Evaluating Taxonomy Free Character Role Labeling (TF-CRL) in News Stories using Large Language Models

David G Hobson[1], **Derek Ruths**[1], **Andrew Piper**[2]
[1]School of Computer Science
[2]Department of Languages, Literatures, and Cultures
McGill University

Abstract

We introduce Taxonomy-Free Character Role Labeling (TF-CRL); a novel task that assigns open-ended narrative role labels to characters in news stories based on their functional role in the narrative. Unlike fixed taxonomies, TF-CRL enables more nuanced and comparative analysis by generating compositional labels (e.g., Resilient Leader, Scapegoated Visionary). We evaluate several large language models (LLMs) on this task using human preference rankings and ratings across four criteria: faithfulness, relevance, informativeness, and generalizability. LLMs almost uniformly outperform human annotators across all dimensions. We further show how TF-CRL supports rich narrative analysis by revealing novel latent taxonomies and enabling cross-domain narrative comparisons. Our approach offers new tools for studying media portrayals, character framing, and the socio-political impacts of narrative roles at-scale.[1]

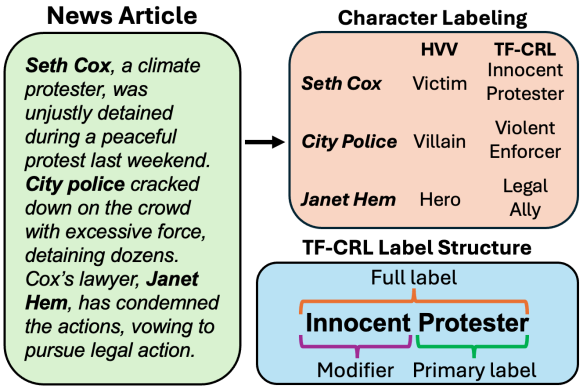


Figure 1: TF-CRL assigns open-ended narrative role labels to characters in news stories based on their functional role in the narrative. Using a compositional structure of primary labels and modifiers, they provide enhanced resolution and nuance compared to fixed typologies like the Hero-Villain-Victim framework (HVV).

1 Introduction

Characters are central to narrative understanding. Whether in fiction or nonfiction, they serve as the primary agents through which events unfold and meanings are constructed (Fludernik, 2002). In the context of news stories, characters often represent real people or collective entities whose actions, experiences, and relationships shape how events are perceived (Johnson-Cartee, 2005). Beyond their functional role in event progression, such entities also provide a bridge to social cognition, enabling readers to attribute mental states, infer motivations, and empathize with real individuals and the causes they represent (Eekhof et al., 2022; Mar, 2018).

Characters in narratives are more than just named individuals performing actions—they serve functional roles within larger narrative structures. A rich tradition in literary and narrative theory, including Propp's Morphology of the Folktale (Propp, 1968), Greimas's actantial model (Greimas, 1983), and the Aarne-Thompson-Uther Index (Uther, 2004), has emphasized that characters frequently embody recurring roles—such as hero, villain, helper, or victim—that organize and propel story dynamics. These roles are not simply descriptive categories; they express the relational and structural logic of the narrative, revealing who initiates change, who resists it, and who is acted upon.

In the context of news stories, recognizing such roles allows for a deeper understanding of how events are framed, how blame or sympathy is distributed, and how larger social narratives (e.g., about justice or conflict) are constructed and sustained. By moving beyond surface-level mentions of particular entities to role-based representations, NLP systems can uncover narrative patterns that shape public perception and collective memory.

A growing body of work in NLP has begun to

[1]Code, data, prompts, and codebooks available at https://github.com/davidghobson1/taxonomy-free-character-role-labeling

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14829–14851
November 4-9, 2025

Sample 1: *The teacher's union continued their strike today in response to recent government cuts. "Teachers continue to be grossly underpaid. We work long hours and the government continues to under-appreciate our service to society," a striking teacher said.*	Sample 2 : *Mark Zuckerberg has always received the brunt of the blame for broader issues with the tech industry. But he's just an innovator. He shouldn't have to deal with all the hate. It's not his fault his company's popular and he's a billionaire.*
Character Role Labels	
Striking teacher	**Mark Zuckerberg**
• Victim	• Victim
• Righteous Protester	• Scapegoated Visionary
• Public Servant*	• Billionaire*

Table 1: Sample character role labels from two fictitious news articles. In terms of our evaluation criteria, the first two labels for each are both faithful and highly relevant. Labels marked with * are faithful but only moderately relevant as they don't capture the element of victimhood central to both depictions.

tackle this challenge. Recent work has focused on reproducing Propp's character archetypes from Russian folktales (Valls-Vargas et al., 2014; Jahan et al., 2021), classified film characters into movie tropes using dialogue (Ziems et al., 2024), and applied the Hero-Villain-Victim (HVV) framework to the news (Gomez-Zara et al., 2018; Stammbach et al., 2022).

One of the principal limitations of this work, however, is the reliance on pre-existing taxonomies (Dundes, 1962). While these may serve specific researcher aims, they can also compromise nuance and resolution, limit knowledge about more diverse latent narrative structures within broader collections of news stories, and hinder certain kinds of large-scale narrative modeling activities. For example, an underpaid teacher and Mark Zuckerberg might both be portrayed as victims in different news stories, but in very different ways and as parts of very different narratives as shown in Table 1.

In this work, we propose "Taxonomy Free Character Role Labeling" (TF-CRL): an open-ended narrative understanding task that assigns a short noun phrase to the function a character performs in a narrative. Unlike traditional schema-driven approaches that rely on predefined taxonomies of roles, TF-CRL allows for a more flexible and exploratory representation of character functions. We show how this approach captures a broader and more nuanced range of social and narrative positions, while also allowing the ability to compare narratives across domains.

At the heart of our method is a compositional labeling framework, where each role consists of a primary label that names the core role (e.g., Leader, Parent, Victim) and an optional modifier that describes how the role is enacted (e.g., Skilled, Understanding, Reluctant) as shown in Figure 1. This structure preserves interpretability and enables contextual distinctions between otherwise similar roles. Crucially, TF-CRL maintains transferability and generalizability across narrative domains while avoiding the need for proscribed ontologies, allowing models to adapt to the diverse nature of roles in real-world narratives such as news stories.

Through the use of case studies, we demonstrate how our approach offers new avenues to study characters at the character-level, by providing enhanced resolution compared to pre-existing taxonomies, at the narrative-level, by aiding in the discovery of new character roles, and also at the cross-narrative level, by comparing roles across domains.

2 Related Work

Characters and character representations have been studied extensively in NLP literature, both in the context of fiction and non-fiction. Zhang et al. (2019) used extractive methods to identify character attributes from fictional stories. Brahman et al. (2021) employed transformer-based approaches to generate unstructured character descriptions from literature summaries. More recently, several works have used large language models (LLMs) for attribute extraction (Baruah and Narayanan, 2024) and structured character description generation (Gurung and Lapata, 2024; Papoudakis et al., 2024; Yuan et al., 2024).

While these methods capture important aspects

of characters, developing simplified representations that capture recurring character patterns remains an open research problem, especially in large-scale domains such as news. Prior work using embeddings and clustering methods based on character descriptions or related representations has made notable progress (Anderson, 2025; Holgate and Erk, 2021; Inoue et al., 2022), though recreating role- or archetypal-based narrative structures remains difficult. Notably, Brahman et al. (2021) identified the commonalities between character descriptions and character summaries, but emphasized the need to go beyond summarization to "abstract out the low-level content of the narrative instead of simply identifying and paraphrasing important details."

In the public policy domain, characters are a core component of the Narrative Policy Framework (NPF; Jones and McBeth, 2010) which has been extensively used to study the news. NPF classifies characters into heroes, villains, and victims (HVV) where heroes are defined as those who fix problems, villains as those who cause them, and victims as those who are harmed by problems. Other variants operationalize additional roles, including beneficiaries (who receive benefits from policies) and allies (who assist hereos or victims) (Kuenzler et al., 2025). In the context of NLP, NFP has been applied to study news using traditional NLP methods (Gomez-Zara et al., 2018), BERT-based methods (Gehring and Grigoletto, 2023) and LLMs (Stammbach et al., 2022).

While pre-defined role taxonomies, like HVV, are useful for their interpretability and ability to bundle character qualities into cohesive units (Jahan et al., 2021), they have their disadvantages. First, they require taxonomies to be known *a priori*. Second, while taxonomies like the HVV classification are applicable to specific areas (like political news), the terminology and framework may not apply as well to other settings. Finally, the fixed nature of these labels restricts the level of resolution available—rendering it of limited value for analysis of more nuanced narratives or character types. All of these limitations call for methods that can learn taxonomies through data-driven means.

Earlier works exploring such taxonomy-free approaches include those by Bamman et al. (2013) and Card et al. (2016) who employed probabilistic Dirichlet models to learn character "personas." Our framework similarly identifies character types in a bottom-up fashion, but is rooted in labelling the functional role a character plays in the narrative. Closely related to our concept of character role labels is that of social labels from social label theory, which studies how labels impact personal identity and people's worldview (McConnell-Ginet, 2003).

Coupled with the compositional design of our labels, which allows for levels of granularity, our goal is to achieve a middle-ground between the detail of descriptive methods and the higher-order insights afforded by pre-defined taxonomies.

We proceed with a more in-depth explanation of character role labels, as well as a definition of our task and evaluation criteria.

3 Task Definition

3.1 Character Role Labels

In narrative theory, a character role is the functional position or purpose a character occupies within a story as defined by the character's actions, relationships, and contribution towards advancing the plot. Some well known fictional examples include the "Sidekick" (e.g. Dr. Watson— *Sherlock Holmes*) or "Reluctant Hero" (e.g. Luke Skywalker—*Star Wars*), but they also apply to portrayals in news, and can differ substantially between outlets. For example, Elon Musk and Jeff Bezos are depicted as "Tech Visionaries" in some outlets (Pandolfo, 2025) but "Corrupt CEOs" in others (Zickgraf, 2022).

Similar to tropes and archetypes, character roles form general classes that group characters into repeatable types based on similar traits, functions, or patterns of behavior. However, akin to character descriptions and summaries, character roles are both descriptive and predictive: Sidekicks can be expected to help out the main character and Corrupt CEOs will disregard ethics in their pursuit of wealth and power. In this way, character role labels help strike a balance between the detail of character descriptions with the transferability and generalizability of archetypes.

3.2 Definition

Given a text and a character from that text, we define the task of *Taxonomy-Free Character Role Labeling* (TF-CRL) as: generating a role label for a given character based on their narrative function as they are depicted by the author in the text.

We define a character as "an animate being that is important to the plot" (Jahan and Finlayson, 2019), including people, nations, organizations, and ethnic groups. "Characters" thus encompass both individuals and collective entities. We define a "role" as a

Criterion	Description
Faithfulness	Is the label consistent with the text? Does it contain hallucinations?
Relevance	Does the label capture the most salient information of the portrayal?
Informativeness	Does the label provide useful information about the portrayal? Can it be used to predict future behavior?
Generalizability	Can the label be applied to characters in other stories?

Table 2: Descriptions of our evaluation criteria for character role labeling.

functional position or purpose a character occupies within a story as defined in section 3.1.

Basing the role label on "authorial depiction" means we do not rely on contextual information outside of the text to determine the label. This is an important part of our definition as our framework relies on the concept of *diegetic function* (Propp, 1968)—the specific role given to the character *in the text* by the narrator.

Below, we identify four evaluation criteria for establishing good role labels. To the best of our knowledge, neither narrative nor social label theory has addressed the question of what makes a "good" role label.

3.3 Evaluation

As an open-ended task, TF-CRL involves an element of subjective interpretation where multiple reasonable answers can coexist. In Table 1, the striking teacher can be seen as both a victim and a protester, just as Mark Zuckerberg can be labeled both a victim and a visionary. A second, related challenge is that of label granularity, where determining the appropriate level of specificity can also be fluid. Should the striking teacher be labeled as a Protester, a Righteous Protester, or a Righteous Government Protester?

These multi-answer and "Goldilocks" problems are common issues in areas such as text summarization (Akkasi et al., 2023; Zhong et al., 2022). We therefore draw inspiration from that field in creating our evaluation criteria.

As a preliminary framework, we propose the following evaluation criteria described in Table 2. *Faithfulness* and *Relevance* are direct analogues from summarization, and evaluate how well a label aligns with the source text. *Faithfulness* refers to the *accuracy* of the label in reflecting the character's actual behavior, attributes, or narrative role, while *Relevance* concerns the *saliency* or thematic importance of the aspects of the character that the label highlights, given the narrative context (Fabbri et al., 2021; Guo et al., 2023).

By contrast, *Informativeness* and *Generalizability* describe the tension between narrative specificity and broader interpretive utility. Just as in the "Protester" versus "Righteous Protester" example, labels can vary in detail and contextual richness. *Informativeness* captures this dimension by indicating how much a label reflects what makes a character unique in a given story (e.g., "Righteous Government Protester" conveys both moral judgment and institutional context). This parallels the informativeness of a summary, which may be more or less detailed. *Generalizability*, on the other hand, has no analog in summarization, and is rooted in the tradition of character archetypes in narrative theory (Jung, 1972; Campbell, 1949). Just as archetypes like "Hero" or "Mentor" are meant to recur across stories, a generalizable role label (e.g., "Protester") abstracts away from specific narrative details to highlight a function that is applicable across different contexts. Such labels may lose nuance but allow for easier aggregation and comparison across texts.

Notably, these are conceptual criteria for which there are a variety of existing, imperfect measures. As in the text summarization literature, refining measurement techniques will be an important direction for future work.

4 Methods

We formulate our task as producing a machine-generated character role label. While there are many ways to do this, we employ LLMs with different prompt pipelines to generate labels. Other approaches, particularly using other transformer-based methods, could be used, however we elect to use LLMs due to the lack of large-scale annotated data and to take advantage of LLMs labeling capacity.

For validation, we employ crowd-workers to rank the LLM labels to those from humans and rate them according to our evaluation criteria.

To further operationalize our task and enhance

our method's flexibility for both informativeness and generalizability, we employ a compositional framework for our labels. Each role consists of a primary label (PRIM) that names the core role, and an optional modifier (MOD) that describes how the role is enacted. For Righteous Protester in Table 1, "Protester" is the primary label and "Righteous" is the modifier. We refer to the combination of the two as the full label, as in Figure 1. In this way, the primary label provides a more generic role while the full label provides a more informative one. Importantly, the primary label must describe the character's role even without the modifier and must represent a single, narratively-relevant function. As a result, primary labels need not be a single word: "Public Servant", for instance, is a single primary label since "Servant" loses essential narrative meaning.

4.1 LLM Implementation

We employ the following LLMs: GPT-4o (OpenAI et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024), and Llama 3.1 (Meta, 2024) since they are among the highest ranked proprietary and open-source LLMs on common leaderboards.[2] We further use large models from each respective LLM provider due to the complexity of the task, namely gpt-4o-2024-08-06, claude-3-5-sonnet-20241022, and meta-llama/llama-3.1-405b-instruct. GPT-4o and Claude were accessed via OpenAI and Anthropic's respective APIs, and Llama through the OpenRouter LLM interface. All experiments use a temperature of zero and the system prompt: "You are a helpful assistant."

For prompting, we use zero-shot generation. We experiment with other prompt pipelines, particularly one adapted from topic discovery research (Lam et al., 2024). However results from crowdworkers indicated no framework to be meaningfully preferable to any other (Appendix A.3). We hence employed the simplest one for cost efficiency.

4.2 Human Evaluation

To validate LLM outputs, we involve human participants at two levels. First, we rely on trained undergraduate research assistants to generate a reference set of character role labels. Four RAs were provided with a codebook and underwent two rounds of practice annotations to ensure consistency of interpretation of definitions (see Appendix A.2.1). Second, we run a crowd-sourced ranked choice study to understand reader preferences for candidate role labels.

Workers from Amazon Mechanical Turk (AMT) were shown the full article, a target character, and the candidate labels for that character from all LLMs and one randomly chosen human label. Workers were asked to rank them from best to worst and rate each label according to the criteria in Table 2. Faithfulness was rated on a binary "yes/no" scale following Zhang et al. (2024), while all other criteria used a 3-point Likert scale.

To ensure quality responses, only workers with a lifetime success rate of 98% were allowed to participate. To ensure basic comprehension, a random label was included among the 4 others which had to be ranked lowest for the response to be accepted. Each article received three independent responses, and label order was randomized per task. Additional details can be found in Appendix A.2.2.

5 Validation

To validate the LLMs thoroughly, our validation dataset is built with a range of news outlets and categories. We use a sample of 120 news articles from the Global News dataset consisting of English articles from four news outlets encompassing both Eastern and Western sources (ABC News, Al Jazeera, BBC News, and The Times of India) and ten news categories (Jobs, Health, Sports, Technology, Climate, Politics, the Israel-Palestine conflict, the Russia-Ukraine conflict, China, and the United States). The mean length of articles is 727 words. More details can be found in Appendix A.1.

Prior to labeling, characters were initially identified using GPT-4o and ranked by importance in the text. Human evaluation deemed GPT proficient at this task (Appendix B).

To assess LLM capabilities across the spectrum of character importance while maintaining reasonable cognitive load for human annotators (RAs), we selected two main and two minor characters randomly from each article for labeling by both LLMs and human annotators. When fewer than four characters were present, all characters were labeled. We defined main characters as those ranking in the top 2 of GPT-4o's importance list, with all others classified as minor characters. While this binary distinction oversimplifies a natural spectrum, Figure 5 in Appendix A.4 empirically shows no qualitative difference between these groups in our

[2] Chatbot Arena

	Rank	Eval. Criteria			
		Faith.	Rel.	Info.	Gen.
GPT	2.05*	0.97*	2.79*	2.75*	2.78*
Llama	2.28*	0.95*	2.60	2.56	2.63
Claude	2.33*	0.95*	2.56	2.55	2.61
Human	2.78	0.81	2.48	2.43	2.58
Correlation with Rank					
Kendall's τ	-0.32	-0.60	-0.60	-0.48	
Spearman's ρ	-0.35	-0.66	-0.66	-0.53	

Table 3: Average rank and evaluation ratings for the character role labels from the AMT survey. Ranking and faithfulness closer to 1 are better. For the other evaluation criteria, closer to 3 is better. * indicates statistical difference compared to humans under a Wilcoxon signed-rank test with $p < 0.01$. Bottom two rows show the correlation of the evaluation criteria with rank. All correlations are statistically significant with $p < 0.01$.

AMT preference survey.

From these four labeled characters, a random character was selected for validation on AMT under the constraint that an equal number of main and minor characters were present. To further validate across label formats, 50% of the articles compared only primary labels with the other half comparing only full labels. This was split evenly between main and minor characters. For the entity type breakdown of the characters in the validation, see Table 7 in Appendix A.4.

Table 3 reports the average rankings and evaluation scores for each labeler. GPT outperformed all other models, including humans, across all metrics. Differences in ranking and faithfulness between LLMs and humans were statistically significant. For the other criteria, only GPT showed a significant advantage over human annotators. GPT also outperformed Claude and Llama in both rankings and evaluation scores.

These findings held across various conditions, including character prominence (main/minor), label format (primary/full), news outlet, article topic, and character entity type (Figure 5 in A.4).

Sample label rankings can be found in Table 10 in Appendix A.4. Overall, crowd-workers consistently favored more relevant and informative labels—for example, choosing "Labor rights champion" over a less descriptive "Labor champion." Rankings were most strongly correlated with relevance and informativeness, suggesting these dimensions were the primary drivers of label quality judgments.

Inter-annotator agreement was moderate with Kendall's W and Krippendorff's alpha values ranging from 0.28 to 0.64 indicating reasonable reliability across both ranking and rating tasks given their inherent subjectivity. See Table 11 in Appendix A.4 for more details.

Given the known output variability of closed-source models at a temperature of zero, we further verified LLM output consistency by collecting 10 independent responses from our validation dataset for each of the LLMs. The results are given in Table 12 of Appendix A.4.2, and show that GPT and Claude produce identical responses in more than 8 out of 10 runs on average.

While all LLMs proved capable, we use GPT in the case study below due to its high performance.

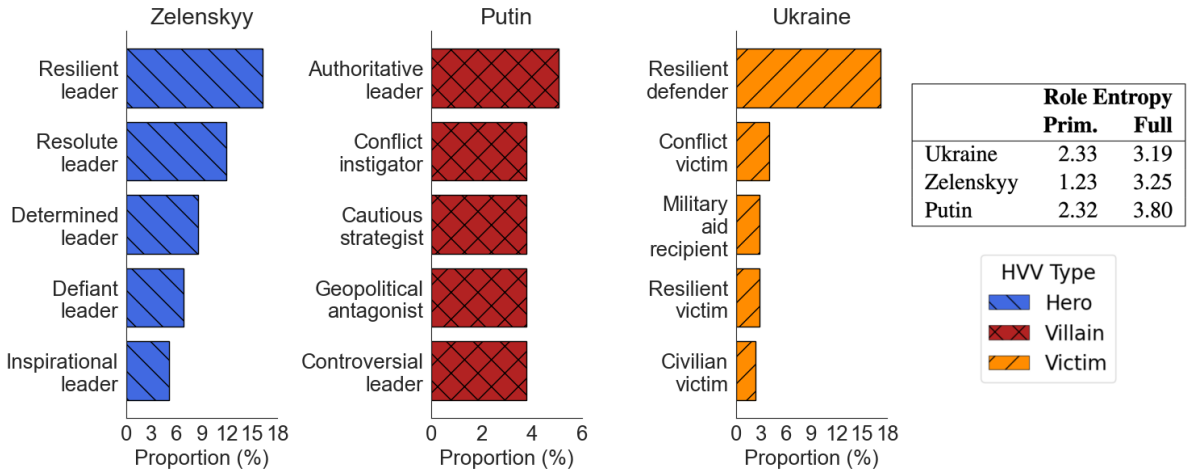
6 Application

We now present a case study to showcase the analytical affordances of character role labelling. We focus first on character-level analyses followed by higher-level narrative and cross-narrative affordances.

We apply our labeling framework to articles from 5 different news topic datasets to illustrate its flexibility: (1) the Russia-Ukraine conflict, (2) immigration in the US, (3) same-sex marriage (SSM) in the US, (4) climate change, and (5) technology. The Russia-Ukraine and Technology articles are from the same source dataset as the validation dataset. The remaining articles are a subset of the Media Frames Corpus (Card et al., 2015), a dataset consisting of articles from U.S.-based outlets, like the Washington Post and the New York Times, originally compiled for media framings research. Specifically we use a random sample of 4000 articles for each category. For more details on the datasets, including a further breakdown of dataset composition, see Table 15 in Appendix C.1. Across all articles, there are an average of 4.6 characters per article.

Prior to post-processing, GPT-4o produced between 479 to 1779 distinct primary labels, 815 to 3754 modifiers, and 1870 to 10,525 full labels in the five datasets, amounting to a total of 3559 unique primary labels, 9994 modifiers, and 31,192 full labels. Although many of these are highly similar, this large number of labels is one of the main limitations of our open vocabulary approach.

To partially address this, we employ clustering



	Role Entropy	
	Prim.	Full
Ukraine	2.33	3.19
Zelenskyy	1.23	3.25
Putin	2.32	3.80

Figure 2: Top 5 most frequent full labels for the top occurring characters in the Russia-Ukraine dataset. Bar color indicates division into the HVV types. Proportions represent the percentage of articles in which the character is assigned that role. Inset shows the role entropy for the top-50 primary and full labels for each character.

of the label embeddings to resolve basic synonymy, specifically using agglomerative clustering with sentence-transformer embeddings (Reimers and Gurevych, 2019). Further details can be found in the Appendix C.2.2.

Still, grouping character roles by semantic similarity faces nuanced difficulties since word embeddings do not naturally capture some important aspects of character roles. Many embedding models, for example, struggle to differentiate opposing roles like "Protagonist/Antagonist" and "Prosecutor/Defendant"—reflecting well-known challenges in distinguishing antonyms with text embeddings (Ali et al., 2024). Additionally, full labels with the same primary label or modifier can have high similarity even when they convey opposing valence (e.g. "Cruel/Heroic Leader" and "Diplomatic Instigator/Mediator"), and embedding models do not always distinguish well between agent versus patient roles (e.g. "Accuser/Accused"). Further examples are compared in Table 18 of Appendix C.2.1.

While resolving some of these issues may be achievable using LLM-based clustering approaches (Pham et al., 2024), a more fundamental challenge lies in defining what it means for character role labels to be the "same." Indeed, the notion of label equivalence, and how best to group labels together, is often application-specific and is closely related to the problem of assessing cluster validity in the topic modeling literature (Hoyle et al., 2022).

Given its complexity, we consider this problem substantial enough to warrant its own study and therefore leave it as future work. For the purposes of this paper, we instead employ clustering to resolve basic label synonymy. To that end, we employ the stsb-roberta-base-v2 embedding model, trained on NLI data, since it better distinguishes antonyms. We further cluster primary labels and modifiers separately to normalize each lexicon. Full labels are then normalized using the normalized MOD + PRIM combination. This reduces the label space from 3559 primary labels to 1190, 9994 modifiers to 3235, and 31,192 full labels to 25,458. For the breakdown by dataset, see Table 19 in Appendix C.2.2.

6.1 Enhanced Character Resolution

We first show how our method can be used in tandem with the Hero-Villain-Victim (HVV) typology to provide greater resolution into specific character roles.

Figure 2 shows the top 5 most frequent full-labels for three of the top characters in the Russia-Ukraine dataset: Volodymyr Zelenskyy, Vladimir Putin, and Ukraine. To label each character within the HVV taxonomy, we use GPT-4o to classify each as "hero", "villain", "victim", or "neither" which is shown by the color of the bars. For this classification, GPT-4o achieved an average F1 score of 88% with more details in Appendix C.3.

Differentiating Role Expressions As seen in Figure 2, Zelenskyy, Putin, and Ukraine strongly correspond to the respective roles of hero, villain, and victim. Across all articles, Zelenskyy, Putin, and Ukraine are respectively portrayed as a hero, villain, and victim, in 82%, 92% and 87% of arti-

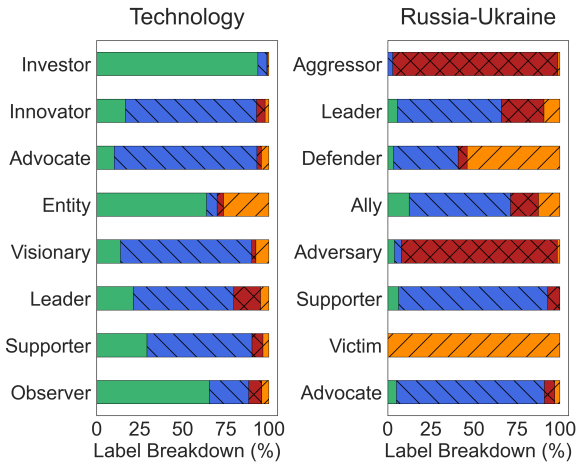


Figure 3: Breakdown of top 8 primary labels into the HVV taxonomy. Green (solid) bars indicate "neither/none." For the other colors, see Figure 2.

cles. These are consistent with the role labels each received. Notably, both Zelenskyy and Putin are identified as "Leaders"—highlighting their functional similarities. However the valence between their leadership roles are distinguished by their modifiers: Zelenskyy as "Resilient" and "Resolute" (heroic attributes), compared to Putin as "Authoritarian" and "Controversial" (antagonistic attributes). In contrast, Ukraine is often explicitly labeled as a victim or a "Resilient defender" flagging distinct functional and semantic roles compared to the other two characters.

The labels in Figure 2 also convey more nuance. Zelenskyy is resilient, defiant, and inspirational, Putin is a strategist in addition to a leader, and Ukraine can be a defender, a conflict victim, and a recipient of military aid. None of these distinctions are observable in a general HVV framework.

While not shown here, Figure 6 in Appendix C.4 illustrates how our method can lend insight into more heterogeneous characters. There, Barack Obama in the Immigration dataset is a mix of hero, villain, and victim roles, with his primary victim role being a "Political target", a very different kind of victimhood than Ukraine's "Resilient defender."

Character Role Entropy An interesting point of analysis concerns the differences in the label distribution shapes which can be interpreted as relating to the narrative consistency surrounding the entity. For instance, while Putin is strongly portrayed as a villain, there is a high level of variance around his villainous roles, whereas the victimhood of Ukraine heavily centers around being a "Resilient defender."

Climate	Immigration	Rus-Ukr
Advocate	**Advocate**	Aggressor
Authority	Enforcer	Leader
Critic	**Authority**	Defender
Leader	**Critic**	Ally
Skeptic	Victim	Adversary

Table 4: Top 5 ranked primary labels in the Climate, Immigration, and Russia-Ukraine domains. Roles between Climate and Immigration are more similar than those in Russia-Ukraine. Bold labels are shared between Climate and Immigration but not with Russia-Ukraine.

To quantify this, we propose using the distribution entropy, which we refer to as the *role entropy*, for the top-k labels. The inset in Figure 2 gives the role entropy values for these 3 characters using the top-50 labels to account for differing tail lengths of the distributions. While these are consistent with the plots in Figure 2, they further illustrate that the entropy of the primary and full labels may differ.

6.2 Novel Roles Beyond HVV

Although we've focused on the HVV framework on the character level, our method can also help uncover other typologies on the narrative level that may not fit into this framework as cleanly. Figure 3 shows the top 8 primary labels in the Technology and Russia-Ukraine datasets as well as how characters occupying those roles were divided into the HVV classes. Green (solid) bars indicates "neither/none." On the right, the Russia-Ukraine labels strongly align with the HVV types in some capacity, however many in Technology, notably investors, observers, and entities (which correspond to corporate entities), do not. This distribution suggests the Technology articles are more centered on investor-innovator relationships rather than standard heroes and villains. This reflects one of the major advantages of TF-CRL, which allows more organic roles and typologies to surface.

6.3 Cross-Narrative Role Comparisons

While the aforementioned affordances of TF-CRL lie in their capacity to differentiate and discover character roles within individual narratives, another major contribution is the ability to generalize across narrative domains. We term this capacity *cross-narrative role similarity*—the extent to which character roles recur in structurally comparable ways across otherwise distinct topical contexts. A role-

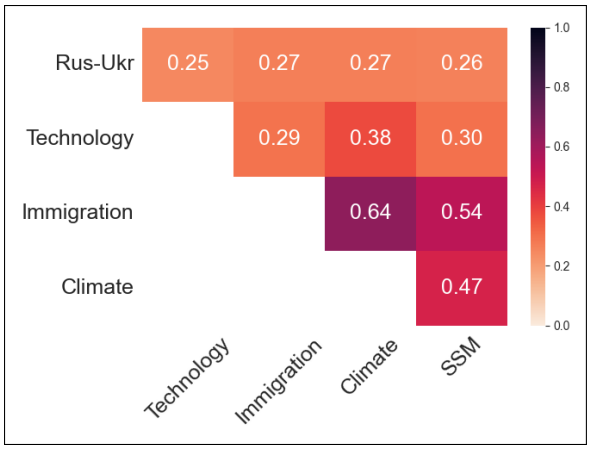


Figure 4: Cross-narrative role similarity measured via RBO. Imigration, Climate, and SMM have more similar roles compared to Technology and Russia-Ukraine. Values closer to 1 indicate higher similarity.

labeling system that only functions at the document level provides limited interpretive power; to be useful at scale, it must also support comparative analysis across narratives, domains, and genres.

To quantify patterns of cross-narrative generalization, we compute pairwise similarity scores between datasets based on their character roles ranked by frequency. Specifically, we employ Rank-Biased Overlap (RBO; Webber et al., 2010), a measure designed for comparing ranked lists of unequal lengths and non-identical items, that is top-weighted to give greater importance to higher-ranked items. Scores range from 0 (no similarity) to 1 (identical rankings). Following Webber et al. (2010), we use a persistence parameter of $p = 0.9$ to prioritize the top 10 roles in each domain.

Figure 4 presents the resulting RBO similarity matrix. Immigration, Climate, and SSM display the highest degrees of similarity, suggesting that these domains tend to foreground comparable role types, particularly those associated with advocacy, authority, and critique. Technology is more weakly aligned, showing modest overlap with Climate, while the Russia-Ukraine dataset diverges most strongly from all others, reflecting its distinctive focus on roles tied to conflict.

To illustrate the underlying role structures driving these scores, Table 4 displays the top five ranked labels in three sample domains. Immigration and Climate both prioritize "Advocate," "Authority," and "Critic," indicating a shared emphasis on institutional actors and value-based conflict—hallmarks of socio-political discourse. By contrast, the Russia-Ukraine dataset is anchored on roles such as "Aggressor," "Defender," and "Ally," highlighting its narrative framing around military and geopolitical conflict. While not shown in this table, the role "Advocate" is prominent across all three domains, but its relative contextual meaning varies significantly.

7 Conclusion

In this paper, we introduced the task of character role labeling, grounded in a character's narrative function, to balance between the descriptiveness of character descriptions and the recurring patterns of archetypes. We formalized the task of TF-CRL and showed that three LLMs are proficient at this task in a news context. We further demonstrated several affordances of this approach: on the character-level, by providing enhanced resolution with pre-existing taxonomies, on the narrative level, by aiding in the discovery of new character roles, and on the cross-narrative level, by comparing roles across domains.

TF-CRL opens many opportunities in comparative narrative analysis and media studies research, particularly for data at-scale. Our approach offers avenues for studying portrayals of particular people and how they vary across outlets, culture, and time. This would be useful for political science research, but also offers applications outside of research, like for NGOs and health organizations interested in depictions of marginalized or stereotyped communities (Chen et al., 2023; Gottipati et al., 2021). In general the versatility of the concepts we introduce is one of its key strengths: role entropy can equally be applied to roles on a narrative-level as to on an individual character-level, and cross-narrative role similarity can be applied to specific characters across outlets instead of general roles across domains.

This work leaves many open questions for future research, including how to reduce the label space effectively, and how to expand these approaches beyond individual roles to capture relationships between characters. These we leave as future work, including validating these methods to cover fiction.

As the world becomes increasingly interconnected, and in many ways, more polarized, it will be essential to understand the narratives people tell and the impacts they have on society. It is hoped this research helps bolster the tools in that arsenal and will help to better understand how narratives shape societal discourse and perceptions.

Limitations

In our work, we employed only large models from each LLM provider and did not test the proficiencies of small- and medium-sized models due to the high costs of the crowd-sourced validation. Although it is hypothesized that smaller models will be less capable at our task due to their limited reasoning abilities, validating and benchmarking smaller models will be an important area of future work to make these techniques more accessible.

Also related to validation, we only validated LLM performance on English articles and not those of any other language. Although non-English articles could be translated into English prior to labeling, generating labels directly from the source language may help preserve cultural nuance and specific character framings that may get lost in translation. Therefore, expanding our validation to additional languages will be an important future step.

As mentioned previously, one of the central limitations of our method is the very large number of labels it produces. Clustering reduced the overall number of primary labels and modifiers by roughly two-thirds, but more than 1000 distinct primary labels and 3000 modifiers remained. Although this high level of detail can be advantageous (e.g. identifying Zelenskyy as a resilient, defiant, and inspirational leader in Figure 2) and can aid in role discovery, it can also introduce interpretive challenges. More research is therefore needed to develop more principled ways to assess character role equivalence and collapse the label space to a more manageable size. It is important to note, however, that there is no single "correct" solution. Prior work using LLMs to provide feedback during clustering represents a promising direction in this regard (Zhang et al., 2023). In particular, approaches that incorporate feedback at multiple stages of the clustering process (Viswanathan et al., 2024) may be especially valuable for specific application domains with targeted needs and requirements.

From a broader perspective, our task's focus on individual narrative roles also comes at the expense of other key dimensions like character relationships. An "Ally", in our case, does not indicate with whom the character is an ally. Although label co-occurrence across articles can be used an as approximation of relationships, extensions of our task to explicitly capture character relationships would create a more complete narrative picture and lend itself to analysis tools from network science.

A final limitation is our method's reliance on LLMs and the potential that creates for biases to emerge in assigned roles, especially in social or political domains where LLMs are known to have specific leanings on certain issues (Feng et al., 2023; Rettenberger et al., 2025). As mentioned in our task definition, it is vital that labels align with the author's portrayal of the character in the text and not any potential LLM biases. While the diversity of the news sources in our validation helps account for this, and shows (at the scale of our validation) that LLMs can accurately assign roles regardless of an article's leaning, further targeted validation of these methods to additional news sources will increase trust and confidence in these systems. This will ideally include validating across a broader range of issues with annotators and crowd-workers from diverse cultural backgrounds.

Ethics Statement

We develop these techniques for the purposes of better understanding how people and groups are portrayed in news as well as the narratives and stories that involve them. However in contexts involving real people, such as these, labels may be closely linked to harmful stereotypes and could be used to typify people in undesirable ways or advance defamatory aims. We therefore advocate for the responsible use of these systems, and for critical thought and thoughtful care to be employed on the part of researchers to avoid the spread of damaging stereotypes.

Acknowledgments

We would like to thank the reviewers for their constructive feedback and suggestions. This research was supported by the Fonds de recherche du Québec (FRQ) in Canada.

References

Abbas Akkasi, Kathleen C Fraser, and Majid Komeili. 2023. Reference-free summarization evaluation with large language models. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 193–201.

Muhammad Asif Ali, Yan Hu, Jianbin Qin, and Di Wang. 2024. Antonym vs synonym distinction using interlaced encoder networks (ICE-NET). *arXiv preprint arXiv:2401.10045*.

Carolyn Jane Anderson. 2025. Components of character: Exploring the computational similarity of austen's characters. *Journal of Data Mining & Digital Humanities*.

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

Sabyasachee Baruah and Shrikanth Narayanan. 2024. Character attribute extraction from movie scripts using llms. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8270–8275. IEEE.

Kelly Bergstrand and James M Jasper. 2018. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81(3):228–247.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "Let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joseph Campbell. 1949. *The Hero with a Thousand Faces*. Pantheon Books, New York, NY.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

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

Dallas Card, Justin H Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1410–1420.

Keyu Chen, Marzieh Babaeianjelodar, Yiwen Shi, Kamila Janmohamed, Rupak Sarkar, Ingmar Weber, Thomas Davidson, Munmun De Choudhury, Jonathan Huang, Shweta Yadav, Ashiqur KhudaBukhsh, Chris T. Bauch, Preslav Nakov, Orestis Papakyriakopoulos, Koustuv Saha, Kaveh Khoshnood, and Navin Kumar. 2023. Partisan US News Media Representations of Syrian Refugees. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:103–113.

Alan Dundes. 1962. From etic to emic units in the structural study of folktales. *The Journal of American Folklore*, 75(296):95–105.

Lynn S Eekhof, Kobie Van Krieken, and Roel M Willems. 2022. Reading about minds: The social-cognitive potential of narratives. *Psychonomic Bulletin & Review*, 29(5):1703–1718.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Monika Fludernik. 2002. *Towards a'natural'narratology*. Routledge.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing Climate Change Policy Narratives with the Character-Role Narrative Framework. *SSRN Electronic Journal*.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the hero, the villain, and the victim? detection of roles in news articles using natural language techniques. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 311–315.

Swapna Gottipati, Mark Chong, Andrew Kiat, and Benny Kawidiredjo. 2021. Exploring Media Portrayals of People with Mental Disorders using NLP:. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 708–715, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Algirdas Julien Greimas. 1983. *Structural semantics: An attempt at a method*. University of Nebraska Press.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2023. Appls: Evaluating evaluation metrics for plain language summarization. *arXiv preprint arXiv:2305.14341*.

Alexander Gurung and Mirella Lapata. 2024. CHIRON: Rich character representations in long-form narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8523–8547, Miami, Florida, USA. Association for Computational Linguistics.

Eric Holgate and Katrin Erk. 2021. "politeness, you simpleton!" retorted [MASK]: Masked prediction of literary characters. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 202–211, Groningen, The Netherlands (online). Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. 2022. Learning and evaluating character representations in novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019.

Labiba Jahan and Mark Finlayson. 2019. Character identification refined: A proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Labiba Jahan, Rahul Mittal, and Mark Finlayson. 2021. Inducing stereotypical character roles from plot structure. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 492–497, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karen S Johnson-Cartee. 2005. *News narratives and news framing: Constructing political reality*. Rowman & Littlefield.

Michael D. Jones and Mark K. McBeth. 2010. A Narrative Policy Framework: Clear Enough to Be Wrong? *Policy Studies Journal*, 38(2):329–353. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0072.2010.00364.x.

Carl Gustav Jung. 1972. *Collected works of CG Jung, volume 7: Two essays in analytical psychology*, volume 20. Princeton University Press.

Johanna Kuenzler, Bettina Stauffer, Caroline Schlaufer, Geoboo Song, Aaron Smith-Walter, and Michael D. Jones. 2025. A systematic review of the narrative policy framework: a future research agenda. *Policy & Politics*, 53:129 – 151.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using LLooM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Raymond A Mar. 2018. Stories and the promotion of social cognition. *Current Directions in Psychological Science*, 27(4):257–262.

Sally McConnell-Ginet. 2003. "What's in a name?" Social labeling and gender practices. *The handbook of language and gender*, pages 69–97.

Meta. 2024. Introducing llama 3.1: Our most capable models to date.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, et al. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Chris Pandolfo. 2025. Elon Musk, AI and tech titans, venture capitalists invited to pre-inauguration dinner at dawn of trump era. *Fox News*.

Argyrios Papoudakis, Mirella Lapata, and Frank Keller. 2024. BookWorm: A dataset for character description and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4471–4500, Miami, Florida, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas press.

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.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

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.

Hans-Jörg Uther. 2004. *The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson*. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, Helsinki.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontanón. 2014. Toward automatic role identification in unannotated folk tales. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 10, pages 188–194.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7476–7483.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised multi-granularity summarization. *arXiv preprint arXiv:2201.12502*.

Ryan Zickgraf. 2022. Jeff bezos's resemblance to lex luthor isn't just skin-deep. *Jacobin*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Validation

A.1 Validation Dataset

The validation dataset was sampled from the Global News dataset; a set of English articles sourced from the NewsAPI between October 1st to November 29th 2023. The dataset is covered under the public domain (CC0).

Articles spanned four news outlets (two from Eastern sources: Al Jazeera English and The Times of India, and two from Western sources: BBC News and ABC News) and ten news categories (Jobs, Health, Sports, Technology, Climate, Politics, the Israel-Palestine conflict, the Russia-Ukraine conflict, China, and the United States).

We used 120 news articles split equally between the four news outlets and ten news categories; hence, 3 articles were present for each outlet-category combination. The mean length of articles is 727 words with a minimum of 200 and a maximum of 2,000 words.

The total number of characters (identified by GPT-4o) was 704, with 5.9 characters found per article on average. For more details on character identification, see Appendix B.

The validation dataset is available on our GitHub repository.

A.2 Human Annotation Details

All codebooks and HTML templates (for our AMT surveys) are available on our GitHub repository.

A.2.1 Human Character Label Annotations

Human label annotations were collected for each of the 120 articles in the label validation dataset. These were provided by trained undergraduate research assistants. Participants were paid an hourly wage above minimum wage and were aware they were participating in a research study.

Annotators were provided with a codebook of definitions and examples, and underwent at two round of practice annotations to affirm consistency of interpretations to the definitions. Human responses adhered to the modifier/primary-label structure, and were made independently of each other and from the LLMs.

A.2.2 Amazon Mechanical Turk Annotations

The following applies for both the prompt selection and label validation AMT surveys.

Three responses were collected for each article. To ensure quality responses, we required workers to have a lifetime success rate of more than 98%. Along with the main labels being considered, a random label was also ranked along with the other labels as a basic comprehension question. This label was manually created for each article by one of the authors, and workers had to give this label the lowest ranking and rating (in terms of faithfulness and relevance) for their answers to be accepted. Workers that had previously answered a HIT incorrectly were excluded from submitting future HITs. All labels were randomly ordered on each survey.

To partially address concerns among researchers of crowd-workers using ChatGPT or other LLMs to answer the questions, all articles were provided as images.

All workers were compensated at a rate of $2/HIT (USD). The estimated time per HIT was 8 minutes thus translating to an hourly wage of $15/hour. Participants were aware they were participating for academic purposes. No geographic restrictions were placed on the AMT workers, and so no cultural representation data for the workers who participated in our survey is available.

A.3 Prompt Selection

This subsection provides details on the prompt selection from section 4.1.

We experimented with 3 prompt pipelines based on the following framework inspired by Lam et al. (2024) for topic discovery:

1. Extract quotes from the source text relevant to the portrayals of the given characters
2. Generate summaries for each of the characters
3. Generate a character role label for each of the characters.

Prompt pipelines accomplished these steps in succession (using separate API calls) where the outputs of one step served as the inputs to the next step. Following Lam et al. (2024), we refer to this as the concept induction framework.

Alternatively, we experimented with generating summaries for the characters directly followed by label generation (steps 2 + 3; two successive API calls), and zero-shot label generation (step 3 only; one API call) which we refer to as the character summary and zero-shot methods, respectively.

Each of these prompt frameworks were used to collect character role labels for each of GPT-4o, Claude, and Llama for a given article and list of

characters. We used crowd-workers from AMT to rank which labels they preferred.

We used a dataset of 60 news articles from the same source dataset as the label validation dataset in A.1. It was disjoint from the label validation dataset, but consisted of the same news outlets and article categories with 6 articles/category and between 14-16 articles/outlet.

Similarly to the main label validation, characters were initially identified in each article using GPT-4o. The total number of characters identified was 348 (5.8 characters/article on average). From these characters, two main and two minor characters were randomly selected in each article for labeling by the LLMs to mimick the label validation (see Appendix B for details).

For the AMT survey, a random character was selected for label comparison for each article. Crowd-workers were provided with the article text, the character name, and the labels generated from each prompt framework. Crowd-workers were tasked with ranking the labels from best to worst. This was done separately for each LLM.

To save on survey costs, only 20 articles were used to test the prompt preferences for Claude and Llama. All 60 articles were used for GPT-4o. In the end, the same trends were observed with Claude and Llama as with GPT-4o, therefore, we did not test with more articles.

The following subsection shows the prompt selection results.

A.3.1 Prompt Selection Results

For GPT-4o, 57 out of 60 articles had a character for which there were at least two distinct labels from the three prompts (i.e. for 3 articles, all characters were given the same label across all the prompts). For Claude, this was 19 out of 20 and for Llama, 20 out of 20. Only these articles were ranked on AMT. When ties occurred in the ranking (because different prompts produced the same label), we employed standard competition ranking ("1224" ranking).

Table 5 shows the average rank for each prompt, as well as the percentages with which each prompt achieved each ranking. Although the character summary prompt achieved the best average ranking, no differences among any prompts were statistically significant according to a Wilcoxon signed-rank test with $p = 0.05$. The only exception was between character summary and concept induction rankings for Llama.

	Avg. Rank	Rank % 1st	2nd	3rd
GPT				
Char. Summary	1.75	42.1	40.4	17.5
Zero Shot	1.77	41.5	40.4	18.1
Con. Induction	1.88	40.9	29.8	29.2
Claude				
Char. Summary	1.86	36.8	40.4	22.8
Zero Shot	1.88	40.4	31.6	28.1
Con. Induction	1.89	40.4	29.8	29.8
Llama				
Char. Summary	1.67	41.7	50.0	8.3
Zero Shot	1.75	46.7	31.7	21.7
Con. Induction	1.98	36.7	28.3	35.0

Table 5: Average rank and rank percentages from the prompt selection AMT survey for the different LLMs and prompt frameworks. Rank percentage is the percent of annotations for which each prompt was ranked 1st, 2nd, and 3rd respectively. Percentages do not sum to 100, row-wise, due the presence of ties. Only character summary and concept induction rankings for Llama were statistically different, according to a Wilcoxon signed-rank test with $p = 0.05$. The number of articles were $n_{\text{GPT}} = 57$, $n_{\text{Claude}} = 19$, $n_{\text{Llama}} = 20$.

Table 6 shows the inter-annotator agreement between the different prompts. For articles in which the prompts produced 3 distinct labels, we use Kendall's W coefficient of concordance as an inter-rater reliability measure among the crowd-workers. This was computed from each article individually and averaged over all articles. Since Kendall's W is only suitable for rankings of 3 or more items, we employ Krippendorff's alpha as the inter-rater reliability measure with only two distinct labels.

For GPT, simulations of Kendall's W with random rankings from 3 annotators gave a p-value of 0.03 based on 50,000 runs ($n_{3\text{ labels,GPT}} = 23$). For articles with 2 labels, a binomial test did not show agreement to be statistically different from random ($n_{2\text{ labels,GPT}} = 34$). The quality of the labels from all prompts was generally very similar, however, and therefore we attribute this (and the low Krippendorff's alpha score) to the difficulty of the task and not to misunderstandings of definitions.

Similar inter-annotator agreement was observed for Claude and Llama. The negative values of Krippendorff's alpha is attributed to the difficulty of the task and the small sample sizes.

No. of labels	Agree on Best (%) 1	2	3	Reliability statistic
GPT				
3	17.4	52.2	30.4	$W = 0.44$*
2	–	67.7	32.3	$\alpha = 0.08$
Claude				
3	8.3	66.7	25.0	$W = 0.56$*
2	–	85.7	14.3	$\alpha = -0.09$
Llama				
3	12.5	37.5	50.0	$W = 0.57$*
2	–	83.3	16.7	$\alpha = -0.11$

Table 6: Inter-annotator agreement from the prompt selection AMT survey. "Agree on Best" percentages are the percentage of articles in which 1, 2, and 3 annotators agreed on the best prompt. The reliability statistics correspond to average Kendall's W for 3 distinct labels and Krippendorff's α for 2 distinct labels. The inter-rater agreement with 2 labels was not significant to a level of $p = 0.05$ according to a binomial test ($n_{2\text{ labels,GPT}} = 34$, $n_{2\text{ labels,Claude}} = 7$, $n_{2\text{ labels,Llama}} = 12$). Inter-rater agreement with 3 labels was statistically significant with a one-sided p-value of 0.05 ($n_{3\text{ labels,GPT}} = 23$, $n_{3\text{ labels,Claude}} = 12$, $n_{3\text{ labels,Llama}} = 8$).

A.4 Label Validation

This subsection provides additional results and details from the label validation in section 5. As mentioned in section 4.2, we used the dataset of 120 articles described in A.1.

Analogous to the prompt selection, among the 4 characters labeled by the humans and LLMs for each article, a random character was selected for label comparison for the AMT survey (under the constraint that an equal number of main/minor characters were represented from the sample). Crowd-workers were provided with article text, the character name, and the labels generated by each LLM and a randomly selected human. Crowd-workers were tasked with ranking the labels from best to worst, and rating the labels according to our evaluation criteria.

Of the 120 articles, 115 had characters with at least two distinct labels (5 articles had all LLMs and humans produce the same labels for all characters). In the cases of labelers producing the same label, we employ standard competition ranking ("1224" ranking). Table 7 gives the breakdown of these characters by prominence and entity type.

Character Type / Entity Type	Main	Minor	Total
PERSON	30	32	62
ORG	11	13	24
GPE	14	7	21
NORP	3	5	8
Total	58	57	115

Table 7: Breakdown of characters by main/minor character and entity type for the characters used in the AMT label validation. PERSON indicates a person, ORG an organization, GPE a geopolitical entity or nation, and NORP an ethnic group.

The entity types of characters were identified using spaCy with errors manually corrected.

A.4.1 Label Validation Results

Table 8 and Table 9 give the ranking and evaluation rating percent breakdowns for each of the LLMs and the humans. For some sample rankings from the AMT survey, see Table 10.

While there were many potential confounding variables in our data including news source, news category, character type (main vs minor), entity type, and labels with/without modifiers (full vs primary labels), the relative rankings of the LLMs were stable across almost all categories as seen in Figure 5. In particular, GPT consistently ranked highest and humans the lowest. Although there was some noticeable interchange between Claude and Llama, they both consistently ranked lower than GPT and higher than humans.

Of particular note, LLMs performed better than

	Avg. Rank	Rank Percentage (%) 1st	2nd	3rd	4th
GPT	2.05	36.5	32.2	21.5	9.9
Llama	2.28	31.6	25.8	25.2	17.4
Claude	2.33	29.9	25.8	25.5	18.8
Human	2.78	22.6	16.5	21.2	39.7

Table 8: Average rank and rank percentage breakdowns for the label validation comparing LLM and human character role labels ($n = 115$). Rank percentage is the percent of annotations for which each labeler was ranked 1st, 2nd, 3rd, and 4th respectively. Percentages do not sum to 100, row-wise, due the presence of ties. Differences between GPT and Llama, and Claude and humans were statistically significant under a Wilcoxon signed-rank test with $p < 0.01$.

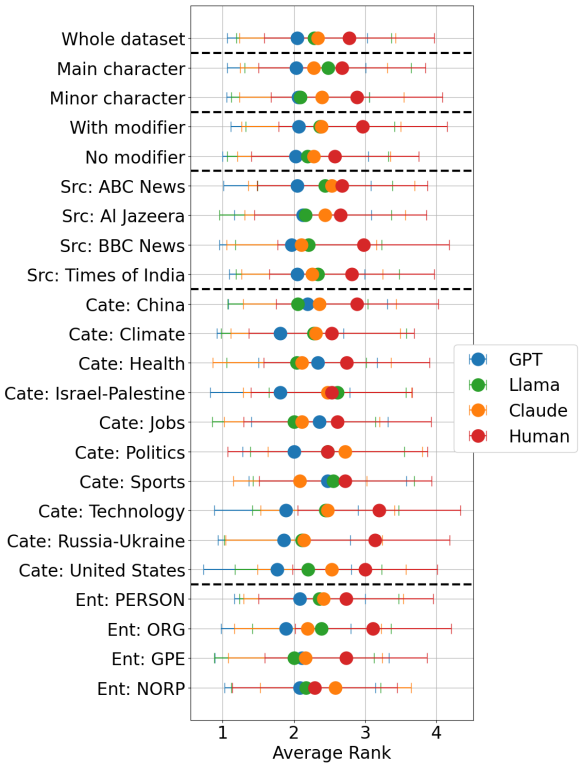


Figure 5: Average AMT rankings by subgroupings. GPT had a consistently high ranking and humans a consistently low ranking. Error bars give standard deviation. Subgroups include main vs. minor characters, labels with/without modifiers, news source (Src), news category (Cate), and entity type. For entity type, PERSON indicates a person, ORG an organization, GPE a geopolitical entity or nation, and NORP an ethnic group.

	Avg. Rating	Rating Percentage (%) 3	2	1
Relevance				
GPT	2.79	80.9	17.4	1.7
Claude	2.56	58.3	39.4	2.3
Llama	2.60	61.5	36.8	1.7
Human	2.48	55.9	36.2	7.8
Informativeness				
GPT	2.75	77.1	21.2	1.7
Claude	2.55	56.8	41.2	2.0
Llama	2.56	58.0	40.0	2.0
Human	2.43	51.0	41.4	7.5
Generalizability				
GPT	2.78	78.8	20.6	0.6
Claude	2.61	61.7	37.4	0.9
Llama	2.63	64.1	35.1	0.9
Human	2.58	62.0	34.2	3.8
			Yes	**No**
Faithfulness				
GPT	0.97		96.8	3.2
Claude	0.95		95.1	4.9
Llama	0.95		94.8	5.2
Human	0.81		80.9	19.1

Table 9: Average evaluation ratings and rating breakdowns for the label validation comparing LLM and human character role labels ($n = 115$). Rating percentages indicate percent of annotations rated 1, 2, or 3 (or Yes/No for Faithfulness). Faithfulness ratings closer to 1 are better. For the other criteria, closer to 3 is better.

humans for both main/minor characters and across entity types (with the exception of Claude on NORP entities) indicating their versatility in handling a variety of different characters.

Table 11 gives the inter-annotator agreement. There, "Agree on Best" gives the percentage of articles in which 1, 2, and 3 annotators agreed on the top-ranked label. For the evaluation criteria, agreement is the percentage in which 1, 2, and 3 annotators agreed on the same rating for all label-article pairs.

In over 50% of cases, at least two annotators agreed on the top-ranked label. Similarly for the evaluation criteria, all three annotators agreed on the same rating at least 44% of the time.

Like for the prompt selection, we employ Kendall's W coefficient of concordance as the inter-rater reliability measure for 3 or more distinct labels and Krippendorff's alpha for articles with two distinct labels. Kendall's W was calculated separately for cases involving 4 distinct labels and 3 distinct labels. We further used Krippendorff's alpha for inter-rater reliability among the evaluation criteria.

Kendall's W values and Krippendorff's alpha values ranged from 0.28 to 0.64 for both the rankings and evaluation ratings, which is reasonable given the inherent subjectivity of the task.

A.4.2 LLM Response Variability

To probe and better quantify LLM output variability (at temperature zero), we collected 10 independent responses from each of the LLMs across our validation dataset. We present the results here.

Table 12 compares the average number of distinct answers across all 10 runs for each of the full labels, primary labels, and modifiers and across all 471 article-character pairs. Included in the table are

Character	1st	2nd	3rd	4th
Iran	Ally^{H}	Backer^{G}	Backer^{L}	Puppeteer^{C}
Dairy Foods Association	Spokesperson^{G}	Messenger^{C}	Informer^{H}	Troubleshooter^{L}
Adidas	Ethical dilemma corporate entity^{G}	Corporate ethics steward^{C}	Fashion tycoon^{H}	Conditional donor^{L}
LGBTQ community	Persecuted vulnerable group^{L}	Targeted vulnerable group^{G}	Persecuted target^{C}	Targeted victim^{H}
Shawn Fain	Labor rights champion^{C}	Worker rights champion^{L}	Labor champion^{G}	Victorious hero^{H}
Sam Altman	Controversial protagonist^{G}	Resurgent protagonist^{L}	Suspicious leader^{H}	Corporate phoenix^{C}

Table 10: Sample label rankings from the AMT label validation. Superscipts on the labels indicate the labeler: C - Claude, G- GPT, H - human, L - Llama.

Ranking				
No. of labels	**Agree on Best (%)** **1**	**2**	**3**	**Reliability statistic**
4	14.3	58.6	27.1	$W = 0.64$
3	22.9	54.3	22.9	$W = 0.44$
2	–	40.0	60.0	$\alpha = 0.48$
Eval. Criteria				
Criterion	**Agreement (%)** **1**	**2**	**3**	$\boldsymbol{\alpha}$
Faith.	–	15.1	84.9	$\alpha = 0.38$
Rel.	0.3	55.8	44.0	$\alpha = 0.29$
Info.	1.2	54.8	44.0	$\alpha = 0.32$
Gen.	0.5	51.6	47.9	$\alpha = 0.28$

Table 11: Inter-annotator agreement for the label validation AMT survey. "No. of labels" represents the number of distinct labels being rated by the AMT workers. "Agree on Best" percentages are the percentage of articles in which 1, 2, and 3 annotators agreed on the best label. For the evaluation criteria, "Agreement" is the percentage in which 1, 2, and 3 annotators agreed when considering all labels and articles. The reliability statistics correspond to average Kendall's W for 3 and 4 distinct labels, and Krippendorff's α for 2 distinct labels ($n_{4\text{ labels}} = 70$, $n_{3\text{ labels}} = 35$, $n_{2\text{ labels}} = 10$).

	GPT	**Claude**	**Llama**
Full Label			
Num. Answers	1.9	1.4	3.5
Maj. Agree. (%)	84.3	92.0	59.4
Full Agree. (%)	43.5	67.9	4.2
Primary Label			
Num. Answers	1.3	1.2	2.3
Maj. Agree. (%)	93.9	96.2	72.8
Full Agree. (%)	74.5	83.4	24.4
Modifier			
Num. Answers	1.7	1.3	2.9
Maj. Agree. (%)	86.9	93.2	66.6
Full Agree. (%)	49.3	72.8	10.6

Table 12: LLM response variability metrics on the validation dataset across 10 runs for each LLM. "Num. Answers" gives the average number of different answers produced across the 10 runs for each article-character pair (closer to 1 is better). "Maj. Agree." (Majority Agreement) is the average percentage of runs which produced the most common output (closer to 100 is better), and "Full Agree." (Full Agreement) is the percentage of article-character pairs in which all 10 runs produced an identical output (closer to 100 is better) ($n = 471$).

also the average percentage of runs that produced the most common (majority) answer, as well as the percentage of article-character pairs in which all 10 runs produced an identical response.

As seen, GPT and Claude exhibited only minor response variability, with more than 8 runs out of 10 producing the same full label on average and more than 9 out of 10 producing the same primary label. Claude was the most consistent of the two, generating an identical full label across all 10 runs for more than 67% of characters. Although GPT

was less consistent, it still produced a single label across all 10 runs with highest likelihood and rarely produced more than two different responses.

Notably, Llama exhibited the most variability. Even still, almost 6 out of 10 runs agreed on the same full label across all 471 characters. Although this is counter-intuitive given that Llama is an open-source model, this can be attributed to our use of the OpenRouter platform (used to host the 405B model) which makes use of different hosting providers across different requests by default. More interestingly, we found that even enforcing a single provider did not guarantee identical responses, and this was true for all providers available through OpenRouter for the 405B model. Pinpointing an exact cause for this is difficult, as it could stem from a multitude of reasons, such as different hardware versions, floating point precision, or quantization on the host platform machines.

Overall however, even when different labels were produced, they were often very similar. As an example, Palestinians were labeled by GPT as "Displaced," "Survivors," and "Endangered group" in one article, all of which were accurate given the article portrayal. Even with Llama, many of the labels were of a similar theme despite the variety of labels. For example, in an article with 6 different labels, NVIDIA was labelled a "High-performing growth stock", "Rate-sensitive high-growth stock", "High-tech growth stock", "Technological growth driver", and a "Technological bellwether".

B Character Identification

Prior to labeling, characters were initially identified by GPT-4o to provide a standard list of characters for each article. Specifically given the article text, GPT was instructed to identify the top 7 characters that contributed meaningfully to the article (following Jahan and Finlayson's (2019) definition of character) or were portrayed strongly by it. As a part of the instructions, GPT was asked to list the characters in descending order of importance in the text. The prompt can be found in our repository.

Human validation of 24 randomly selected articles from the validation dataset showed GPT was proficient at this task. Three graduate students independently assessed (1) whether each character identified by GPT was present in the text and contributed meaningfully to it (on a binary "yes/no" scale), and (2) assessed the quality of GPT's ranking of the characters on a 3-point Likert scale. The results are in Table 13.

	Avg. Rating	Rating Breakdown (%) –	Yes	No
Identif.	0.99		99.1	0.9
		3	**2**	**1**
Ranking	2.69	70.83	27.78	1.39

Table 13: Average evaluation rating and rating breakdowns for the human validation of GPT's character identification and ranking ability ($n = 24$). Rating percentages indicate percent of annotations rated 1, 2, or 3 (or Yes/No for character identification). Identification ratings closer to 1 are better. For Ranking, closer to 3 is better.

	Agreement (%) 1	2	3	α
Identif.	–	2.0	98.0	0.24
Ranking	0	54.2	45.8	0.23

Table 14: Inter-annotator agreement for the human validation of GPT's character identification and ranking ability ($n = 24$). Agreement is the percentage in which 1, 2, and 3 annotators agreed when considering all characters and articles. The reliability statistic is Krippendorff's alpha.

GPT was very proficient at identifying relevant characters, even among articles with many characters or potentially confounding individuals, such as people being interviewed in the article. Annotators agreed with 99% of the characters identified by GPT.

For the rankings of characters by importance, annotators gave GPT an average rating of 2.70/3 with 70% of its rankings given the highest rating. Only 1% of its rankings were assigned the lowest rating. In cases where GPT made mistakes, it was most often associated with lead bias, where a character appearing early on in the article would be assigned a high importance even if they had little influence later on.

Table 14 gives the inter-annotator agreement which was very strong for character identification and moderate for assessing character ranking. While Krippendorff's alpha was low for character identification in spite of strong overall agreement between annotators (98%), this was because in the rare cases where an annotator thought a character was not important, the other annotators would often disagree.

	R-U	Tech.	Imm.	Clim.	SSM
N	496	1852	4000	4000	4000
μ_{char}	5.6	5.3	4.3	4.4	4.4

Table 15: The total number of articles and the average number of characters per article for the datasets in the case study. R-U and SSM are shorthands for the Russia-Ukraine and Same-Sex Marriage datasets.

B.1 Classifying Main vs. Minor Characters

Two main and two minor characters were randomly selected from each article for labeling by both LLMs and human annotators. This was to assess LLM capabilities across the spectrum of character importance while also maintaining reasonable cognitive load for human annotators (RAs).

Although character importance is a spectrum, we used a binary classification for simplicity, and used the following heuristic to classify characters. If the character was within the top 2 characters in GPT's importance ranking, it was considered a main character, otherwise, it was a minor character. For each article, the first two characters were always selected for labeling as the major characters, as well as 2 randomly selected minor characters. If fewer than 4 characters were identified in an article, all characters were used.

Although this method may be imprecise, our empirical results from AMT showed no substantial differences in performance between these two groupings as seen in Figure 5. In particular, GPT still ranked highest for main and minor characters and all LLMs performed better than humans across both groupings.

C Application

C.1 Application Datasets

As mentioned in section 6, the Russia-Ukraine and Technology datasets were sourced from the Global News dataset (see A.1), and the Immigration, Climate, and SSM datasets were sourced from the Media Frames Corpus (MFC; Card et al., 2015). This data was obtained from the authors under the condition of academic use, which we adhered to. For access to our sample, please contact the original authors.

Table 15 gives the breakdown of the number of articles and characters for the datasets in the Application section. Table 16 and Table 17 give the news outlet breakdowns for these datasets.

News Source	Imm. (%)	Clim. (%)	SSM (%)
Washington Post	23.4	28.4	21.7
New York Times	18.6	25.2	20.2
Atlanta Journal and Constitution	8.8	3.7	5.2
San Jose Mercury News	8.0	10.7	10.2
St. Louis Post-Dispatch	6.6	7.7	7.5
Daily News	6.3	1.3	4.3
Denver Post	5.0	4.8	4.1
Usa Today	5.0	5.5	5.1
Philadelphia Inquirer	4.9	3.7	6.5
Palm Beach Post	4.0	0.2	1.8
Saint Paul Pioneer Press	3.4	3.8	6.5
St. Petersburg Times	3.1	1.9	2.5
Tampa Bay Times	1.9	1.3	1.7
Herald-Sun	1.2	1.8	2.7

Table 16: News source breakdowns for the Immigration, Climate, and SSM datasets used in the Application section. All datasets here were sourced from the Media Frames Corpus (Card et al., 2015).

The article publication dates for the MFC articles were between January 5th, 2000 to February 5th, 2018. As mentioned in A.1, the time frames of Russia-Ukraine and Technology articles were between October 1st to November 29th 2023.

The mean article length was 260 words for the MFC articles and 727 words for the Russia-Ukraine and Technology articles.

C.2 Character Label Clustering

C.2.1 Challenges with Label Embeddings

As mentioned in section 6, clustering the character role labels faced challenges since many embedding models did not consistently capture important aspects to character role equivalence. Table 18 gives some additional examples of this, showing the cosine similarities between label pairs when embedded using different embedding models from Sentence Transformers. Note that this table is intended for illustrative purposes and is not meant to be exhaustive.

The first two rows, "Victim/Victims" and "Establishment/Institution", are intended as baselines

News Source	R-U (%)	Tech. (%)
Business Insider	29.9	7.1
Globalsecurity.org	22.4	3.5
RT	12.5	1.5
ABC News	9.9	3.6
BBC News	6.7	3.6
The Times of India	5.8	17.3
International Business Times	2.9	1.9
NPR	2.3	1.9
Al Jazeera English	1.8	2.5
Time	1.8	1.7
Boing Boing	0.8	0.6
GlobeNewswire	0.8	4.6
ReadWrite	0.8	2.0
Wired	0.6	0.7
Forbes	0.4	16.9
Gizmodo.com	0.3	0.7
The Verge	0.3	0.4
The Punch	0.2	1.5
ETF Daily News	0.0	19.4
The Indian Express	0.0	3.2
Digital Trends	0.0	1.3
Phys.Org	0.0	1.3
Deadline	0.0	1.1
Marketscreener.com	0.0	0.5
Android Central	0.0	0.4
CNA	0.0	0.4
CNN	0.0	0.2
AllAfrica - Top Africa News	0.0	0.1

Table 17: News source breakdowns for the Russia-Ukraine and Technology datasets used in the Application section. Both datasets were sources from the Global News dataset.

as they represent the same or highly similar roles, whereas the remaining six rows correspond to role pairs that should be dissimilar compared to the baselines. The chosen embedding models are the top 2 highest ranked general-purpose[3] and NLI[4] models from Sentence Transformers, as well as GloVe embeddings (Pennington et al., 2014) for additional comparison.

As seen, opposing roles like "Protagonist/Antagonist," "Prosecutor/Defendant," and "Attacker/Victim" often had comparable similarity to "Establishment/Institution." Agent and patient roles like "Accuser/Accused" also had relatively high similarity even those these roles should ideally be narratively distinct in order to differentiate between degrees of agency (i.e., doing vs. receiving an action). In addition, full labels possessing the same primary label but opposing modifiers (e.g. "Cruel/Heroic Leader") or the same modifier with opposing primary labels (e.g. "Diplomatic Instigator/Mediator") also had high similarity.

While separating labels by their HVV classification prior to clustering may help alleviate some of these issues, it is not a robust solution, as some roles like "Prosecutor/Defendant" and "Accuser/Accused" can function as heroes, villains, or victims depending on the context.

C.2.2 Label Clustering Procedure

To avoid inconsistent embedding similarity bias of full labels by their primary labels versus their modifiers, primary labels and modifiers were clustered separately to establish normalized lexicons for each type. Full labels were then normalized using the normalized MOD + PRIM combination.

All primary labels and modifiers were initially singularized and lemmatized using NLTK. Words were embedded using Sentence Transformer embeddings (Reimers and Gurevych, 2019) and clustered using agglomerative clustering with complete linkage. Alternative clustering approaches, such as HDBSCAN (Campello et al., 2013), were also evaluated but produced lower-quality results. All labels from all 5 Application datasets were clustered together, and distance thresholds were chosen by maximizing for Silhouette score (Rousseeuw, 1987). All labels/modifiers within a cluster were normalized to the most frequently occurring label/modifier within the cluster.

We experimented with clustering using all of the embedding models from Table 18. These were chosen since they are the top 2 highest ranked general-purpose and NLI models on Sentence Transformers. Ultimately, the stsb-roberta-base-v2 model was used since it better distinguished antonyms and produced the best qualitative groupings for both primary labels and modifiers. For the primary label clusters, select clusters were manually split to improve cluster quality (e.g. separating "Protagonist" and "Antagonist"). The results are available on our repository. For the full breakdown of numbers of labels before and after clustering, see Table 19.

[3] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

[4] https://www.sbert.net/docs/pretrained-models/nli-models.html

		GloVe	General purpose		NLI	
			all-mpnet-base-v2	multi-qa-mpnet-base-dot-v1	stsb-mpnet-base-v2	stsb-roberta-base-v2
Victim	Victims	0.69	0.71	**0.92**	**0.85**	0.75
Establishment	Institution	0.46	0.74	0.68	0.61	0.54
Protagonist	Antagonist	0.56	0.64	0.61	0.50	**0.91**
Prosecutor	Defendant	0.40	0.75	0.74	0.66	0.58
Attacker	Victim	0.46	0.64	0.71	0.64	0.29
Accuser	Accused	0.21	0.72	0.85	0.63	0.78
Cruel leader	Heroic leader	0.67	0.50	0.60	0.31	0.56
Diplomatic instigator	Diplomatic mediator	**0.71**	**0.84**	0.81	0.69	0.60

Table 18: Cosine similarities of the embeddings between pairs of role labels. Opposing role pairs have comparable similarities to roles pairs like "Victim/Victims" and "Establishment/Institution" which should, intuitively, have higher similarity. Bold indicates the most similar score for each model (column). Models (aside from GloVe) are categorized as general-purpose embedding models or those trained using NLI data.

	R-U	Tech.	Imm.	Clim.	SSM
PRIM					
Orig.	479	1089	1779	1440	1265
Clus.	334	654	847	734	675
%	30.3	39.9	52.4	49.0	46.6
MOD					
Orig.	815	3223	3754	3124	2590
Clus.	631	1730	1870	1576	1400
%	22.6	46.3	50.2	49.5	46.0
Full					
Orig.	1866	6382	10525	9120	8126
Clus.	1778	5823	9062	7739	6792
%	4.7	8.8	13.9	15.1	16.4

Table 19: Breakdown of the number of generated labels by dataset in the case study. "Orig." gives the original number of unique labels generated by GPT-4o, "Clus." is the number of labels after clustering, and "%" is the compression percentage. R-U and SSM are shorthands for the Russia-Ukraine and Same-Sex Marriage datasets.

C.3 HVV classification

We operationalize heroes, villains, and victims based on the definitions from the Narrative Policy Framework (Jones and McBeth, 2010) as well as Bergstrand and Jasper (2018). They are defined as follows:

- **Hero**: a fixer to a problem. They may increase agreement among members, boost commitment for a cause, or serve as a rallying point for a cause.
- **Villain**: a causer of a problem. They may focus blame, provide a clear target for action, intensify negative emotions, or solidify group identities.
- **Victim**: one harmed by a problem. They are innocent, good, and in need of protection.

We used a sample of 10 randomly selected articles from the validation dataset, and one author manually labeled each of the characters into these categories (with an additional "neither" category) which were treated as the gold-standard. The articles contained 63 characters in total with 3 articles from each of the BBC and the Times of India, and 2 articles from ABC and Al Jazeera. Seventeen characters were labeled as "hero," 18 as "villain," 13 as "victim," and 15 as "neither."

Class	Prec.	Recall	F1	*N*
Hero	73.9	100	85.0	17
Villain	94.7	100	97.3	18
Victim	100	92.3	96.0	13
Neither	100	60.0	75.0	15
Avg. (Macro)	92.2	88.1	88.3	

Table 20: Precision, recall, and F1 scores for the HVV classification using GPT-4o ($N_{total} = 63$).

We tried 4 different prompt variants with each of GPT-4o, Claude, and Llama. GPT achieved the best performance with an accuracy of 88%, which was the prompt used in the Application section. Table 20 gives the precision, recall, and F1 scores.

C.4 Additional Results – Heterogeneous Characters

While the Russia-Ukraine conflict represents a standard case for the HVV framing, our method can also lend insight in more ambiguous situations. Figure 6 shows the full labels for Barack Obama within the Immigration dataset, which shows a more heterogeneous mixture of heroic and villainous portrayals. Across those articles, he is portrayed as a hero, villain, and victim 55%, 22%, and 9% of the time, respectively (the remaining 14% was neither). His labels also exhibit more nuance: while he is a villain for being a "Controversial protagonist", he is also a hero as a "Controversial decision maker." The contrast between Obama's labels and those of Zelenskyy, Putin, and Ukraine from Figure 2 also showcases the resolution of our method. Obama and Zelenskyy are both primarily heroes, however Obama's role as an "Immigration reform advocate" shows he is a very different kind of hero than Zelenskyy's "Resolute leader." Similarly Obama's "Political target" role is a very different kind of victimhood than Ukraine's "Resilient defender."

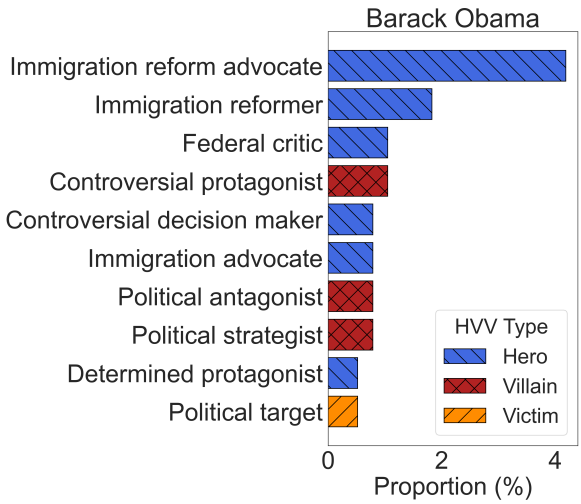


Figure 6: Most frequent full labels for Obama in the Immigration dataset. The proportions represent the percentage of articles where Obama is assigned that label.