

Heroes, Villains, and Victims, and GPT-3

Automated Extraction of Character Roles Without Training Data

Dominik Stammbach
ETH Zurich
dominsta@ethz.ch

Maria Antoniak
Cornell University
maa343@cornell.edu

Elliott Ash
ETH Zurich
ashe@ethz.ch

Abstract

This paper shows how to use large-scale pre-trained language models to extract character roles from narrative texts without domain-specific training data. Queried with a zero-shot question-answering prompt, GPT-3 can identify the *hero*, *villain*, and *victim* in diverse domains: newspaper articles, movie plot summaries, and political speeches.

1 Introduction

What makes a good story? According to some leading theories of narrative (e.g. Propp, 1968), a good story has at least three ingredients: a *hero*, a *villain*, and a *victim*. In a classic recipe, the hero defeats the villain to gain justice for the victim. The cohesiveness of these character roles extends even beyond fiction, with social-science research showing the effectiveness of the hero-villain-victim framework in explaining what motivates consumer behavior and political processes (Jones and McBeth, 2010; Clément et al., 2017; Bergstrand and Jasper, 2018).

Motivated by the relevance of narratives to culture and society, a literature in natural language processing (NLP) has arisen to automatically detect characters in texts and annotate their character roles (Bamman et al., 2013a; Jahan and Finlayson, 2019; Piper et al., 2021). In particular, prior work has used dictionary methods to identify heroes, villains, and victims in texts (Gomez-Zara et al., 2018). The previous methods have mixed results, motivating the present research.

This paper explores the use of large pre-trained language models for the task of character role labeling. Operationalizing the problem as a Machine Reading Comprehension Task (MRCP), we provide an input document and ask the language model who is the hero (or villain or victim). As illustrated by the prompt in Figure 1, we directly ask “Who is the hero” (or villain or victim). Thus, we can extract character roles from plain-text documents without in-domain training data.

Who is the villain in the following text?

Text: In questioning the use of hydraulic fracturing in New York to help produce natural gas, you do not note that the technology has been employed and continuously improved for more than 50 years and that studies by the Environmental Protection Agency and the Ground Water Protection Council have not identified a single instance of groundwater contamination. Wells where fracturing is used are specially constructed to protect drinking water sources. Regulatory oversight is extensive. The fluids mostly water that are forced into a well to create pressure to fracture rock are pushed back out by the oil and gas flowing upward for safe processing. Protecting our water supplies is important, as are reductions in greenhouse gas emissions through use of clean-burning natural gas. Banning hydraulic fracturing would be unwarranted and shortsighted, preventing production of large amounts of natural gas that could directly benefit New York consumers for decades and create thousands of good jobs.

Villain: The villain in this text is the person who is questioning the use of hydraulic fracturing in New York.

Hero: The hero in this text is the person who wrote it, who is trying to convince the reader that hydraulic fracturing is a safe and effective way to produce natural gas

Victim: The New York consumers

Figure 1: Zero-Shot Example of the NFP using GPT-3. Prompt in italics, GPT-3 completion in bold (Analogous results for the hero and victim below the dashed line)

We find that a large pre-trained language model, GPT-3 (Brown et al., 2020), is more effective in labeling these character roles than previous methods, across a diverse set of narrative domains. First, we investigate our method on a labeled corpus of newspaper articles about fracking where the three roles—hero, villain, victim—are manually annotated. In that dataset, our language-model approach is over twice as accurate as an existing baseline using a dictionary approach. Second, as an example of possible memorization during pre-training, we report the annotations produced on a selection of Disney movie plot summaries. These results comport well with subjective judgment.

In our third experiment, we apply the method to a corpus of U.S. State of the Union Addresses, 2001-2018. Matching up the character role annotations with the party affiliation of the president, we explore partisan differences in the framing of heroes, victims, and villains. To make this process feasible, we explore clustering of the GPT-3

output, which produces more legible sets of character assignments. These results demonstrate the promise of the method for empirical research in social science and the digital humanities.

These results are of broad interest given the narrative centrality of character archetypes (Propp, 1968). They are of more specific interest in the literature analyzing narrative framing in news media and policy discourse (Jones and McBeth, 2010; Blair and McCormack, 2016). A robust and efficient computational method to extract character roles in text without training data opens up a wide array of research questions to quantitative analysis.

2 Related Work

This paper adds to the work in NLP on automated extraction of character roles from natural language accounts, and in particular the identification of heroes, villains, and victims. The closest paper is Gomez-Zara et al. (2018), who similarly focus on the detection of heroes, villains, and victims in news articles and provide a dictionary-based approach which we will use as a baseline.

On the broader problem of extracting stereotypical character roles, prior work has explored a variety of methods, including the detection of personas using annotated data combined with feature engineering and regression (Bamman et al., 2013b); parsing and lexical matching tools to identify a consistent set of personas (e.g. doctor, nurse, doula) across testimonials about childbirth and then assess the relative power dynamics (Antoniak et al., 2019); annotations of German news and social media sentences for villains and rogues and transformer models to machine-tag these roles (Klenner et al., 2021); clustering of structural plot information from folktales (Jahan et al., 2021); and a combination of NER and clustered phrase embeddings to identify repeatedly occurring entities, along with semantic role labeling to identify how entities are connected by actions (Ash et al., 2021). Our method does not rely on labeled data, but we employ some of these techniques (e.g., clustering) to support the legibility of our results.

The second related literature is treating role extraction as a machine reading comprehension (MRCP) task, which for example has been proposed for semantic role labeling (He et al., 2015). Most related to our work, Liu et al. (2020) and Du and Cardie (2020) interpret event extraction as an MRCP task and leverage pre-trained language

models to extract events, producing state-of-the-art results in event extraction and leading us to apply this method for detection of character roles.

In principle, any NLP task can be framed as MRCP or question answering (QA) tasks (see e.g. Kumar et al., 2016; McCann et al., 2018). Interpreting tasks (such as event extraction) as question answering enables us to leverage zero-shot capabilities of pre-trained models. Moreover, these methods are not necessarily dependant on domain-specific features, but solely on plain text. Given the zero-shot capabilities of MRCP tasks across domains (Brown et al., 2020), it is more likely that this procedure transfers across domains.

Our proposed task has many similarities with the computational identification of *framing* (Card et al., 2015) and *agenda setting* (Tsur et al., 2015; Field et al., 2018), as well as with automated *bias measurement* (Bolukbasi et al., 2016; Caliskan et al., 2017). These various tasks all seek to identify the author’s written perspective; the same topic can be portrayed differently by different authors, just as the hero or victim might be assigned differently by different authors (Bergstrand and Jasper, 2018). Our identification of the hero, villain, and victim provides yet another method to describe the particular viewpoint expressed in a particular text and to draw comparisons between these various viewpoints over large datasets.

3 Methods

3.1 Labeling Character Roles

Our approach is to frame the labeling of narrative character roles as a machine reading comprehension or closed question answering task. We use auto-regressive language models, i.e., we provide the question and context as prompts to a pre-trained model and decode the answer span token-by-token. We use GPT-3 (Brown et al., 2020), which has proven proficiency in various question-answering tasks (e.g. Rajpurkar et al., 2018). This method allows us to directly leverage knowledge acquired in pre-training on vast amounts of text.

Figure 1 shows an example prompt. We directly ask, “Who is the villain [or hero or victim] in the following text?”. That question is followed by the story text, and then the respective character role is repeated to nudge the model to generate the most likely completion of this prompt. We use the same prompt across all experiments in this study, only varying the story text. We use the 175B-parameter

davinci model with default decoding parameters.¹

To benchmark our new model’s performance, we consider as a baseline the dictionary-based model from Gomez-Zara et al. (2018). First, they use named entity recognition (NER) to extract important entities from news articles. Second, for each entity, they use the surrounding text and its sentiment polarity and dictionary matching to decide whether an entity is a hero, villain, or victim.²

3.2 Corpora

We apply our labeling approach to three corpora, described here. These corpora span three domains and types of narratives: descriptions of current events, fictional stories, and political speeches.

Newspaper Articles. The first domain is newspaper articles. We use a corpus of 66 newspaper articles about fracking published in the *Boulder Daily Camera*, a local Colorado newspaper, from the years 2008-2013. Blair and McCormack (2016) hand-code the three character roles (hero, villain, victim) in these articles.³ The average length of each article is 682 words.

Disney Movie Plots. The second domain is Disney Movie plots. We selected eleven Disney movies based on a “most well-known classics” list (see Table 2 below). We then downloaded the plot summary section for these movies from Wikipedia. The average plot summary length is 670 words.

U.S. Presidential Speeches. Our third corpus includes presidential speeches given at the annual U.S. State of the Union Address, for the years 2001 to 2018.⁴ We split each speech into paragraphs and skip paragraphs containing fewer than 20 words. The final corpus contains $N = 1,379$ paragraphs. Each paragraph contains on average 73 words.

4 Results

This section presents the results, with the three empirical domains reported in turn.

¹except for the state-of-the-union address speeches where we use the 13B-parameter curie model for cost reasons.

²We could not find the original source code for Gomez-Zara et al. (2018), so we used the implementation available at <https://github.com/meganzhao10/Hero-Villain-Victim>.

³Note that the original article also had articles from the *Colorado Springs Gazette*, but we were unable to reproduce that part of the corpus. We contacted the authors to provide the articles, but without success.

⁴These are hosted on Kaggle at [https://www.kaggle.com/datasets/rtatman/](https://www.kaggle.com/datasets/rtatman/state-of-the-union-corpus-1989-2017)

Character	Accuracy GPT-3	Accuracy Baseline	N
Hero	50%	15%	20
Victim	90%	65%	20
Villain	47%	18%	17
All	63%	33%	57

Table 1: Main Evaluation Results: Accuracy of GPT-3 for extracting heroes, villains and victims from The Boulder Daily Camera articles, compared to a dictionary-based baseline described in (Gomez-Zara et al., 2018). In the last column N , we show the number of annotations for each character type present in the data.

4.1 Newspaper Articles about Fracking

Our first analysis applies our GPT-3 method to the collection of news articles about fracking from Blair and McCormack (2016). That paper uses the manual annotations of character roles to analyze framing differences between liberal and conservative media. Regardless of the political leaning, the media outlets in that study framed the public as the victim and the oil and gas industry as the villain. However, the role of hero differed: the liberal media outlet often presented environmental organizations as the hero, while this role is instantiated by specific actors of the oil and gas industry in the conservative outlet.

To automate the annotation process, we use the prompt shown in Figure 1 for each article, the difference being that the text now is the article in question. In the manually annotated data, the authors only find 20 heroes, 20 victims and 17 villains, but our method produces a result for every character role in every article. For evaluation, we only consider model outputs in cases where a true gold annotation exists, and discard all other articles.

In the gold labels from Blair and McCormack (2016), annotations are coarsened such that each specific role (e.g. hero, villain) is mapped to one of a finite set of classes: the public, the government, environmental organizations, or the oil and gas industry. A challenge in the model evaluation is that the language model is not constrained to the finite label set, so the generated text output often does not exactly match the gold labels, even when the output is semantically correct. For the purposes of evaluation, we manually map each GPT-3-generated answer to one of the four categories. The set of GPT-3 outputs and our annotated labels are shown

[state-of-the-union-corpus-1989-2017](https://www.kaggle.com/datasets/rtatman/state-of-the-union-corpus-1989-2017).

in Appendix Table 4.

The evaluation results are shown in Table 1. We achieve an overall accuracy of 63%, a large improvement over the dictionary baseline from Gomez-Zara et al. (2018) (33%). While both methods provide decent results for the victims (which is usually assigned to the public), our approach achieves strong gains in detecting the heroes and villains. We observe an almost three-times improvement for both villains and heroes. More detailed metrics (precision, recall, and F1, per annotation type and character role) are reported in Appendix Table 5. To investigate the stability of our method, we replicated the GPT-3 experiment twice. We achieved 65% and 70% overall accuracy in the replication runs.

As an alternative to GPT-3, we also replicate our results with a pre-trained QA model fine-tuned on the union of 8 existing QA datasets (Khashabi et al., 2020).⁵ Using the same prompts, we achieve an overall accuracy of 55%, which is not far from GPT-3’s performance while using a much smaller model (also having the advantage of being free software). Again, we see the benefits of approaching character role extraction as a QA task leveraging pre-trained models.

4.2 Disney Movie Plot Summaries

Next, we provide qualitative evidence that our method also works in a second domain of popular movie plot summaries. We extract heroes, villains and victims from Wikipedia plot descriptions for widely known Disney movies. Given that these movies contain well-known heroes and villains (if not always victims), it is straight-forward to manually evaluate the quality of the extracted roles. For the same reason, this task also provides some insight into the memorization capabilities of GPT-3, which would have learned about these movies from the training corpus.

The list of annotations for the Disney moves are reported in Table 2. Readers who are familiar with the movies can see that the method works very well in this setting. While some of these annotations are limited or arguable, none are indefensible—there is some reasonable argument for each of these 33 annotations being correct.

As mentioned, these results could be due in part to memorization. We found, for example, that GPT-

⁵We use the unifiedQA-T5-large model found on [huggingface](https://huggingface.co).

3 can correctly complete the prompt “Who’s the hero in Aladdin?” without the additional narrative text. This memorization seems to be important, because the UnifiedQA model (which would not have a memorization capacity) does make more errors (Appendix Table 6). For example, for *Aladdin*, the model mistakes "Aladdin" for the villain.

4.3 U.S. State of the Union Addresses

In our last application, we show how the method can be used to analyze political discourse in the context of U.S. State of the Union Address speeches, where there is no labeled data, as in the fracking articles, or easily verified set of roles, as in the Disney movies. As we have no ground-truth labels, this section follows a descriptive social-science approach and includes adaptations to our previous methods to improve the legibility of the results.

As before, we apply the method to extract a victim, hero, and villain in each paragraph from the corpus of recent U.S. State of the Union Addresses. The free-form texts generated for the character roles are diverse. We have hundreds of unique answers for each role, with many singletons. To reduce the dimensionality of these outputs and make them more interpretable, we encode the phrases using S-BERT (Reimers and Gurevych, 2019) and apply k -means clustering to the resulting vectors (Jahan et al., 2021; Ash et al., 2021). After manual inspection for different k , we select $k=20$.

We then use the partisan affiliation of the speakers to score the most Democrat-associated and most Republican-associated clusters in each character role. Formally, we compute the log odds ratio of each cluster w.r.t. the party affiliation of the president giving the speech and show the cluster with the highest and lowest odds ratio.

Table 3 displays the clusters with the highest partisan log odds ratio by character role—that is, the entities taking on this role more often for one or the other party. For Republican presidents (Bush and Trump), the heroes, victims, and villains in SOTU addresses are connected to the U.S. military and wars in the Middle East. Democratic speeches (by Obama) have a more populist flavor, with the average American portrayed as a hero. Intriguingly, for Democrats the villains and victims are both associated with the education system.

Movie	Hero	Victim	Villain
<i>101 Dalmatians</i>	Roger Dearly	The Dalmatian Puppies	Cruella de Vil
<i>Aladdin</i>	Aladdin	Aladdin	Jafar
<i>Cinderella</i>	Cinderella	Cinderella	Lady Tremaine
<i>Alice In Wonderland</i>	Alice	Alice	The Queen of Hearts
<i>The Jungle Book</i>	Mowgli	Mowgli	Shere Khan, a man-eating Bengal tiger
<i>Sleeping Beauty</i>	Prince Phillip	Aurora	Maleficent
<i>The Lion King</i>	Simba	Mufasa	Scar
<i>Peter Pan</i>	Peter Pan	Wendy, John, Michael, and the Lost Boys	Captain Hook
<i>Mary Poppins</i>	Mary Poppins	Mr. Banks	Mr. Dawes
<i>The Little Mermaid</i>	Ariel	Ariel	Ursula
<i>Snow White</i>	Snow White	Snow White	The Queen

Table 2: Results for Wikipedia plots of widely known Disney Movies

Role	Democrats	Republicans
Hero	The average family watching tonight, the average person, The average American household, The average person, The average worker, Average American. Log Odds Ratio: -0.88	The men and women of the 9/11 generation who have served in Afghanistan and Iraq, The United States military, The military, The veterans, The Cajun Navy volunteers, The man who lost four of his brothers at war, The troops, The troops and civilians who sacrifice every day to protect us, America’s veterans . . . Log Odds Ratio: 1.0
Victim	The American students, The community colleges, The American student, The person who pays for the good education., The school district, A student, The American public school system, The school, The students who are not American citizens, The students, High school students, The high school graduates in Germany, the American student, The teacher, The school system, Every high school diploma is a ticket to success. Log Odds Ratio: -1.43	the coalition to defeat ISIS, ISIS leader, al-Baghdadi, Assad, The UN concluded that Saddam Hussein had biological weapons sufficient to produce over 25,000 liters of anthrax, enough doses to kill several million people, The President of the Iraqi Governing Council, Safia Taleb al-Suhail, Prime Minister Allawi, Iraqi security forces, Iraqi interpreter, Iraqi Government, The Iraqi Government, The American and Iraqi surges have achieved results few of us could have imagined just one year ago. Log Odds Ratio: 1.42
Villain	The college, The teacher who comes in early because he knows she might someday cure a disease., The school administration, The educational system, The school in Dillon, South Carolina, The national competition to improve schools is the villain in this text., The school, The Education Secretary, The education reformer, The school system. Log Odds Ratio: -1.37	The Taliban, Islamic State, ISIS leader, al-Baghdadi, Intel, The terrorists, The unnamed terrorist group, The terrorist underworld, Al Qaida, International terrorism, Iraqi officials, Iraqi intelligence officers, The enemies of freedom, Radical Islam, Marines, Al Qaeda, Sunni extremists, Syrian occupiers, Radical Shia elements. Log Odds Ratio: 1.81

Table 3: Heroes, Victims and Villains extracted from State of the Union speeches. Shown in this table are the entries for the cluster with the highest/lowest odds ratio for Democratic and Republican Presidents

5 Discussion and Future Work

Task formulation. Perhaps the highest-priority limitation of our study is that the method will try to extract a character role from a text, when prompted, even when the role is not present. The newspaper-article evaluation metrics would be much worse if we included the articles missing a role in the test set. In the presidential speeches, in particular, we frequently found that our model assigned the same agent to all three roles—even though villain is mutually exclusive from hero or victim in our evaluation—because there was only one agent mentioned in the speech. There are a number of ways to address this issue. Perhaps the simplest would be to adjust the prompt to allow for a “not applicable” answer, or to ask a preliminary question: “Does this text contain a [role]?” For both of these adjustments, a few-shot approach where the model is provided with some examples could improve performance.

Prompt engineering. Prior work has shown that prompts with subtle differences can product significantly different results (Holtzman et al., 2021; Zhao et al., 2021). Besides few-shot learning, the language-model prompting could also be adjusted to potentially improve performance. Rather than asking about the three roles in three separate prompts, the model could be asked to identify all three simultaneously, for example. The question could be asked in different ways and then the answers aggregated. One could also explore adapting the prompt to constrain the set of entities to a finite set; e.g., in the fracking articles corpus, there was a pre-specified list of four possible entities. Finally, GPT-3 has some decoding hyperparameters that could be tweaked.

Limitations of large language models. Like other NLP models (Bolukbasi et al., 2016), large pre-trained language models can encode harmful human biases (Bender et al., 2021). For example, prior work has shown that narratives generated by

GPT-3 explicitly portray feminine characters as less powerful (Lucy and Bamman, 2021) while also encoding implicit gender biases (Huang et al., 2021). GPT-3 is trained on multiple large datasets, including scraped web text, book texts, and Wikipedia articles. Because of their size, these datasets are difficult to document; even quantifying the number of duplicate documents can be a challenge (Lee et al., 2021) and even more difficult are detailed descriptions, like those called for in data documentation best practices (Gebru et al., 2021). We use GPT-3 to measure authors’ framing biases, but it is simultaneously likely that biases encoded in GPT-3 influence our results in ways that are difficult to measure.

Another major limitation to the use of the GPT-3 API is the cost of OpenAI API queries. The queries for our relatively small-scale analysis of state-of-the-union address speeches cost nine dollars using the 13B-parameter GPT-3 model. Scaling this up to larger corpora of thousands or millions of documents, such as the *U.S. Congressional Record* (Ash et al., 2021), would be prohibitively expensive. Hence, using even larger pre-trained models, such as PaLM (Chowdhery et al., 2022), is likely not cost-effective for most academic research. Exploring smaller open-source language models, such as GPT-Neo Black et al. (2021), which can be implemented at scale, is a promising alternative.

Moving beyond pre-trained language models, performance and scalability could be improved through further model training. Fine-tuning GPT-3 for this task is one possibility. A less expensive option would be to use GPT-3 to create a labeled dataset, perhaps with human supervision, for training a smaller student model. That student model could be a distilled autoregressive model or an encoder model like BERT. For MRCP, BERT-like approaches work well for question-answering tasks where the answer is a span of tokens in the input text—in our case, the character being assigned a character role. A limitation of this approach is that character roles are often implicit, rather than explicitly mentioned in the text. BERT-like models work for explicit mentions, but for implicit mentions a generative model like GPT is needed.

Concluding note. In this work, we used the zero-shot capabilities of a large-scale language model to automatically extract heroes, villains and victims from newspaper articles, movie plot summaries, and U.S. presidential speeches. Large pre-

trained language models can solve machine reading comprehension tasks for the purposes of labeling short to medium-sized documents, without hand-annotated training data. This approach could be useful for many projects in computational social science and digital humanities.

While promising, our results are still a proof of concept. We have introduced a basic version of the method, which performs better than prior work using a dictionary baseline. But our method’s sufficiency for social-science applications is not yet assured, and more work is needed to build up the method and assess its robustness in the field.

References

- Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.
- Elliott Ash, Germain Gauthier, and Philine Widmer. 2021. [Text semantics capture political and economic narratives](#).
- David Bamman, Brendan O’Connor, and Noah A. Smith. 2013a. [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, Sofia, Bulgaria. Association for Computational Linguistics.
- David Bamman, Brendan O’Connor, and Noah A. Smith. 2013b. [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, Sofia, Bulgaria. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- 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.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow](#). If you use this software, please cite it using these metadata.
- Benjamin D Blair and Larkin McCormack. 2016. [Applying the narrative policy framework to the issues surrounding hydraulic fracturing within the](#)

- news media: A research note. *Research & Politics*, 3(1):2053168016628334.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to home-maker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Dallas Card, Amber E. Boydston, 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.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Maéva Clément, Thomas Lindemann, and Eric Sangar. 2017. The “hero-protector narrative”: Manufacturing emotional consent for the use of force. *Political psychology*, 38(6):991–1008.
- Xinya Du and Claire Cardie. 2020. [Event extraction by answering \(almost\) natural questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.
- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. [Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- 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 *23rd International Conference on Intelligent User Interfaces, IUI '18*, page 311–315, New York, NY, USA. Association for Computing Machinery.
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. [Question-answer driven semantic role labeling: Using natural language to annotate natural language](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. [Surface form competition: Why the highest probability answer isn’t always right](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. [Uncovering implicit gender bias in narratives through commonsense inference](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Michael D. Jones and Mark K. McBeth. 2010. [A narrative policy framework: Clear enough to be wrong?](#) *Policy Studies Journal*, 38(2):329–353.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- 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.
- Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Roman Paulus, and Richard Socher. 2016. [Ask me anything: Dynamic memory networks for natural language processing](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387, New York, New York, USA. PMLR.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. [Deduplicating training data makes language models better](#). *arXiv preprint arXiv:2107.06499*.
- Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. [Event extraction as machine reading comprehension](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.
- Li Lucy and David Bamman. 2021. [Gender and representation bias in GPT-3 generated stories](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. [The natural language decathlon: Multitask learning as question answering](#). *arXiv preprint arXiv:1806.08730*.
- Andrew Piper, Richard Jean So, and David Bamman. 2021. [Narrative theory for computational narrative understanding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Vladimir Iakovlevich Propp. 1968. *Morphology of the folktale*. University of Texas Press.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Oren Tsur, Dan Calacci, and David Lazer. 2015. [A frame of mind: Using statistical models for detection of framing and agenda setting campaigns](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Article title	Role	Annotated	GPT-3 answer	Equivalent
No. 3: Gloves come off in the fight over fracking	villain	Oil and gas industry	the oil and gas industry	1
Polis to oil and gas: please stop suing the communities I represent	hero	Fed Govt	Jared Polis	1
Anti-fracking measures win in Lafayette, Boulder, Fort Collins	victim	General or specific public	Lafayette	1
Anti-fracking measures win in Lafayette, Boulder, Fort Collins	villain	Oil and gas industry	the oil and gas industry	1
Colorado Oil and Gas Association gives \$600K to fight fracking bans on Front Range Ballots	hero	Oil and gas industry	Merrily Mazza	0
Colorado Oil and Gas Association gives \$600K to fight fracking bans on Front Range Ballots	victim	General or specific public	Lafayette Campaign for Energy Choice	1
Colorado Oil and Gas Association gives \$600K to fight fracking bans on Front Range Ballots	villain	Environmental Orgs	The Colorado Oil and Gas Association	0
Lafayette to draft resolutions of opposition to fracking ban, utility occupation tax	victim	General or specific public	Lafayette	1
Broomfield approves agreement with oil and gas operator	hero	Environmental Orgs	The City Council	0
Broomfield approves agreement with oil and gas operator	victim	General or specific public	Anthem residents	1
Broomfield approves agreement with oil and gas operator	villain	Local Govt	Sovereign	0
Ballot initiative to ban fracking in Lafayette garners enough signatures to move forward	victim	General or specific public	Lafayette	1
Extension of Boulder County fracking moratorium deemed a win for many	victim	General or specific public	the citizens of Boulder County	1
Lafayette anti-fracking effort draws canvassers	hero	Environmental Orgs	Cliff Willmeng	1
Lafayette anti-fracking effort draws canvassers	victim	Oil and gas industry	Lafayette	0
Lafayette anti-fracking effort draws canvassers	villain	Environmental Orgs	East Boulder County United	1
Broomfield unlikely to pursue fracking moratorium	hero	Environmental Orgs	The City Council	0
Broomfield unlikely to pursue fracking moratorium	victim	General or specific public	the community in Broomfield	1
Broomfield unlikely to pursue fracking moratorium	villain	Oil and gas industry	The City Council	0
Boulder County Commissioner Elise Jhones to debate Gov John Hickenlooper on fracking	hero	Local Govt	Boulder County Commissioner Elise Jones	1
Boulder County Commissioner Elise Jhones to debate Gov John Hickenlooper on fracking	victim	General or specific public	Boulder County Commissioner Elise Jones	0
Boulder County Commissioner Elise Jhones to debate Gov John Hickenlooper on fracking	villain	State Govt	Elise Jones	1
Anti-fracking group to present Lafayette with petition for ban	hero	Environmental Orgs	Anti-fracking group	1
Anti-fracking group to present Lafayette with petition for ban	victim	General or specific public	Lafayette	1
Anti-fracking group to present Lafayette with petition for ban	villain	Oil and gas industry	East Boulder County United	0
Lafayette to pursue moratorium on new oil and gas permits	hero	Oil and gas industry	Lafayette	0
Group urges ban on fracking Activists ask Lafayette for moratorium	hero	Environmental Orgs	John Chinnery	0
Group urges ban on fracking Activists ask Lafayette for moratorium	victim	General or specific public	the community of Lafayette	1
Group urges ban on fracking Activists ask Lafayette for moratorium	villain	Environmental Orgs	John Chinnery	0
In Erie, oil and gas companies to pay twice as much for water	hero	Environmental Orgs	The town last week doubled its commercial water rate . . .	0
In Erie, oil and gas companies to pay twice as much for water	victim	General or specific public	The town of Erie	1
In Erie, oil and gas companies to pay twice as much for water	villain	Oil and gas industry	The oil and gas companies	1
Anti-fracking group to provide air-monitoring equipment to erie residents	hero	General or specific public	The anti-fracking group Erie Rising.	0
Anti-fracking group to provide air-monitoring equipment to erie residents	hero	Environmental Orgs	The anti-fracking group Erie Rising.	1
Anti-fracking groups to protest Encanas drilling operation near Erie schools Saturday	hero	Environmental Orgs	Anti-fracking groups	1
Anti-fracking groups to protest Encanas drilling operation near Erie schools Saturday	victim	General or specific public	Erie schools	1
Anti-fracking groups to protest Encanas drilling operation near Erie schools Saturday	villain	Oil and gas industry	Encana	1
Erie mothers, children descend on Encana headquarters with anti-fracking petition	victim	General or specific public	Erie mothers and children	1
Dirty air in Erie linked to gas drilling	hero	Environmental Orgs	The researcher with the National Oceanic and Atmospheric Administration	0
Dirty air in Erie linked to gas drilling	victim	General or specific public	Erie residents	1
Dirty air in Erie linked to gas drilling	villain	Oil and gas industry	Industry	1
Lawsuit from COGA brings Lafayette into thick of fight over oil and gas drilling	hero	Fed Govt	Lafayette	0
Pro-fracking group alleges 'systemic failures' in Broomfield election	hero	Environmental Orgs	BJ Nikkel	1
Colorado Oil and Gas Association gives \$600K to fight fracking bans on Front Range ballots	victim	General or specific public	Lafayette campaign for energy choice	1
Voter frustration fills ballot with citizen initiatives in Boulder, Broomfield counties	hero	General or specific public	A frustrated voter	1
Voter frustration fills ballot with citizen initiatives in Boulder, Broomfield counties	villain	Other	-	0
Broomfield tightening requirements for oil and gas companies	villain	Oil and gas industry	The City Council	0
Coming off win at Boulder County level, anti-frackers turn focus statewide	hero	Environmental Orgs	Anti-frackers	1
In talk at CU-Boulder, Hickenlooper says he is 'constantly attacked now for being in the pocket of oil and gas'	hero	Oil and gas industry	Governor John Hickenlooper	0
Broomfield postpones hearing on North Park fracking application	hero	Environmental Orgs	Jackie Houle	1
Broomfield postpones hearing on North Park fracking application	victim	General or specific public	Concerned residents of Broomfield	1
'Bucket Brigade': Anti-fracking citizen effort to monitor the air in Erie	victim	General or specific public	Erie residents	1
'Bucket Brigade': Anti-fracking citizen effort to monitor the air in Erie	villain	Oil and gas industry	Global Community Monitor	0
Hundreds gather to protest Encana Corp.'s fracking operation in Erie	victim	General or specific public	The community	1
Hundreds gather to protest Encana Corp.'s fracking operation in Erie	villain	Oil and gas industry	Encana Corp.	1
Erie eyes agreements with oil and gas operators	villain	Oil and gas industry	Erie	0
Fracking discussion packs Erie Town Hall, no action taken on moratorium	victim	General or specific public	The community of Erie	1

Table 4: Article title, annotated label from (Blair and McCormack, 2016), the GPT-3 output, and the author's determination whether the generated output is equivalent to the manual annotation.

Role	Entity	Pr	Rc	F1	N
Hero	Envirommental Orgs	0.88	0.58	0.70	12
	(local, State or Fed) Government	0.50	1.00	0.67	3
	Oil and gas industry	0.00	0.00	0.00	3
	Other	0.00	0.00	0.0	0
	General or specific public	0.20	0.50	0.29	2
Villain	Envirommental Orgs	0.50	0.33	0.40	3
	(local, State or Fed) Government	0.33	0.50	0.40	2
	Oil and gas industry	0.86	0.55	0.67	11
	Other	0.50	1.00	0.67	1
	General or specific public	0.00	0.00	0.00	0
Victim	Envirommental Orgs	0.00	0.00	0.00	0
	(local, State or Fed) Government	0.00	0.00	0.00	0
	Oil and gas industry	0.00	0.00	0.00	1
	General or specific public	0.94	0.89	0.92	19

Table 5: Detailed precision, recall and F1 scores for the different annotation types and roles. N denotes the number of annotated examples in the data, e.g. the data contains 12 Environmental Orgs as heroes.

Movie	Hero	Victim	Villain
<i>101 Dalmations</i>	Roger Dearly	Anita	Cruella de Vil
<i>Aladdin</i>	Aladdin	Aladdin	Aladdin
<i>Cinderella</i>	the hero is Cinderella.	Lucifer	Lucifer
<i>Alice In Wonderland</i>	Alice	the Queen of Hearts	the Queen of Hearts
<i>The Jungle Book</i>	Mowgli	Shere Khan	Shere Khan
<i>Sleeping Beauty</i>	Phillip	Prince Phillip	Maleficent
<i>Lion King</i>	Simba	Scar	Scar
<i>Peter Pan</i>	Peter Pan	Peter Pan	Hook
<i>Mary Poppins</i>	Mary Poppins	the bank	banker
<i>The Little Mermaid</i>	Ariel	Ariel	Ursula
<i>Snow White</i>	the dwarfs	the queen	the queen

Table 6: Results for Wikipedia plots of widely known Disney Movies using the unifiedqa-t5-large model, a T5 model fine-tuned on 8 existing QA datasets.