement on labels for the same text span, where both annotators agreed on the entity but assigned different roles. For instance, in tweets, more than half of the disagreements between GPT-4 and Annotator 2 (522 instances) resulted from mismatched labels, while fewer disagreements (452) stemmed from mismatched text spans. This indicates that while role assignment may be subjective, the model effectively identified prominent characters. Importantly, we observed that disagreements between human annotators and between humans and models often arose from similar challenges. **Reliance on world knowledge**: Both models and annotators sometimes relied on external world knowledge rather than strictly following the author's intent. For example, GPT-4 and Annotator 1 labeled "fossil fuel companies"

Model	**Setting**	**Annotator1**			**Annotator2**		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
GPT-4	one-shot	0.52	0.64	0.57	**0.57**	0.68	0.62
	zero-shot	0.48	0.61	0.54	0.54	0.66	0.59
	few-shot	0.52	**0.72**	**0.60**	**0.57**	**0.76**	**0.65**
GPT-4o	one-shot	0.48	0.62	0.54	0.53	0.67	0.59
	zero-shot	0.44	0.58	0.50	0.50	0.65	0.56
	few-shot	0.47	0.59	0.52	0.53	0.65	0.58
GPT-4-Turbo	one-shot	**0.53**	0.55	0.54	**0.57**	0.57	0.57
	zero-shot	0.46	0.34	0.39	0.51	0.37	0.43
	few-shot	0.52	0.63	0.57	0.55	0.65	0.60
Llama3.1	one-shot	0.41	0.44	0.42	0.45	0.46	0.45
	zero-shot	0.37	0.27	0.31	0.44	0.32	0.37
	few-shot	0.37	0.65	0.47	0.40	0.68	0.51

Table 2: Models Performance on Tweets test set in terms of IAA with human annotators.

Model	**Setting**	**Annotator1**			**Annotator2**		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
GPT-4	one-shot	0.73	0.54	0.62	0.85	**0.59**	**0.70**
	zero-shot	0.75	0.49	0.59	0.88	0.55	0.68
	few-shot	0.72	0.61	0.66	0.78	**0.59**	0.67
GPT-4o	one-shot	0.71	0.55	0.62	0.86	0.58	0.69
	zero-shot	**0.79**	0.42	0.55	**0.89**	0.53	0.66
	few-shot	0.73	0.61	**0.67**	0.82	0.55	0.66
GPT-4-Turbo	one-shot	0.66	0.51	0.58	0.80	0.49	0.61
	zero-shot	0.69	0.37	0.48	0.85	0.38	0.53
	few-shot	0.66	**0.67**	0.66	0.72	0.62	0.67
Llama3.1	one-shot	0.61	0.52	0.56	0.68	0.33	0.44
	zero-shot	0.64	0.25	0.36	0.76	0.42	0.54
	few-shot	0.56	0.64	0.60	0.65	0.57	0.61

Table 3: Models Performance on Paragraphs test set in terms of IAA with human annotators.

Model	**Precision**	**Recall**	**F1-score**
Tweets 50 test set			
GPT-4	**0.86**	0.89	0.87
GPT-4o	**0.86**	**0.90**	**0.88**
GPT-4-Turbo	0.77	0.79	0.78
LLama3.1	0.63	0.75	0.68
Editorials 50 test set			
GPT-4	0.76	0.76	0.76
GPT-4o	0.76	**0.79**	**0.78**
GPT-4-Turbo	**0.77**	**0.79**	**0.78**
LLama3.1	0.65	0.65	0.65

Table 4: Model performance on the categorization task for the 50-tweet and 50-editorial paragraph datasets (few-shot setting).

as villains, likely due to their general environmental impact, and "poor countries" as victims, possibly influenced by the linguistic cue "poor", even when the narrative did not explicitly frame them as such. **Subtle framing of characters**: Some entities played subtle yet important roles, making it difficult to decide whether to annotate them. Models, especially in tweets, tended to over-annotate compared to humans (as seen in Tables 1 and 7), though this over-annotation followed consistent patterns. For instance, in a tweet calling Croatia a "remarkable biodiversity hot spot", the model labeled Croatia as a beneficiary, or in "we can develop #green habits", it labeled "green habits" as a hero. However, model and human alignment was much closer in editorials, with GPT-4's label distribution closely matching Annotator 1 and 2. **Role ambiguity**: Some entities played multiple roles within the same text. For example, "endangered species" in a tweet could be labeled as victims of habitat destruction but also as beneficiaries of conservation efforts. Finally, **calculation strategy** also influenced disagreement rates. Partial overlap of text spans (e.g., "*the* president" vs. "president *of the USA*") was not counted as an agreement, though they referred to the same entity.

Label	**Tweets**				**Editorials**			
	Tot	Human	Instr.	Natural	Tot	Human	Instr.	Natural
Hero	722	295	271	156	3906	2013	1596	297
Victim	989	301	223	465	1974	931	511	532
Villain	693	275	259	159	2599	843	1192	564
Beneficiary	898	340	168	335	876	520	245	111

Table 5: Label distribution among categories for Final Tweets and Editorials datasets

Statistic	**Tweets**	**Editorials**
Mean tags per text	2.24	4.39
Mean words per text	35.08	89.22
Texts with only 1 entity	313	137
Texts with more than 1 entity	1021	1972
Texts with no entities	225	41
Total number of texts	**1559**	**2150**

Table 6: Statistics for both the Tweets and Editorials final datasets.

6.2 Preliminary Datasets Insights

To gain an initial understanding of the narrative structure across the datasets, we (i) calculated the co-occurrence of <category-role> pairs within the texts and (ii) observed the most frequently occurring character/role combinations by category (after lemmatizing the labeled entities). Figure 1 in Appendix A.2 presents the co-occurrence matrices of label pairs, showing how frequently each <category-role> pair occurs together within the same text. In the *Paragraph-matrix*, **Instrumental-Hero** is the most frequent pair, often co-occurring with **Human-Hero** and **Human-Villain**. This suggests that policies, technologies, or interventions are portrayed as solutions to problems driven by human actions or institutions. **Human-Victim** is another prominent category, frequently paired with **Human-Villain** and **Instrumental-Villain**, reflecting that humans are often depicted as suffering due to harmful policies or corporate actions. The frequent pairing of **Natural-Victim** with **Instrumental-Villain** reinforces the idea of natural entities (e.g., animals, biodiversity) being victims of human or institutional harm. In the *Tweets-matrix*, **Human-Victim** and **Human-Villain** frequently co-occur, indicating a strong narrative of communities suffering from human-made harm. **Natural-Victim** also appears frequently, especially alongside **Human-Villain** and **Instrumental-Villain**, reflecting the recurrent theme of environmental damage due to human-driven actions. Both matrices highlight the central role of **Instrumental-Hero** in environmental narratives, emphasizing the importance of policy measures and technologies as solutions. Meanwhile, **Natural-Victim** and **Human-Villain** are strongly linked in both datasets, underscoring a consistent framing of human-driven environmental harm.
Cross-referencing character roles with the stance of tweets reveals further nuances. In tweets with a *supportive* stance, “greenwashing”, “fossil fuel companies” and “plastic” are frequent villains, while heroes include “Inflation Reduction Act”, “climate movements”, “scientists”. In contrast, in *skeptical/opposing* tweets, entities like “government” and “climate scientists” are often framed as villains, with abstract concepts like “freedom” and “societal values” portrayed as victims. Editorials reflect similar patterns but within a more formal, structured narrative, with “scientists”, “developing countries”, and “biodiversity” frequently appearing as victims, and “climate change," “coal” and “Trump” as villains.

7 Conclusion

In this paper, we introduced a novel approach to analyzing climate change (CC) and environmental narratives using the Character-Role Framework across social media (from the *Ecoverse* dataset (Grasso et al., 2024a)) and scientific editorials (from the *NatSciEdCC* corpus (Stede et al., 2023)). We extended the framework by adopting a bottom-up approach and performing fine-grained entity-level analysis. After manually annotating two test sets of editorial paragraphs and tweets for characters in four roles (villain, hero, beneficiary, victim) and three categories (natural, human, instrumental), we evaluated four Large Language Models (LLMs) (GPT-4, GPT-4o, GPT-4-turbo, Llama-3.1-8B) on character-role detection and categorization tasks. GPT-4, the best-performing model, was then applied to create two fully annotated datasets: 1,559 tweets and 2,150 editorial paragraphs. Finally, we conducted a qualitative error analysis and explored the narrative patterns emerging from the datasets.

8 Limitations

Some limitations of our work include the following:

1. Our approach does not incorporate co-reference resolution, which may result in different mentions of the same entity (e.g., “he”, “the president”) being treated as separate characters, or conversely, labeled multiple times. As we detailed in Section 4.3.2, our analysis and annotation guidelines include personal pronouns, which leads to this limitation.

2. The datasets used (scientific editorials and tweets) are primarily from English-speaking regions, which might not capture the full range of CC and environmental narratives across different cultures. This limits the generalizability of our findings to other linguistic or cultural contexts.

3. A broader range of open models, differing in type and size, would offer a stronger basis for evaluating their performance. This highlights a limitation of our work, as it does not fully capture the diversity of available open models. While we acknowledge that comparing LLaMA-3.1-8B to much larger GPT models is inherently imbalanced, it provides an initial perspective on how their performance differs for these tasks.

4. Expanding the dataset with LLM-generated annotations carries inherent risks, as these models can reflect biases or limitations in their training data, including outdated or incomplete world knowledge (Blodgett et al., 2020). However, as discussed, we observed that LLM biases remarkably often aligned with those of human annotators, resulting in similar disagreement patterns. This suggests that the models can perform reliably on such a subjective and complex task as character-role detection. For example, in the first batch of 100 tweets, agreement between Annotator 2 and GPT-4 (F1: 0.77) exceeded the inter-annotator agreement between humans (F1: 0.72). While encouraging, we recognize that LLMs are not perfect substitutes for human annotation. Future work should include further validation, such as cross-checking expanded datasets with smaller manually labeled subsets or assessing their robustness in downstream tasks to ensure reliability in practical scenarios.

5. The preliminary dataset insights presented in Section 6.2 primarily focus on category-level trends (i.e., distributions of human, instrumental, and natural entities across character roles and their co-occurrence). While this provides an important first look into broader framing patterns, a more fine-grained analysis of the most frequent entities within each category and the specific narrative structures they form (e.g., recurring villain-victim or hero-beneficiary pairings) remains an open area for future work. Investigating these patterns could offer a deeper understanding of how environmental narratives are constructed, revealing more nuanced ’plot’ dynamics among characters.

References

Richard J. Alexander and Arran Stibbe. 2014. From the analysis of ecological discourse to the ecological analysis of discourse. *Language Sciences*, 41:104–110.

E. Armstrong and A. Ferguson. 2010. Language, meaning, context, and functional communication. *Aphasiology*, 24:480 – 496.

Kelly Bergstrand and James M. Jasper. 2018. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81:228 – 247.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the university of turin. In *CEUR Workshop Proceedings*, volume 3596. CEUR-WS.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

Jannis Bulian, Mike S Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen

Huebscher, Christian Buck, Niels Mede, Markus Leippold, et al. 2023. Assessing large language models on climate information. *arXiv preprint arXiv:2310.02932*.

Yashar Deldjoo. 2023. Fairness of chatgpt and the role of explainable-guided prompts. *ArXiv*, abs/2307.11761.

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alwin Fill and Peter Muhlhausler. 2006. *Ecolinguistics reader: Language, ecology and environment*. A&C Black.

Kjersti Fløttum and Øyvind Gjerstad. 2013. Arguing for climate policy through the linguistic construction of narratives and voices: the case of the south-african green paper "national climate change response". *Climatic Change*, 118:417–430.

Kjersti Fløttum and Øyvind Gjerstad. 2017. Narratives in climate change discourse. *Wiley Interdisciplinary Reviews: Climate Change*, 8.

Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. Unlearning climate misinformation in large language models. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 178–192, Bangkok, Thailand. Association for Computational Linguistics.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework. *SSRN Electronic Journal*.

Shoshanah B.D. Goldberg-Miller and Rachel Skaggs. 2022. The story and the data: Using narrative policy framework to analyze creative economy reports. *Artivate*, 10:–.

Francesca Grasso and Stefano Locci. 2024. Assessing generative language models in classification tasks: Performance and self-evaluation capabilities in the environmental and climate change domain. In *International Conference on Applications of Natural Language to Information Systems*, pages 302–313. Springer.

Francesca Grasso, Stefano Locci, and Luigi Di Caro. 2025. Towards addressing anthropocentric bias in large language models. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 84–93, Tallinn, Estonia. University of Tartu Library.

Francesca Grasso, Stefano Locci, Giovanni Siragusa, and Luigi Di Caro. 2024a. Ecoverse: An annotated twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5461–5472.

Francesca Grasso, Ronny Patz, and Manfred Stede. 2024b. Nytac-cc: A climate change subcorpus of new york times articles. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 403–409, Pisa, Italy. CEUR Workshop Proceedings.

Reiner Grundmann and Ramesh Krishnamurthy. 2010. The discourse of climate change: A corpus-based approach. *Critical approaches to discourse analysis across disciplines*, 4(2):125–146.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Mike Hulme. 2015. Why we disagree about climate change. *Zygon*, 50(4).

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Michael D Jones. 2018. Advancing the narrative policy framework? the musings of a potentially unreliable narrator. *Policy Studies Journal*, 46(4):724–746.

Michael D. Jones and Mark K. McBeth. 2010. A narrative policy framework: Clear enough to be wrong? *Policy Studies Journal*, 38:329–353.

Michael D Jones, Mark K McBeth, and Elizabeth A Shanahan. 2022. Narratives and the policy process: Applications of the narrative policy framework.

Manfred Klenner, Anne Göhring, and Sophia Conrad. 2021. Getting hold of villains and other rogues. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 435–439, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Nikolay Koldunov and Thomas Jung. 2024. Local climate services for all, courtesy of large language models. *Communications Earth & Environment*, 5(1):13.

Mai Kuha. 2017. The treatment of environmental topics in the language of politics. In *The Routledge handbook of ecolinguistics*, pages 249–260. Routledge.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, et al. 2024. Automated fact-checking of climate change claims with large language models. *arXiv preprint arXiv:2401.12566*.

Ashish Mittal, Rudra Murthy, Vishwajeet Kumar, and Riyaz Bhat. 2024. Towards understanding and mitigating the hallucinations in nlp and speech. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 489–492.

Brigitte Nerlich, Nelya Koteyko, and Brian Brown. 2010. Theory and language of climate change communication. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1):97–110.

Holly L. Peterson. 2021. Narrative policy images: Intersecting narrative & attention in presidential stories about the environment. *Policy Studies Journal*.

Holly L. Peterson, Chad Zanocco, and Aaron Smith-Walter. 2021. Lost in translation: Narrative salience of fear > hope in prevention of covid-19. *Narratives and the Policy Process : Applications of the Narrative Policy Framework*.

Donald E Polkinghorne. 1988. Narrative knowing and the human sciences. *SUNY Pres*.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2023. Entity-level sentiment analysis (elsa): An exploratory task survey. *arXiv preprint arXiv:2304.14241*.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Kalyan Dutia, Alok Singh, Julia Bingler, Christophe Christiaen, Neetu Kushwaha, Veruska Muccione, Saeid A. Vaghefi, and Markus Leippold, editors. 2024. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Association for Computational Linguistics, Bangkok, Thailand.

Manfred Stede, Yannic Bracke, Luka Borec, Neele Charlotte Kinkel, and Maria Skeppstedt. 2023. Framing climate change in nature and science editorials: applications of supervised and unsupervised text categorization. *Journal of Computational Social Science*.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. *Proceedings of the 1st Workshop on NLP for Positive Impact*.

Arran Stibbe. 2015, 2021. *Ecolinguistics: Language, ecology and the stories we live by*. Routledge.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *ArXiv*, abs/2401.09646.

Fabián Villena, Luis Miranda, and Claudio Aracena. 2024. llmner: (zero|few)-shot named entity recognition, exploiting the power of large language models. *ArXiv*, abs/2406.04528.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *ArXiv*, abs/2304.10428.

Erika Allen Wolters, Michael D. Jones, and Kathryn L. Duvall. 2021. A narrative policy framework solution to understanding climate change framing research. *Narratives and the Policy Process : Applications of the Narrative Policy Framework*.

Haiqi Zhou, David Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155, Bangkok, Thailand. Association for Computational Linguistics.

Hongyin Zhu and Prayag Tiwari. 2023. Climate change from large language models. *ArXiv*, abs/2312.11985.

A Appendix

A.1 Datasets Statistics

Statistic	Tweets			Editorials		
	A1	A2	GPT4	A1	A2	GPT4
Avg. label/txt	1.93	2.15	2.80	5.96	7.00	5.09
texts w/1 label	32	42	41	8	2	0
texts w/>1 labels	176	182	257	138	145	150
texts w/no labels	92	76	2	5	3	0

Table 7: Statistics for Tweets and Editorials Test sets for A1, A2, and best model (A1: Annotator 1, A2: Annotator 2.)

Label	Editorials (50)	Tweets (50)
Natural	94	40
Human	125	42
Instrumental	154	19

Table 8: Character Categorization Label Distribution for 50 Editorial Paragraphs and 50 Tweets.

A.2 Co-occurrences Matrices

Figure 1 shows the co-occurrence matrices of category-role pairs.

A.3 Annotation Guidelines

Annotation Guidelines for Character-Role and Character categorization tasks.

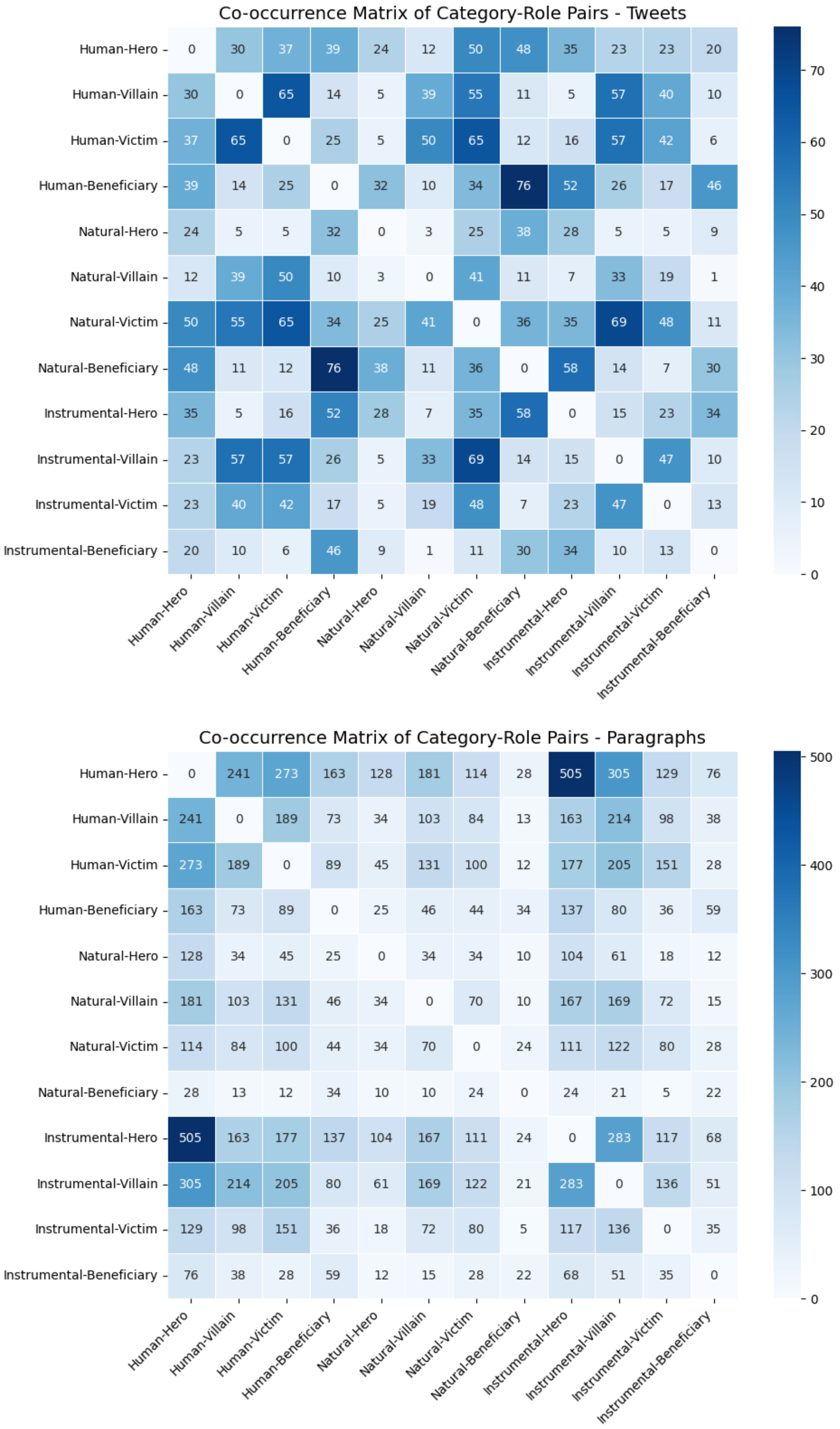

Figure 1: Co-occurrence matrices.

Character Roles: Annotation Guidelines

Introduction The aim is to leverage and adapt the Character-Role Narrative Framework (Gehring & Grigoletto 2023) which in turn stems from the so-called Policy Narrative Framework (e.g. Jones et al. 2022), to analyze the narratives underlying two different textual contexts: (i) A set of tweets (from the EcoVerse Dataset), all linked to ecology/environment-related themes; (ii) A set of editorials (from *Nature* & *Science*), all linked to the climate change topic.

Task Overview The goal of this annotation is to identify the entities that contribute to the text's narrative, i.e., to the core message being conveyed. We identify these entities as characters bearing specific roles from a small inventory. Crucially, we analyze the *author's narrative*, so we must always keep in mind to rely on the perception of its communicative intention and limit access to our world knowledge.

Character Roles: Definitions

Typically, characters can assume one of three (in our project, four) fundamental roles in the "drama triangle": **hero**, **villain**, **victim**, or **beneficiary**. *Heroes* actively contribute to, endorse, or are portrayed as having the potential to determine positive actions or events. Importantly, the hero is also assigned a determinant role within the narrative; they are provided with the potential to do something, regardless of whether they actually pursue their mission. *Villains* contribute to, endorse, favor, or determine negative actions or events. *Victims* are harmed, endangered, potentially harmed, or suffer from the consequences of events or actions, typically playing a passive role in the narrative. *Beneficiaries* play a passive role and benefit or potentially benefit from events or actions being described.

Types of Characters: Definitions

Characters are entities that play an identifiable role within the narrative and determine its essence. Characters can be:

Human Characters: These include humans or entities made up of people, such as corporations, governments, organizations of any type (e.g., religious), and political movements, whose actions, inactions, or beliefs are crucial for the narrative and message of the text. *Instrumental Characters*: These are more abstract entities such as policies, laws, technologies, measures, or objects that (i) have been produced by human characters, (ii) are important for the narrative and message of the text, and (iii) can be assigned a character role, as they are determining for the narrative being told. *Natural Characters*: These comprise non-human entities such as natural elements (e.g., soil, oceans), animals, nature itself, the planet, and so on. They can also be processes ("city growth") and phenomena ("climate change," "pandemic") on the condition that they clearly have inherited some agentive role within the narrative.

Annotation Procedure

Before starting the annotation, read the entire text to understand the core message. Then, proceed with the annotations from beginning to end, sentence by sentence. For each sentence: (i) Decide whether there are significant characters with prominent roles within the text that can be assigned the type human, instrumental, or natural. (ii) Do not label if no particular narrative (and subsequently no characters/roles) can be identified, and/or if the text is too vague or lacks the author's perspective. (iii) Assign the character role based on contextual information (villain, hero, beneficiary, victim). (iv) Once completed, click on "submit."

Notes on Annotation: Borderline Cases for Editorials

In editorials, the narrative is often spread across larger portions of text. Annotators should consider the entire paragraph to understand the overall perception and narrative conveyed. The assignment of roles to characters by the author is often subtle and implicit. Annotators may need to infer roles based on the overall narrative, allowing for reasonable implications or assumptions, especially when strong linguistic indicators are absent. Abstract concepts, such as "decision," cannot be annotated as characters. However, instrumental characters like "measures," "reports," "laws," or "policies" should be annotated. When both an instrumental character (e.g., the name of a report) and the human character responsible for it (e.g., the government) are mentioned, annotate BOTH.

Character Categorization: Annotation Guidelines

Introduction The goal of this annotation task is to categorize each entity labeled with a character role—*hero*, *villain*, *victim*, or *beneficiary*—into one of three predefined supercategories: **human**, **instrumental**, or **natural**. This categorization helps identify the broader nature of entities contributing to the text's narrative.

Task Overview For each entity previously annotated with a character role, you will assign one of the following categories: **Human**; **Instrumental**; **Natural**.

This categorization does not depend on the entity's role in the text but solely on the entity's *type* based on the definitions provided below.

Category Definitions

1. Human Characters: Humans or entities representing humans or groups of humans. They include:

- Individuals: *e.g., "scientists", "activists", "farmers".*
- Organizations, governments, corporations, or institutions: *e.g., "United Nations", "fossil fuel companies", "NGOs".*
- Groups of people: *e.g., "communities", "developing countries".*

Example:

- *The **government** decided to implement new measures.* → **Human**

2. Instrumental Characters: More abstract entities that are human-made or human-driven, such as:

- Policies, laws, reports, measures: *e.g., "climate policies", "30x30 report".*
- Technologies or objects: *e.g., "dams", "wind turbines", "plastic".*
- Processes initiated or controlled by humans: *e.g., "urbanization", "deforestation", "city growth".*

3. Natural Characters: Non-human entities from the natural world or natural phenomena, such as:

- Animals, plants, natural elements: *e.g., "biodiversity", "oceans", "forests".*
- The environment as a whole: *e.g., "the planet", "nature".*
- Natural processes or phenomena: *e.g., "climate change", "pandemics", "wildfires".*

Annotation Procedure

1. **Review the entity:** For each entity already labeled with a role, identify its type (human, instrumental, or natural) based on the provided definitions.
2. **Assign a category:** Use the definitions and examples to determine the appropriate category.
3. **Handle borderline cases:**
   - If an entity fits more than one category, prioritize the most contextually relevant type.
   - For processes (e.g., "city growth"), determine if it is human-driven (Instrumental) or naturally occurring (Natural).

Notes for Annotators If the type remains unclear, discuss it with the coordinators to ensure consistency.

A.4 Prompts for LLMs

Prompt GPT family for character-role detection task - one shot

Task Overview: You are given a text. Your task is to identify and label characters within the narrative. Characters are entities playing a clear role in the story, contributing to its core message. Label each identified character with one of the following roles: **Hero**: Actively contributes to or endorses positive actions/events. **Villain**: Responsible for negative actions or harm. **Victim**: Suffers from or is endangered by actions/events, typically playing a passive role. **Beneficiary**: Passively benefits from actions/events. **Character Types: Human Characters**: Humans or entities made up of people (e.g., corporations, governments, organizations). **Instrumental Characters**: Abstract entities (e.g., policies, laws, technologies) produced by human characters that play a crucial role in the narrative. **Natural Characters**: Non-human entities (e.g., animals, nature, natural processes) given agentive or passive roles within the narrative.

Instructions: (i) Identify Characters: Assess each sentence, sentence by sentence, to identify characters (there can be 0 to N characters per sentence).

(ii) Assign Roles: Label identified characters with the appropriate role based on how the narrative portrays them. Do not infer or imply any roles based on common knowledge or assumptions. Only label characters if the text explicitly describes them in a way that fits a specific role (Hero, Villain, Victim, Beneficiary). Rely strictly on what is explicitly stated in the text—avoid making interpretations or assumptions.

(iii) Use Linguistic Indicators: Pay close attention to linguistic cues such as “heal," “save," “suffer from," “endangered by," “protect," and other similar phrases. These indicators will help determine the role of a character. If the text does not explicitly use such indicators or similar language, do not assign a role based on presumed implications.

(iv) Be Aware of Role Shifts: A character’s role can change as the sentence or paragraph progresses. Even if a character starts neutral, it might take on a role later in the sentence. Similarly, a character can switch roles within the same sentence or paragraph. Assign roles based on how the character is portrayed at each point in the text.

(v)Focus on Narrative Perspective: Use linguistic indicators and context within the text to determine roles, strictly reflecting the author’s intended perspective. Avoid relying on external knowledge or common narratives—only label characters based on the explicit narrative context provided.

(vi) Label Nouns Only: Only label nouns or noun phrases, excluding articles (e.g., “the” in “the President”) and other parts of speech. Personal pronouns (e.g., “we,” “they”) can be labeled too.

(vii) Multiword Expressions: For multiword expressions (e.g., “President of the United States”), label the entire phrase, but avoid including unnecessary extensions.

(viii) Avoid Labeling Abstract Entities: Do not label overly abstract entities such as “decision”.

No Labeling If: No clear narrative or characters/roles are identifiable. The text is too short, vague, or the narrative is too implicit. The text does not express the author’s perspective (e.g., reporting someone else’s perspective).

Output format: You must return the input text with each character labeled using in-line tag annotations (`<start_token>text<end_token>`), where the tag corresponds to a role name. The only available tags are: **Hero**: `<HER>text</HER>` **Villain**: `<VIL>text</VIL>` **Victim**: `<VIC>text</VIC>` **Beneficiary**: `<BEN>text</BEN>`

For example, if the input text is “The Government saved the environment." the output text should be: “The `<HER>Government</HER>` saved the `<BEN>environment</BEN>`."

IMPORTANT: **DO NOT CHANGE THE INPUT TEXT, ONLY ADD THE TAGS.**

Note: Be attentive to the linguistic cues and specific wording used by the author, as they will guide you in assigning the correct roles. Avoid inferring roles based on outside knowledge or assumptions.

Here is the text to annotate:

Prompt Llama for character-role detection task - zero shot

Task Overview: You are given a text. Your task is to identify and label characters within the narrative. Characters are entities playing a clear role in the story, contributing to its core message. Label each identified character with one of the following roles:

Hero: Actively contributes to or endorses positive actions/events.

Villain: Responsible for negative actions or harm.

Victim: Suffers from or is endangered by actions/events, typically playing a passive role.

Beneficiary: Passively benefits from actions/events.

Here's how you should do it:

Look at each sentence: Go through the text sentence by sentence to find any characters. There can be no characters, one character, or many characters in a sentence.

Label the characters: When you find a character, give them the correct label (Hero, Villain, Victim, or Beneficiary) based only on what the text says. Do not guess or assume anything. Only label a character if the text clearly shows their role.

Pay attention to words: Words like “heal," “save," “suffer," “endangered," “protect," or similar words can help you decide the character's role. If these or similar words are not there, do not label based on your assumptions.

Roles can change: A character's role might change in the same sentence. Label them based on how they are described at that moment.

Use tags in the text: When you label a character, use the following tags directly in the text:

Hero: `<HER>text</HER>`

Villain: `<VIL>text</VIL>`

Victim: `<VIC>text</VIC>`

Beneficiary: `<BEN>text</BEN>`

IMPORTANT: Rewrite the entire text with these tags included. Do not change the original text—just add the tags around the characters.

Remember, only label characters based on what is clearly stated in the text. Do not use your own knowledge or assumptions.

Here is the text to annotate:

Prompt for GPT family and Llama for character categorization task - few-shot

Task Overview: You are given a text. Each text represents a unique entity, and your task is to categorize the entity based on one of three categories: *Human*, *Instrumental*, or *Natural*. Below are the definitions for each category, along with relevant examples.

Categories:

Human Characters: These include humans or entities made up of people, such as corporations, governments, organizations of any type (e.g., religious), and political movements. Examples:

“Oil and gas industry” (categorized as Human because it refers to a group of businesses).

“World” (categorized as Human when referring to governments, organizations, or companies).

Instrumental Characters: These are more abstract entities such as policies, laws, technologies, measures, objects, or human-driven processes (e.g., “urbanization," “deforestation," “city growth”). They can also be artifacts or processes that have been produced or initiated by human characters. Examples:

“Pesticides and fertilizers” (categorized as Instrumental because they are human-made technologies).

“Carbon emissions” (categorized as Instrumental because they result from human processes).

Natural Characters: These comprise non-human entities such as natural elements (e.g., soil, oceans), animals, nature itself, and the planet. They can also include natural processes or phenomena (e.g., “biodiversity loss,” “climate change,” “pandemic”). Examples:

“Europe” (categorized as Natural when referring to the geographical region and its natural elements, rather than its people).

“Smoke” (categorized as Natural when referring to poor air quality from smoke, assuming it is not human-caused).

Output Format: You must return the input text with each entity labeled using in-line tag annotations (`<start_token>`text`<end_token>`), where the tag corresponds to a category name. The only available tags are:

```
Human: <HUM>text</HUM>
Instrumental: <INS>text</INS>
Natural: <NAT>text</NAT>
```

Examples:

```
<HUM>oil and gas industry</HUM>
<HUM>low-income communities</HUM>
<INS>30x30 policy</INS>
<INS>pesticides and fertilizers</INS>
<NAT>climate change</NAT>
<NAT>the ocean</NAT>
```

Here is the text to annotate: