The Empirical Variability of Narrative Perceptions of Social Media Texts

Joel Mire[♣] **Maria Antoniak**[♢] **Elliott Ash**[♠] **Andrew Piper**[♡] **Maarten Sap**[♣△]

[♣]Carnegie Mellon University [♢]University of Copenhagen [♠]ETH Zürich
[♡]McGill University [△]Allen Institute for AI

Abstract

Most NLP work on narrative detection has focused on prescriptive definitions of stories crafted by researchers, leaving open the questions: how do crowd workers perceive texts to be a story, and why? We investigate this by building STORYPERCEPTIONS, a dataset of 2,496 perceptions of storytelling in 502 social media texts from 255 crowd workers, including categorical labels along with free-text storytelling rationales, authorial intent, and more. We construct a fine-grained bottom-up taxonomy of crowd workers' varied and nuanced perceptions of storytelling by open-coding their free-text rationales. Through comparative analyses at the label and code level, we illuminate patterns of disagreement *among* crowd workers and *across* other annotation contexts, including prescriptive labeling from researchers and LLM-based predictions. Notably, plot complexity, references to generalized or abstract actions, and holistic aesthetic judgments (such as a sense of cohesion) are especially important in disagreements. Our empirical findings broaden understanding of the types, relative importance, and contentiousness of features relevant to narrative detection, highlighting opportunities for future work on reader-contextualized models of narrative reception.

1 Introduction

Identifying stories in social media texts provides a lens through which we can study how individuals and communities process and communicate experiences (Dirkson et al., 2019; Ganti et al., 2022; Falk and Lapesa, 2024). However, despite narrative's omnipresence in our private (Bruner, 1991) and public lives (Shiller, 2019; Dillon and Craig, 2019), its generality and multi-faceted complexity makes modeling and detecting it a major challenge in NLP (Piper et al., 2021; Piper and Bagga, 2022; Antoniak et al., 2024).

Thus far, most approaches to narrative detection in NLP have involved a small number of researchers and/or trained students who use a prescriptive annotation guideline concerned with textual features to obtain a gold label. In this way, prior efforts exhibit relative uniformity in terms of annotator *types* (researchers, sometimes alongside trained students), *number of annotators* (single to few), annotation *paradigm* (prescriptive), narrative *feature type* (textual), and *label aggregation* (single gold).[1] These prescriptive approaches can lead to annotations (and, by extension, trained models) that reflect singular definitions of storytelling, failing to reflect the true variability of narrative perceptions.

There are alternative approaches worth considering across these dimensions, such as crowd workers' perceptions, *descriptive* (codebook-free) annotations (Rottger et al., 2022), extra-textual features concerned with reader response, and different ways of handling human label variation (Gordon et al., 2021; Plank, 2022). Exploring these alternatives

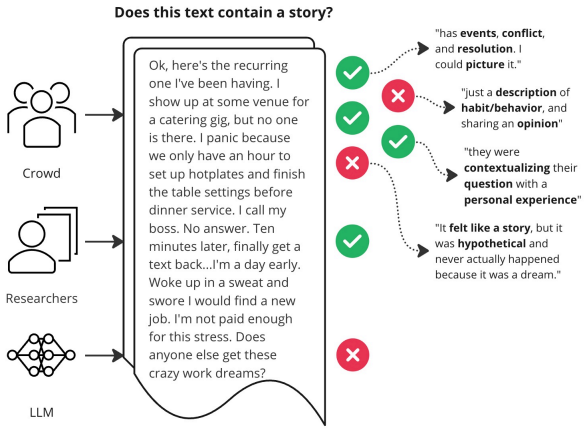

Figure 1: We investigate descriptive perceptions of storytelling from crowd workers, presenting a new dataset, STORYPERCEPTIONS, and compare the crowd annotations to prescriptive labels from researchers and LLM-assisted annotations to explore the complexities and generalizability of narrative detection.

[1]Notable exceptions, which differ across one of these dimensions, are discussed in Section 2.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19940–19968
November 12-16, 2024

would help determine the extent to which narrative perceptions are invariant across substantially different annotation contexts, which affects the generalizability of LLMs and other models trained on prescriptive labels or prompted with prescriptive codebooks using in-context learning. Moreover, insights into which features shape or divide crowd perceptions of narrative can guide model development concerned with the psychological, pragmatic, and social aspects of narrative reception.

To these ends, we investigate *descriptive perceptions* of English-language stories from crowd workers, broadening beyond prescriptive labeling from researchers. We present STORYPERCEPTIONS, a dataset of 2,496 free-text perceptions of storytelling from 255 crowd workers, based on the 502 social media texts in the StorySeeker dataset (Antoniak et al., 2024). Through open coding of the crowd workers' perceptions, we develop a detailed taxonomy of 30 codes representative of the discourse categories (e.g., explanation), textual features (e.g., events), and extra-textual features (e.g., suspense) that crowd workers associate with narratives. Using the taxonomy, we augment the dataset with meta-annotations of both the crowd workers' label `rationales` and perceptions of authorial `goal`.

With STORYPERCEPTIONS, we explore the following questions. First, to bring a broad and fresh set of perspectives to the narrative detection task, we ask **[RQ1] What are crowd workers' descriptive perceptions of storytelling in social media texts?** Second, to illuminate the degree and nature of variation in descriptive perceptions, we ask **[RQ2] How do narrative perceptions differ *among* crowd workers?** Third, and finally, to understand how invariant perceptions are across annotation contexts, we ask **[RQ3] How do narrative perceptions differ *across* prescriptive labels from researchers, descriptive annotations from crowd workers, and predictions from LLM-based classifiers?** Pairwise comparison of descriptive crowd annotations with Antoniak et al.'s (2024) prescriptive labels provides insight into the generalizability of prescriptive approaches. Our motivation for comparing human annotations with LLMs is that deep learning models have been argued to capture "average" perceptions of concepts (Richardson, 2021) or otherwise fail to accurately represent the distributions of human opinions for many tasks (Pavlovic and Poesio, 2024). However, the extent to which this is true for subjective concepts like stories is not fully understood. We evaluate, for the first time, how well descriptive story annotations from LLMs align with human labels from different contexts (crowd workers' descriptive labels and researchers' prescriptive labels). While we do not consider any set of human annotations as the gold standard, understanding the alignment between LLM and crowd worker annotations is crucial for future computational social science research on narrative perceptions in social media. Our comparisons illuminate the validity and risks of using LLM-assisted annotations in this domain.

We find that while crowd workers' label `rationales` frequently refer to a few core textual features (events, characters, plot, setting), they also appeal to cognitive and aesthetic responses (sense of conflict, cohesion, feeling like a story, etc), often alongside references to textual features. Regarding variation among crowd workers, we find that disagreements are more likely to revolve around holistic assessments (e.g., about plot structure) rather than more straightforward textual features. Furthermore, comparing crowd labels to the prescriptive labels from researchers or LLM-based predictions shows that crowd workers have the highest requirements concerning sequential events and plot structure across annotator types. Finally, we find that models in the GPT-4 family (OpenAI et al., 2023) are typically less likely to identify stories in texts that describe abstract activities (e.g., habits, behaviors, processes) relative to human annotators.

Altogether, this study offers new insights from crowd workers on the narrative detection task, underscores the intricate nature of narrative discourse through its diverse features and varying levels of inter-annotator agreement, and identifies both opportunities and important questions for future work at the intersection of narrative modeling, reader response, and LLM-assisted annotation.

2 Background & Related Work

2.1 Annotation Paradigms for Narrative Detection

Significant prior work in NLP has shown that annotator disagreement for subjective tasks is both common and often justified (Aroyo and Welty, 2015; Basile et al., 2021). This has also led to efforts in dataset construction and modeling that resist collapsing variation among perspectives onto one dominant interpretation or label (Uma et al., 2021). Attending to disagreement can lead to an expanded understanding of the task itself, such as the role

of identity and psychological attitudes in labeling (Sap et al., 2022b; Homan et al., 2024).

Rottger et al. (2022) distinguishes between two paradigms for subjective annotation tasks, including *prescriptive* approaches which aim to minimize annotator subjectivity and *descriptive* paradigms which embrace it. As Table 3 in Appendix A shows, the predominant approach to annotation for narrative detection datasets has been overwhelmingly *prescriptive*, that is, the goal has been to follow guidelines in pursuit of a single gold label.

Moreover, the annotator pools for narrative detection datasets have been very small and limited to researchers (Ceran et al., 2012; Dirkson et al., 2019; Antoniak et al., 2024; Abdessamed et al., 2024), trained students (Gordon and Ganesan, 2005; Yao and Huang, 2018; Piper and Bagga, 2022), or other domain experts (Dirkson et al., 2019; Ganti et al., 2022). dos Santos et al.'s (2017) dataset of Portuguese blogs annotated for narrative status is one exception, as it is based on annotations from a large number of crowd workers (167); however, the crowd workers still follow a prescriptive approach.

Existing approaches thus leave open the question of how well the predominant conceptions of narrative associated with prescriptive definitions from researchers map onto crowd workers' perceptions.

2.2 Features for Narrative Detection

Narrative features can be categorized into two types. The first consists of textual features and includes syntactic, semantic, and other structural aspects of texts (Barthes and Duisit, 1975). The second type concerns the cognitive and aesthetic effects that the text has on a reader (e.g., suspense) (Brewer and Lichtenstein, 1982). Bortolussi and Dixon (2002) frame this division as foundational for research in narrative understanding, proposing that the field embrace an experimental framework based on a unidirectional causal model in which *objective* textual features influence *constructed* reader responses. Pianzola and Passalacqua (2016) connects this division to a philosophical distinction between objectivist and constructivist approaches and, advocating for the constructivist view, complicates the notion that textual features precede reader constructions.

Most prior annotation guidelines in NLP have focused on what are traditionally considered textual features—especially characters and event sequences—as summarized in Table 3 in Appendix A. Modeling efforts have historically leveraged lexical (n-grams, lexica), syntactic (parts of speech, named entities), semantic (subject-verb-object triplets), and other structural (event chains) textual features. Several recent approaches have relied on LLM-based models, such as BERT-style models and GPTs with in-context learning (Antoniak et al., 2024; Abdessamed et al., 2024). Notable exceptions include Piper and Bagga's (2022) emphasis on world-making, which relates to constructed features of concretization and experientiality, and Steg et al.'s (2022) attempt to model narrativity via the constructed features Sternberg (2001) defines as suspense, curiosity, and surprise.

StoryPerceptions illuminates how crowd workers perceive stories, from textual features to extra-textual elements linked to readers' cognitive and aesthetic reactions. Understanding the role of extra-textual, constructed reader responses in narrative perceptions is crucial for future research on the pragmatic and social dynamics of stories.

2.3 Discursive Roles of Storytelling

Considering the flexibility and prevalence of narrative, significant studies have explored its discursive roles. These include analyses of how stories are composed of discourse forms, such as description (Bal, 1997), and how storytelling contributes to broader discursive aims. Philosophers have examined narrative's role in self-understanding (Pellauer and Dauenhauer, 2022; Pereira Rodrigues, 2023), while psychologists view narrative as a foundational, automatic mechanism for organizing sense experiences (Bruner, 1991). Sociologists and humanists have investigated storytelling's role in collective sense-making and democratic processes (Polletta and Lee, 2006; Bietti et al., 2019; Dillon and Craig, 2019), and others have highlighted the explanatory power of narrative discourse across specialized domains, from historiography (White, 1980) to scientific communication (Dahlstrom, 2014).

Additionally, there has been much attention to the relationship between storytelling and persuasive forms of writing. For example, psychologists have found that narrative transportation (Green and Brock, 2000) and emotional flow (Nabi and Green, 2015) contribute to persuasive outcomes.

In NLP, researchers have mostly studied social storytelling in connection to sensemaking processes (Verberne et al., 2018; Antoniak et al., 2019) and argumentation (Falk and Lapesa, 2023, 2024). While these works often focus deeply on a particular discursive mechanism of storytelling or a spe-

cific community, our work offers a broad view of the discursive roles of storytelling across topically diverse social media texts.

3 The STORYPERCEPTIONS Dataset

Our texts are drawn from the StorySeeker dataset (Antoniak et al., 2024), which contains posts and comments sampled from Reddit communities (Völske et al., 2017).[2] The dataset includes researcher-annotated binary labels indicating whether a text contains a story, based on a prescriptive codebook that emphasizes characters and event sequences ("A story describes a sequence of events involving one or more people"). We refer the reader to Antoniak et al. (2024) for a complete discussion of their codebook.

Building STORYPERCEPTIONS on the StorySeeker dataset has several advantages: the texts originate from diverse online communities, providing a broad range of features relevant to narrative perception; they are nearly evenly split between stories and non-stories; and the prescriptive labels, grounded in the researchers' codebook, serve as a useful comparison point with the descriptive labels and rationales collected from crowd workers.

3.1 Collecting Crowd Perceptions

We design an annotation task to illuminate how people interpret texts as containing or not containing stories. We serve our task via the Portable Text Annotation Tool (Potato) (Pei et al., 2022) and recruit US-based participants with an undergraduate degree via Prolific.[3]

The task presents a text from the StorySeeker dataset and asks about the following aspects:

1. story `label` (binary)
2. label `rationale` (free-text)
3. label `confidence` (Likert)
4. story `span`, if text contains a story (free-text)
5. perceived `goal` of author (free-text)
6. `alternative` classification, if text does not contain a story (free-text)
7. text topic `familiarity` (Likert)

Our study was considered exempt by the IRB at the Allen Institute for AI. See Appendix B for the survey details, e.g., questions, recruitment filters, and demographics.

[2]Each text is between 100 and 500 tokens, consists of coherent sentences, and was accompanied by a short summary.

[3]www.prolific.com

Our final STORYPERCEPTIONS dataset consists of 2,496 survey responses from 255 crowd workers, with 5 responses from different crowd workers for each of the 502 StorySeeker posts. We share our dataset and code publicly.[4]

3.2 Open Coding

We analyze the crowd workers' nuanced and multifaceted free-text responses using open coding (Saldaña, 2013), fully detailed in Appendix C.

Through extensive rounds of reviewing and refining codes among coauthors, we observed significant conceptual overlap across crowd workers' responses to the three free-text questions (perceived `goal`, story label `rationale`, and `alternative` if no story was identified). We therefore opt to develop a unified taxonomy of 30 codes, consisting of 20 *feature* codes and 10 *discourse* codes. The *discourse* codes describe broad categories of writing modes, such as explanation, argument, and inquiry. The *feature* codes include both relatively textual features (e.g., characters, events) and extra-textual features associated with reader response, such as descriptions of reading experiences and aesthetic judgments (e.g., "evocative," "cohesive"). Each code can have a positive or negative polarity, indicating either its stated presence or absence.[5]

Using the final taxonomy, the first author annotated all 2,496 responses, each composed of 3 distinct free-text sub-responses. A coauthor independently annotated 25 responses using the final taxonomy for validation purposes. We measure the agreement among annotators using the Jaccard index, defined as the size of the intersection of code annotations divided by the size of their union. The Jaccard indices for the questions about the (1) perceived `goal` of the author, (2) label `rationale`, and (3) `alternative` classification (if not a story) were 0.515, 0.613, and 0.75, respectively. Considering the large number of codes in the taxonomy,[6] the moderate agreement scores suggest that the taxonomy provides an interpretable scaffold for annotations, with clear distinctions among the codes.

For illustration, we paraphrase the `rationale` a crowd worker provided to justify their claim that a comment (concerning strategies for navigating the

[4]https://github.com/joel-mire/story-perceptions

[5]Having both positive and negative versions of each code enables, for instance, distinguishing "There **weren't** any **characters**" from "There **were characters**."

[6]60, when considering both positive and negative variants of each code

medical profession) contains a story:

> i think the part where the writer shares a scenario of someone who has chosen the wrong specialty counts as a story, just the scenario part, because i imagined the nephrologist getting out of bed, going to work, and doing these tasks in some vague way. Does that make a story? I'm not sure but it feels like one

Using our taxonomy, we assigned the meta-annotations to the `rationale`: CHARACTER_PERSON, PLOT_SEQUENCE, EVOCATIVE_TRANSPORTING, EVENT_EXPERIENCE, FICTIONAL_HYPOTHETICAL, THEME_MORAL.

See Appendix D for additional information about the taxonomy, the complete list of feature (Table 6) and discourse (Table 5) codes, annotation rules (D.3), and complete example annotations (D.4).

4 Examining Crowd Perceptions [RQ1]

To understand the crowd workers' perceptions of storytelling, we first analyze the feature and discourse codes associated with stories and examine their co-occurrences.

4.1 Relative Feature Prevalence in Stories (vs. Non-Stories)

To identify which features crowd workers associate with storytelling (as opposed to non-storytelling texts), we examine the relative prevalence of feature codes in label `rationales` between these two groups.[7] Specifically, we measure the difference in the empirical probabilities that a feature code is referenced in a `rationale` for a story label versus a non-story label.

We use this and other variants of the relative prevalence metric throughout the paper because we are primarily interested in the relationship between codes and distinct rationale groups (e.g., stories vs. non-stories), independent of other codes. Beyond merely showing that a code is prevalent in stories in an absolute sense, the relative prevalence metrics bring greater attention to the tail of the distribution of codes, which supports our goal of broadening understanding of the types, relative importance, and contentiousness of features relevant to narrative perception.

As shown in Fig. 2, a few textual features like EVENT_EXPERIENCE, CHARACTER_PERSON, and PLOT_SEQUENCE have the highest relative prevalence in story `rationales`. These features are

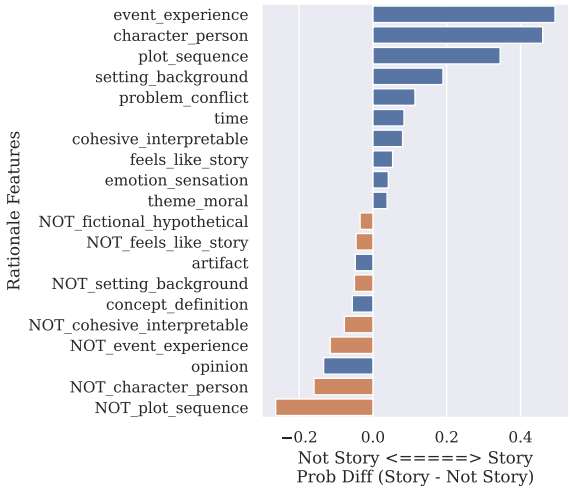

Figure 2: The relative prevalence of feature codes in story (vs. non-story) `rationales`. Positive values indicate greater prevalence in stories, and negative polarity codes are shown in orange. Features mentioned fewer than 50 times or with a probability difference magnitude less than 0.03 are excluded from this and subsequent plots (unless stated otherwise).

largely consistent with prior work on narrative detection (Piper and Bagga, 2022; Antoniak et al., 2024), suggesting a shared understanding of the core components of storytelling.

Moreover, we observe that numerous other features are prevalent in `rationales`, many of which are *constructed* extra-textual features. Examples include PROBLEM_CONFLICT, COHESIVE_INTERPRETABLE, and FEELS_LIKE_STORY, all of which point to the cognitive and aesthetic experiences or judgments. While these kinds of reader-constructed features correspond to prior work in narrative theory (Herman, 2009), psychology (Green and Brock, 2000; Graesser et al., 1994), and some recent efforts in NLP (Steg et al., 2022), our work shows the importance of a broad set of these features to crowd workers, foregrounding their relevance for future work in computational narrative understanding.

4.2 Feature Co-occurrence in Stories

Since multiple feature codes applied to the label `rationales`, we also examine how feature codes co-occur using normalized pointwise mutual information (Church and Hanks, 1990; Bouma, 2009).

Among expected pairings of core features (or their shared absence), such as the pairing of NOT_CHARACTER_PERSON and NOT_PLOT_SEQUENCE, which has the highest

[7] See Appendix F.1 for the independent feature prevalence metrics.

Most Co-occurring Feature Pairs (Constrained)	
Feature Pair	NPMI
cohesive_interpretable & plot_sequence	0.4
NOT_plot_sequence & NOT_problem_conflict	0.32
NOT_cohesive_interpretable & NOT_plot_sequence	0.27
plot_sequence & problem_conflict	0.23
plot_sequence & theme_moral	0.23
character_person & problem_conflict	0.23
feels_like_story & plot_sequence	0.21
problem_conflict & setting_background	0.18
character_person & emotion_sensation	0.17
cohesive_interpretable & event_experience	0.17

Table 1: The most co-occuring pairs of features that consist of both a textual feature and an extra-textual feature. Scores can range from -1 (never co-occur) to 0 (independent) to 1 (always co-occur). We filter out pairs that occur less than 20 times.

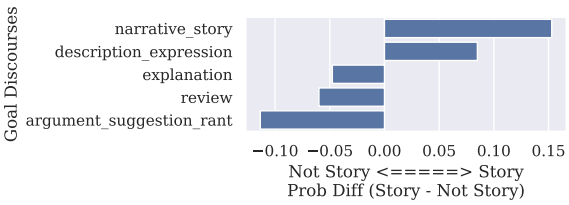

Figure 3: The relative association of discourse categories with the perceived `goal` of the text.

overall NPMI score (0.5), we find many interesting connections between relatively textual features and relatively *constructed* extra-textual features that rely on cognitive or aesthetic judgments. Table 1 lists the 10 most co-occuring pairs composed of one textual feature and one extra-textual feature. The pairing of PLOT_SEQUENCE and COHESIVE_INTERPRETABLE, for example, suggests that the extent to which a story comes together as a meaningful whole (e.g., via resolution) depends on the presence of a plot sequence. In general, PLOT_SEQUENCE is associated with various other reader constructions ranging from conflict detection, thematic inference, moralization, and more intuitive feelings of story-ness. This suggests that global narrative structures play an important role in eliciting extra-textual responses from readers that, in combination with textual features, help explain why a text is perceived as a story.

4.3 Relative Discourse Associations with Stories (vs. Not Stories)

In addition to feature-level analyses, understanding narrative perception requires examining narrative communication within a broader pragmatic framework (Prince, 1983; Herman, 2009). While our entire crowd annotation task is concerned with audience reception, the question of authorial `goal` is especially tied to pragmatic aspects of narrative communication, as it centers on the audience's perception of the author's overall intent (e.g., explanation, argumentation). Analyzing these responses can shed light on the associations between storytelling and the perceived discursive aims of authors.

We compare the relative prevalence of the foremost discourse code for the authorial `goal` question for posts labeled as containing stories versus not.[8]

As illustrated in Fig. 3, storytelling is relatively more prevalent in descriptive posts, while argumentative, review, and explanatory posts are relatively more likely not to contain storytelling. This suggests that storytelling is especially *representational* discourse useful for conveying the appearance or state of things, for example, by providing an account of a sequence of events leading to a current situation or outcome. In contrast, other discourses may involve more logical forms of evidence (argumentation), comparisons of qualities and value statements (reviews), or clarifications (explanation). Our findings offer an overarching perspective on multiple broad discursive aims, complementing prior work's focused attention on storytelling in relation to specific discourses (see Section 2.3), thus providing an important background context for future work on the discursive functions of storytelling in social media.

5 Disagreement Among the Crowd [RQ2]

To further examine crowd perceptions of storytelling, we examine divergent perceptions among crowd workers. In STORYPERCEPTIONS, we observe substantial disagreement among the descriptive crowd labels (alpha=0.426) (Krippendorff, 2011), compared to Antoniak et al.'s (2024) prescriptive annotations (alpha=0.655), underscoring the subjectivity of the descriptive narrative detection task in the absence of prescriptive guidelines.

5.1 Majority vs. Minority

Our first lens into internal disagreement among the crowd is exploring why crowd workers disagree

[8]For `goal` responses with multiple discourse codes, we qualitatively select the dominant discourse, typically based on the first verb. For instance, in "to ask for recommendations," which maps to QUESTION_REQUEST and ARGUMENT_SUGGESTION_RANT, we identify QUESTION_REQUEST as the primary discourse.

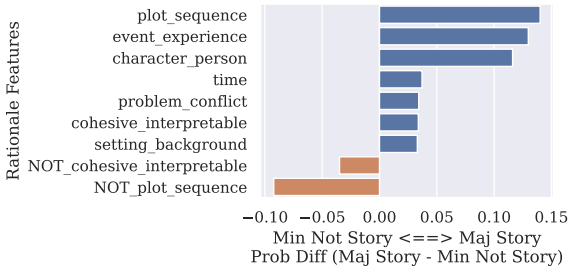

Figure 4: Relative feature code prevalence in majority story (vs. minority non-story) `rationales`. We exclude feature codes that appear fewer than 25 times due to the relatively smaller number (N=886) of `rationales` relevant to this comparison.

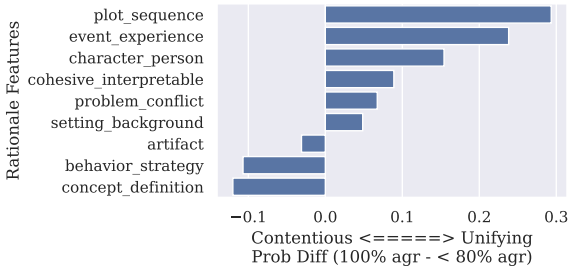

Figure 5: Relative feature prevalence with unanimously-voted stories versus substantially divided-vote stories.

with the majority vote.

First, we examine the relative prevalence of feature codes in majority story `rationales` versus minority non-story `rationales`. As shown in Fig. 4, NOT_PLOT_SEQUENCE and NOT_COHESIVE_INTERPRETABLE are relatively more prevalent in the minority non-story votes. Since PLOT_SEQUENCE and COHESIVE_INTERPRETABLE show the opposite trend, we conclude that a sense of cohesion and especially plot are contentious features for crowd workers. Notably, both feature codes concern global or holistic aspects of texts, indicating that those who disagree with the majority story vote diverge in their big-picture assessments rather than in their perceptions of more localized feature codes.

Finally, when we inspect the converse scenario – comparing the relative prevalence of feature codes in minority story `rationales` versus majority non-story `rationales` – we similarly find NOT_PLOT_SEQUENCE to be particularly contentious (see Appendix F.2 for details).

5.2 Unanimous vs. Divided Votes

We next examine why certain posts are unanimously seen as (not) stories vs. contentious.

First, Fig. 5 shows the relative prevalence of feature codes in unanimously-voted stories versus stories with substantial division, defined as a story vote rate in the range $[0.5, 0.8)$. We find that PLOT_SEQUENCE and EVENT_EXPERIENCE are more prevalent in unanimously-voted stories. Meanwhile, BEHAVIOR_STRATEGY, which covers generalized behaviors (e.g., "walking the dog every day," "smoking a pack a day") and references to abstract processes (e.g., "how to get married"), and CONCEPT_DEFINITION, which capture references to abstract ideas (e.g., 'theology', 'climate change') are more prevalent in divided-vote stories.

The contentiousness of BEHAVIOR_STRATEGY contrasts with the relative unanimity associated with EVENT_EXPERIENCE. This distinction aligns with prior work emphasizing the distinction between general descriptions of behavior and *realis* events that occur at a particular time and place (Sims et al., 2019; Antoniak et al., 2019). Overall, the findings and background context highlight that in contrast to consensus surrounding particular actions (events), abstract references to generalized actions (routines, procedures) have a more ambiguous relationship to storytelling, often leading to disagreements among crowd workers about whether a text qualifies as a story. See Appendix F.3 for the complementary plot for non-stories.

6 Disagreement Across Annotation Contexts [RQ3]

Finally, we use paired sets of annotations to explore disagreements across annotation contexts, including descriptive crowd annotations, prescriptive labels from researchers, and predictions from LLMs, including several models from the GPT-4 family[9] and the instruction-tuned Llama 3 8B (Dubey et al., 2024). We describe our prompts in Appendix E.

As shown in Table 2, pairwise agreements between researchers, crowd majority, and GPT-4 (gpt-4-0613) fall at or just below 0.6 Cohen's κ. We conclude that at the label level, the prescriptive annotations from researchers and descriptive GPT-4 predictions can generalize moderately well to certain aggregated crowd perceptions of whether a social media post contains a story. See Appendix F.5 for an illustrative comparison of story classifiers trained on researcher vs crowd majority labels, which shows that despite similar inference behav-

[9] GPT-4 (gpt-4-0613), GPT-4-Turbo (gpt-4-turbo-2024-04-09), GPT-4o (gpt-4o-2024-05-13)

Annotator Type Pair *More (L) / Less (R) Stories*	Cohen's κ	*Story (L) / Not Story (R)*			*Not Story (L) / Story (R)*		
		Freq.	Most Relatively Prevalent Code	Code Prob	Freq.	Most Relatively Prevalent Code	Code Prob
GPT-4 / crowd_maj	0.604	14%	NOT_plot_sequence	-0.26	4%	character_person	-0.42
researcher / GPT-4	0.592	10%	behavior_strategy	-0.18	9%	character_person	-0.21
researcher / crowd_maj	0.574	16%	NOT_plot_sequence	-0.22	4%	character_person	-0.44
crowd_maj / GPT-4t	0.523	17%	behavior_strategy	-0.10	2%	-	-
crowd_maj / GPT-4o	0.496	19%	behavior_strategy	-0.12	0%	-	-
researcher / GPT-4t	0.379	28%	behavior_strategy	-0.11	1%	-	-
researcher / GPT-4o	0.355	30%	behavior_strategy	-0.12	0%	-	-
Llama3 / researcher	0.302	30%	NOT_plot_sequence	-0.15	5%	event_experience	-0.21
Llama3 / crowd_maj	0.290	38%	NOT_plot_sequence	-0.27	1%	-	-

Table 2: Cohen's κ agreement metrics across pairs of descriptive annotations from crowd workers, prescriptive annotations from researchers, and descriptive predictions from GPT-4, GPT-4 Turbo (GPT-4t), GPT-4o, and the instruction-tuned Llama 3 8B. Each forward slash-separated annotator type pair in the leftmost column is ordered by which annotator type perceived storytelling more. The center set of columns details the most frequent (Freq.) disagreement scenario, in which the left (L) annotator type identified a story, but the right (R) annotator type did not. The rightmost column section details the converse scenario, in which the left annotator identified less stories than the right annotator. If the label disagreement frequency is $\geq 3\%$, we also display the feature code most associated with the disagreement scenario (lower relative probabilities indicate greater prevalence in disagreements).

ior overall, there are textual subdomains for which storytelling prediction rates diverge across models.

GPT-4t and GPT-4o agree with all human annotators less, especially with the researchers. Without access to detailed model specifications, pretraining data, and alignment procedures for each of these models, we cannot fully explain the stark difference between GPT-4 and both GPT-4 Turbo and GPT-4o.[10] Nonetheless, the variation among GPT models highlights the significant differences in LLMs' approaches to narrative classification, underscoring the importance of validating model outputs with human annotations as a necessary (though not necessarily sufficient (Agnew et al., 2024)) step before treating them as representative of aggregate perceptions from specific groups of annotators.

Llama 3 8B achieves the lowest agreement scores with human annotators, primarily due to an overprediction of stories, which indicates the model's susceptibility to confirmation bias.

Examining story labeling rates, we find that prescriptive labeling from researchers and descriptive GPT-4 predictions identify more stories than crowd workers. The feature code most relatively prevalent in scenarios where crowd workers do not identify a story (but researchers or GPT-4 do) is NOT_PLOT_SEQUENCE.[11] This suggests that crowd workers require greater structural complexity (e.g., with respect to the sequential chain of events comprising a plot) in stories.[12]

Although fully explaining divergences across annotator types and the influence of the term 'story' on crowd worker annotations is beyond our scope, we offer initial conjecture based on common intuitions and empirical findings. Although storytelling is woven into much of human discourse, we do not always name it as such. Typically, when we encounter the word 'story' (e.g., 'let me tell you a story'), there is a connoted expectation of significance or interestingness to the story. Thus, crowd workers may be more hesitant to attach the term 'story' to a text unless that text surpasses a baseline threshold for interestingness and purpose, differing from Antoniak et al.'s (2024) prescriptive labels, which focus on character and event sequences rather than story purpose or interestingness.

Concerning feature-level preferences in cases

[10]However, with evidence that larger models retain long-tail knowledge better than smaller models (Wei et al., 2022; Kandpal et al., 2023), we could conjecture that *if* GPT-4 is the largest among these models, its relatively larger size could partially explain its apparent retention of more nuance for the narrative detection task from training.

[11]Because we only have survey responses from crowd workers, there is not a straightforward way to compare features based on our qualitative coding method to researcher or LLM annotations. However, we can leverage the crowd codes and the paired labels to analyze how feature importance changes when the crowd agrees vs. disagrees with the labels from another annotator type (e.g., researchers, LLM). Separately, in Appendix F.6, we leverage the basic feature-level metrics available in the StorySeeker corpus for comparison across annotation contexts.

[12]For all other comparisons involving GPT models, where the overall agreement is lower, the GPT model predicts stories less frequently than the human counterpart, with this skew being particularly pronounced for both GPT-4t and GPT-4o.

of disagreement between a human annotator type (researcher or crowd) and a GPT-4 family of models (including GPT-4o, GPT-4t, and GPT-4), the feature code most relatively prevalent in disagreements is BEHAVIOR_STRATEGY. This bolsters and extends our finding in Section 5.2 regarding the contentiousness of BEHAVIOR_STRATEGY *among* crowd workers. Evidently, BEHAVIOR_STRATEGY is also contentious *across* annotation contexts. Excepting the GPT-4 / crowd comparison, where the disparity is less pronounced, the models in the GPT-4 family frequently avoid labeling texts as stories when they describe behaviors, habits, or plans, relative to human annotators. While this tendency ostensibly aligns with Antoniak et al.'s (2024) codebook guidance not to conflate general descriptions of behavior with events bounded in space and time, this guidance does not preclude other aspects of the text from justifying a story label. We conclude that models in the GPT-4 family lack nuance in their ability to identify stories that both (1) contain general descriptions of behavior and (2) particularized events and/or other features that contribute to storytelling, relative to human annotators who, in aggregate, approach these texts in a more balanced manner.

7 Summary of Findings

RQ1: What are crowd workers' descriptive perceptions of storytelling in social media texts? We find that while crowd workers' label `rationales` are based primarily on a few core textual features (events, characters, plot), extra-textual features, such as cognitive and aesthetic experiences while reading (sense of conflict, cohesion, feeling like a story), are also important. We additionally identify associations between crowd workers' aesthetic experiences and textual features (e.g., between a sense of wholeness and plot), and we demonstrate that crowd workers find storytelling relatively more prevalent in descriptive writing than argumentative, review, or explanatory posts.

RQ2: How do narrative perceptions differ *among* crowd workers? We find that while crowd workers generally agree on basic textual features, their holistic assessments of complex textual features (such as plot) and extra-textual aesthetic judgments (such as sense of cohesion) can diverge from one another. Additionally, distinguishing between events and more general descriptions of behavior is a particularly challenging and contentious aspect of narrative detection for crowd workers.

RQ3: How do narrative perceptions differ *across* prescriptive labels from researchers, descriptive annotations from crowd workers, and predictions from LLM-based classifiers? Through pairwise label comparisons across annotation contexts, we conclude that prescriptive labels from researchers and descriptive GPT-4 predictions can approximate certain aggregated crowd perceptions of narrative status reasonably well for short text-based social media posts. Important qualifications include differing thresholds for structural complexity (with crowd workers having a stricter definition of plot) and a tendency of GPT-4 family models to diverge from human perceptions for texts that describe behaviors, habits, or abstract plans.

8 Conclusion & Future Work

In this paper, we introduced the STORYPERCEPTIONS dataset to bring crowd workers' descriptive perceptions to bear on the narrative detection task. Complementing prior work that uses prescriptive annotations from a small number of researchers, our empirical findings highlight the types, relative importance, and contentiousness of a broad range of features for narrative perception.

Our study points to several opportunities for further research. First, while we looked at simple co-occurrence of features, deeper statistical analysis and experimentation could more clearly illuminate interactions and causal relationships between features. While we offer a pilot experiment in training a story classifier with aggregated labels from crowd workers (see Appendix F.5), future studies could survey more sophisticated methods to incorporate multiple perceptions during training. Further research could also integrate a broader range of features from our taxonomy—both textual and extra-textual—into narrative detection models. Moreover, replicating our crowd studies using alternative terminology, such as 'narrative,' would help assess whether our findings generalize or reveal important distinctions across different terms. Finally, as attention in computational narrative understanding broadens beyond textual features to include reader reception, there are major outstanding ethical and epistemic questions concerning the use of LLM-assisted annotation; an exploration of these questions tailored for the subfield of computational narrative understanding would be invaluable.

9 Limitations

We follow Antoniak et al. (2024) in adopting a simple binary definition of stories, in contrast to scaled labels, such as in Piper and Bagga's (2022) proposal to use a Likert scale for annotating the degree of narrativity in a text.

We broadly define "researcher" as any researcher who develops or uses an annotation guideline with reference to prior work in the field of narrative theory (and optionally NLP). We do not consider differences among this broad category of researchers, e.g., between an NLP researcher working on narrative detection and a subject matter expert from the field of narrative theory. Future work could compare annotation tendencies across more finely partitioned expertise levels.

While we compare prescriptive labeling from researchers with descriptive annotations from crowd workers, we do not disentangle these pairings to investigate other combinations, such as descriptive annotations from researchers or prescriptive annotations from crowd workers.

Moreover, our feature-level analyses rely primarily on our qualitative coding of crowd workers' free-text `rationales`. While we understand from the researchers' codebook that their prescriptive labels emphasize the presence of event chains and characters, we lack comparable instance-level free-text rationales for our dataset. Additionally, for the LLM-based predictions, we do not examine generated rationales or apply other interpretability methods, such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), which could yield deeper feature-level insights into LLM-based narrative detection.

As described in Appendix B.1, the demographics of our 255 crowd workers skew toward U.S.-based, English-speaking white adults, aged 22–44, with undergraduate-level educations. Consequently, we do not expect our findings to generalize to all readers. For example, a study on narrative perception in children might produce different results. Along with variations in crowd perceptions of narratives, different demographic samples may align more or less with either prescriptive annotations from researchers or descriptive predictions from LLMs. Future work should avoid over-generalizing by assuming our results apply to an "ideal lay reader" or that the observed correlations across annotator type pairs will hold across different samples within those populations.

Our work relies on a dataset of English-language texts sampled from Reddit. We do not necessarily expect our results to generalize beyond this setting, as different languages, cultures, and data sources might bias the crowd workers in various ways. The stories in the StorySeeker dataset are relatively short, informal, and typically nonfictional accounts of personal experiences, written in the 1st-person perspective. We expect that the degree to which our findings about narrative perceptions will generalize beyond the social media context will depend on how closely the profiles of the target stories (i.e., in terms of formality, point-of-view, and length) correspond to our dataset. One advantage of our dataset in this regard is the topical diversity of its stories. Furthermore, while we do not have access to the data, crowd workers, or expertise needed to run a multilingual study, we hope our work can support future work that draws comparisons across languages.

10 Ethical Considerations

Our study was considered exempt by the IRB at the Allen Institute for AI, as no information was collected that could identify the workers. Workers were paid an average of $15/hour and were given a description of the study before opting in and could exit at any time.

The StorySeeker dataset, the source of our annotation texts, contains posts and comments from diverse subreddits. These subreddits were filtered for toxicity and sensitivity, and the individual posts and comments were also filtered for toxicity to protect the annotators.

11 Acknowledgements

We thank our anonymous reviewers for their detailed feedback and suggestions. We also thank Jocelyn Shen for her insightful comments on an early draft of this work.

References

Yosra Abdessamed, Shadi Rezapour, and Steven Wilson. 2024. Identifying Narrative Content in Podcast Transcripts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2643, St. Julian's, Malta. Association for Computational Linguistics.

William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir

Mohamed, and Kevin R. McKee. 2024. The illusion of artificial inclusion. ArXiv:2401.08572 [cs].

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):88:1–88:27.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where Do People Tell Stories Online? Story Detection Across Online Communities. ArXiv:2311.09675 [cs].

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24. Number: 1.

Mieke Bal. 1997. *Narratology: Introduction to the Theory of Narrative*, 2nd edition edition. University of Toronto Press, Scholarly Publishing Division, Toronto ; Buffalo.

Roland Barthes and Lionel Duisit. 1975. An introduction to the structural analysis of narrative. *New literary history*, 6(2):237–272.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Lucas M. Bietti, Ottilie Tilston, and Adrian Bangerter. 2019. Storytelling as Adaptive Collective Sensemaking. *Topics in Cognitive Science*, 11(4):710–732. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12358.

Marisa Bortolussi and Peter Dixon. 2002. *Psychonarratology: Foundations for the Empirical Study of Literary Response*. Cambridge University Press, Cambridge.

G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

William F. Brewer and Edward H. Lichtenstein. 1982. Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics*, 6(5):473–486.

Jerome Bruner. 1991. The Narrative Construction of Reality. *Critical Inquiry*, 18(1):1–21. Publisher: The University of Chicago Press.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

B. Ceran, R. Karad, A. Mandvekar, S. R. Corman, and H. Davulcu. 2012. A Semantic Triplet Based Story Classifier. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 573–580, Istanbul. IEEE.

Betul Ceran, Nitesh Kedia, Steven R. Corman, and Hasan Davulcu. 2015. Story Detection Using Generalized Concepts and Relations. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 942–949, Paris France. ACM.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Michael F. Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(supplement_4):13614–13620. Publisher: Proceedings of the National Academy of Sciences.

Sarah Dillon and Claire Craig. 2019. Storylistening: Narrative Evidence and Public Reasoning.

A. Dirkson, S. Verberne, and Wessel Kraaij. 2019. Narrative Detection in Online Patient Communities.

Henrique D. P. dos Santos, Vinicius Woloszyn, and Renata Vieira. 2017. Portuguese personal story analysis and detection in blogs. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 709–715, New York, NY, USA. Association for Computing Machinery.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo,

Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick,

Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models.

Joshua Eisenberg and Mark Finlayson. 2017. A Simpler and More Generalizable Story Detector using Verb and Character Features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2708–2715, Copenhagen, Denmark. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2024. Stories and Personal Experiences in the COVID-19 Discourse. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15320–15340, Torino, Italy. ELRA and ICCL.

Achyutarama Ganti, Steven Wilson, Zexin Ma, Xinyan Zhao, and Rong Ma. 2022. Narrative Detection and Feature Analysis in Online Health Communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science*, 48(4):1093–1118.

Evelyn Gius and Michael Vauth. 2022. Towards an Event Based Plot Model. A Computational Narratology Approach. *Journal of Computational Literary Studies*, 1(1). Number: 1 Publisher: Universitäts- und Landesbibliothek Darmstadt.

A. Gordon and Reid Swanson. 2009. Identifying Personal Stories in Millions of Weblog Entries.

Andrew Gordon, Luwen Huangfu, Kenji Sagae, Wenji Mao, and Wen Chen. 2013. Identifying Personal Narratives in Chinese Weblog Posts. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 9(4):23–29. Number: 4.

Andrew S. Gordon, Qun Cao, and Reid Swanson. 2007. Automated story capture from internet weblogs. In *Proceedings of the 4th international conference on Knowledge capture*, pages 167–168, Whistler BC Canada. ACM.

Andrew S. Gordon and Kavita Ganesan. 2005. Automated story capture from conversational speech. In *Proceedings of the 3rd international conference on Knowledge capture*, pages 145–152, Banff Alberta Canada. ACM.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan. ACM.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395.

Melanie C. Green and Timothy C. Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, 79(5):701–721.

David Herman. 2009. *Basic Elements of Narrative*. John Wiley & Sons, Incorporated, Newark, UNITED STATES.

Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2024. Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia. ELRA and ICCL.

Peter Hühn. 2009. Event and Eventfulness.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. ArXiv:2211.08411 [cs].

K. Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions.

Robin L. Nabi and Melanie C. Green. 2015. The Role of a Narrative's Emotional Flow in Promoting Persuasive Outcomes. *Media Psychology*, 18(2):137–162.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report.

Maja Pavlovic and Massimo Poesio. 2024. The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

David Pellauer and Bernard Dauenhauer. 2022. Paul Ricoeur. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2022 edition. Metaphysics Research Lab, Stanford University.

Inês Pereira Rodrigues. 2023. "Who do you say that I am?" Truth in Narrative Identity. *Études Ricoeuriennes / Ricoeur Studies*, 14(1):132–150.

Federico Pianzola and Franco Passalacqua. 2016. Epistemological Problems in Narratve Theory: Objectvist vs. Constructvist Paradigm. pages 195–217.

Andrew Piper and Sunyam Bagga. 2022. Toward a Data-Driven Theory of Narrativity. *New Literary History*, 54(1):879–901. Publisher: Johns Hopkins University Press.

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.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francesca Polletta and John Lee. 2006. Is Telling Stories Good for Democracy? Rhetoric in Public Deliberation after 9/II. *American Sociological Review*, 71(5):699–723. Publisher: [American Sociological Association, Sage Publications, Inc.].

Gerald Prince. 1983. Narrative pragmatics, message, and point. *Poetics*, 12(6):527–536.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Conference Name: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ISBN: 9781450342322 Place: San Francisco California USA Publisher: ACM.

Sharon Richardson. 2021. Against generalisation: Data-driven decisions need context to be human-compatible. *Business Information Review*, 38(4):162–169.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Johnny Saldaña. 2013. *The coding manual for qualitative researchers*, 2nd ed edition. SAGE, Los Angeles. OCLC: ocn796279115.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A. Smith, James W. Pennebaker, and Eric Horvitz. 2022a. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022b. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Robert J. Shiller. 2019. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Wei Song, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Anecdote Recognition and Recommendation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2592–2602, Osaka, Japan. The COLING 2016 Organizing Committee.

Max Steg, Karlo Slot, and Federico Pianzola. 2022. Computational Detection of Narrativity: A Comparison Using Textual Features and Reader Response. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Meir Sternberg. 2001. How Narrativity Makes a Difference. *Narrative*, 9(2):115–122. Publisher: Ohio State University Press.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Suzan Verberne, Anika Batenburg, Remco Sanders, Mies Eenbergen, Enny Das, and Mattijs Lambooij. 2018. Social processes of online empowerment on a cancer patient discussion form: using text mining to analyze linguistic patterns of empowerment processes (Preprint). *JMIR Cancer*, 5.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023a. Prevalence and prevention of large language model use in crowd work.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023b. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. ArXiv:2206.07682 [cs].

Hayden White. 1980. The Value of Narrativity in the Representation of Reality. *Critical Inquiry*, 7(1):5–27. Publisher: The University of Chicago Press.

A Prior Work in Narrative Detection

Table 3 summarizes prior work in narrative detection. Generally, small numbers of researchers have followed prescriptive codebooks to assign gold labels to texts.

Prior Work	Ann Type	# Ann	# Ann / Text	Codebook Emphasis	Model Features	Lang
Gordon and Ganesan (2005)	grad students, staff	5	unk	event sequence, purpose	n-grams	en
Gordon et al. (2007)	grad students, staff	5	unk	event sequence, purpose	n-grams, POS	en
Gordon and Swanson (2009)	first author	1	1	event sequence, purpose, characters	n-grams	en
Ceran et al. (2012)	domain experts	unk	1	actions, characters	n-grams, POS, NE, stative verb rate, semantic triplets	en
Gordon et al. (2013)	native speakers grad. from Chinese Uni.	6	1	event sequence, characters	n-grams	zh
Ceran et al. (2015)	domain experts	unk	1	actions, characters	n-grams, semantic triplets, misc. generalized extensions of semantic triplets	en
Song et al. (2016)	Literature students	2	2	N/A, but un-supervised methods based on characters, event chains	POS, semantic triplets	en
dos Santos et al. (2017)	crowd workers	167	3	personal experiences, characters, actions/behaviors	LIWC, TF-IDF n-grams, LDA topics, syllable count, connectives, POS	pt
Gerber et al. (2018)	research team	2	2	examples	N/A	en
Yao and Huang (2018)	trained annotators	2	2	N/A, but un- and semi-supervised methods based on linear event sequences, characters	grammar production rules, perplexity of event chains (wrt corpus), NE, LIWC, POS	en
Dirkson et al. (2019)	domain expert, first author	2	2	personal experiences	n-grams, LIWC	en
Falk and Lapesa (2022)	N/A	N/A	N/A	N/A. Aggregation of pre-existing datasets prescriptively labeled for 'storytelling' and 'testimony'	BERT tokens	en
Ganti et al. (2022)	experts	2	2	unk	BERT tokens	en
Piper and Bagga (2022)	trained students	3	3	agency, event sequences, world-making	NER, TimeML, TLINKS, WordNet, concreteness lexicon, event rate, LIWC, animacy, entity recurrence, POS, n-grams	en
Steg et al. (2022)	Uni students, professor	7	3	suspense, curiosity, surprise	TF-IDF n-grams, concreteness lexicon	en
Falk and Lapesa (2023)	trained students	4	3-4	event sequence	N/A	en
Antoniak et al. (2024)	research team	2	2	characters, event sequence	RoBERTa tokens	en

Abdessamed et al. (2024)	research team	2-3	2-3	characters, event sequence	LIWC Narrative Arc, DeBERTa tokens	en

Table 3: Summary of prior work in story detection, with respect to annotation procedure. Prior work has favored prescriptive annotations from experts over descriptive crowd-sourced annotations for developing narrative detection datasets. The one work that involves a large crowd study is based on Portuguese blogs (dos Santos et al., 2017). All other works concern English texts.

B Survey Details

We recruited workers via the Prolific[13] platform.

Crowdwork with free text responses has faced increasing challenges as public-facing text generation tools like ChatGPT have become both more fluent and more accessible. Prior work has measured alarmingly high rates at which crowd workers use LLM-based tools to generate their free text responses (Veselovsky et al., 2023b). However, related work has also found that simple mitigation strategies can substantially reduce this rate (Veselovsky et al., 2023a). Following this prior work, we have removed the ability to paste into the free text boxes in our annotation interface. We have also added an explicit request not to use LLM-based tools like ChatGPT, and we have tried to keep the free text responses short and easy to fill out. After making these changes in our pilot studies, we observed a significant increase in task completion time, suggesting that fewer participants were using text-generation tools to expedite the annotation process. Finally, our qualitative inspection of responses through the open coding process further strengthened our confidence that virtually all submissions are human-written.

B.1 Demographics

We required that workers be located in the U.S., over the age of 18, and fluent in English. To improve the quality of responses, we also required that the workers have an approval rating of 99-100, have completed at least 100 prior submissions, and have at least an undergraduate degree. We found that removing the undergraduate degree requirement resulted in significantly lower-quality annotations.

[13] https://www.prolific.com/

Gender	51.4%	Woman (including Trans Female/Trans Woman)
	45.5%	Man (including Trans Male/Trans Man)
	3.1%	Non-binary (would like to give more detail)
Education	78.4%	Undergraduate degree (BA/BSc/other)
	16.9%	Graduate degree (MA/MSc/MPhil/other)
	4.7%	Doctorate degree (PhD/other)
Race/Ethnicity	54.5%	White/Caucasian
	11.8%	Black/African American
	9.4%	East Asian
	8.6%	Latino/Hispanic
	6.3%	Mixed
	3.1%	South East Asian
	2.0%	Native American or Alaskan Native
	2.0%	South Asian
	0.8%	African
	0.8%	White / Sephardic Jew
	0.8%	Other
Age	32.5%	24-34
	31.4%	34-44
	16.1%	44-54
	9.8%	54-64
	7.5%	18-24
	2.4%	64-74
Degree Subject	17.2%	Social Sciences
	12.9%	Arts & Humanities
	11.0%	Other
	10.6%	Information and Communication Technologies
	10.2%	Health and welfare
	7.8%	Engineering, manufacturing and construction
	7.8%	Natural sciences
	7.1%	Administration & Law
	6.3%	Education
	3.9%	Mathematics and statistics
	2.0%	Services
	1.6%	Agriculture, forestry, fisheries and veterinary
	1.2%	Journalism & Information Business
	0.4%	History
Reddit Use	70.2%	Regular use ($>$ once per month)
	29.8%	$<$ once per month

Table 4: Demographic information for our 255 crowdworkers. The categories and their descriptions are not designed by us; they are prescreening questions that Prolific asks of all their workers.

B.2 Task Description

The following task description was used to advertise the task to workers.

> Welcome! This is a study about storytelling on the internet.
>
> We will show you some example texts, and for each text, we will ask you whether the text contains a story and to explain your reasoning.
>
> We have applied some content filters, but because the texts come from online forums, there might be content that could be upsetting or NSFW.
>
> We will use this dataset to study stories computationally, and the final dataset of labels and texts (without any identifiers) will be released for other researchers.
>
> This study involves writing short text responses, and we have disabled the ability to paste into the the response boxes.
>
> Please do note use AI tools like ChatGPT to answer these questions. We really appreciate your work! We'd prefer that you write short responses rather than

use AI to write responses that would really mess up our scientific results. We're interested in your opinion, not a bot's opinion!

Feedback: If you have any questions, feedback, or concerns about this study, please feel free to send us a message. We're very happy to talk with you to improve our study!

About Us: We're researchers at [redacted for privacy]. Our team includes researchers in AI, English literature, and political science.

B.3 Survey Questions

1. "How familiar are you with the topic of this text?" (Likert)
2. "What is the author's goal in writing this text? Finish the sentence: The author of this text wants to _____." (Free-text)
3. "Does this text contain a story?" (Binary)
4. "How confident are you in your answer to Question 3" (Likert)
5. "Explain your answer to Question 3 by writing a short list of reasons." (Free-text)
6. "If you answered YES to Question 3, copy and paste the part of the text that IS A STORY into this box." (Free-text)
7. "If you answered NO to Question 3, what is this text? Finish the sentence: This is not a story, it's a _____." (Free-text)

C Methodological Approach to Coding Crowdworkers' Free-Text Responses

The variety, nuance, and mix of both positive and negative assertions about the presence of features in the crowd workers' free-text responses led us to open and axial coding as a primary analytical lens in this work. Open and axial coding refer to a bottom-up, manual, and cyclical process of surfacing ideas and claims from a population of texts and abstracting those ideas and claims into a set of themes or codes appropriate to the data (Saldaña, 2013). After a set of codes is developed and validated, a researcher assigns the codes to the data samples, which then allows for basic quantitative analyses of the data, grounded in attentive qualitative description.

C.1 Open Coding Process

Initially, one author read through a batch (N=100) of the crowd workers' free-text survey responses, noting down unique observations and claims for why or why not a given text contains a story. The author then marked which of the ideas seemed to have repeated mentions across the batch. The author then restarted, but this time on a larger batch (N=200). This process continued, jumping to a large sample size (N=1000) by the fourth iteration. Next, the author, reviewing their notes and scanning through the data as needed, attempted to abstract 30-40 core ideas or claims from the notes. We arrived at our initial taxonomy after associating each of those ideas/claims with an I.D., associated keywords, and a short description.

The author used this taxonomy to annotate a batch of free-text responses in a multi-label fashion. They repeated this on increasingly large batches of data, slightly revising the taxonomy and re-coding data samples. After labeling N=1000 data samples with one version of the taxonomy, the author presented the taxonomy and initial results to coauthors for feedback and discussion.

After two coauthors, one of whom used the taxonomy to annotate a small batch of samples (N=25), provided a final round of feedback, we developed a final version of the taxonomy, which is described in Appendix D.

D Taxonomy of Features and Discourse Categories Used to Explain the Presence or Absence of Storytelling

D.1 Introduction to the Taxonomy

We construct the taxonomy below through a process of open coding of free-text rationales from crowd workers reasoning about the intents of social media posts and explaining why or why not a text contains a

story. See Section C for background on our motivation and process for using open coding to analyze the crowd workers' responses.

To understand what exactly the taxonomy represents, its important to relate it back to the crowd work annotation task and subsequent open coding procedure that created it. We list a few key observations from our experience conducting open coding that help contextualize and explain the structure and content of the taxonomy.

1. The taxonomy was developed in a bottom-up fashion based on crowd workers' perceptions about what a text is or contains (e.g.,a story) and the writer's goal was in posting the text. While the primary author who developed the taxonomy is familiar with narrative theory to some degree, and that could influence interpretation the crowd workers' responses, the goal was to impose as little theoretical background onto the codes, especially in the early stages of open coding.
2. Despite that there were three distinct free-text questions in our survey,[14] the author performing open coding observed that many ideas and claims in the free-text responses manifested in not just one, but in two or three of the questions. For this reason, we developed a unified taxonomy, based on the free-text rationales for all questions. Consequently, one should not assume that the presence of a code in the taxonomy necessarily means that it is positively associated with storytelling.
3. The taxonomy is relatively flat, in that we do not organize the codes into a large number of subcategories arranged hierarchically. Rather, we define two basic categories into which all codes fall. First, there are *discourse categories* as a distinct set of codes. These refer to broad types of communication, distinguished in part by their pragmatic purpose and the associated syntax. Examples include "description", "explanation", "argument", and "inquiry". As a rule of thumb, these are different from the rest of the codes because they can function as either a noun or a verb (e.g., "description" vs "describe"). The second category, called "features", is purposefully generic. This category contains codes that are typically invoked as textual features of stories (e.g., "characters", "events", "setting"), the aesthetic or interpretative experience of the reader (e.g., "evocative", "cohesive"), or highly abstract things that crowd workers often refer to but do not neatly fall into the other categories (e.g., "artifact", "emotion", "plan"). However, we often find that these distinctions are porous, for instance events are not always associated with stories, and emotions can manifest in stories or in other kinds of texts. For this reason, we opt to keep the "features" category flat.

To summarize, the taxonomy depicts the key discourse categories, textual features, and reading experiences that crowd workers refer to when reasoning about the goal of social media texts and why or why not those texts contain stories.

D.2 The Taxonomy

The taxonomy is organized into two groups of codes: *features* and *discourse categories*. The short names of the codes in our taxonomy conjoin 1-3 representative keywords for the concept represented by the code. We also include a list of keywords associated with code as well as a short comment that define the codes in greater detail and provide concrete annotation tips. Table 5 lists the discourse codes and Table 6 lists the feature codes.

D.2.1 Discourses

Table 5 lists the discourse codes in our taxonomy.

D.2.2 Features

Table 6 lists the feature codes in our taxonomy.

[14] We asked about perceived goal of the post, an explanation for why or why not the text is a story, and, if the user decided that the text did not contain a story, we asked what else they thought it was (e.g.,a review). See B.3 for the precise language.

Short Name	Keywords	Comment
narrative_story	narrative, narration, narrate, story, storytelling, retelling, recount, account, anecdote	An account of "a sequence of events involving one or more people" (Antoniak et al., 2024).
question_request	question, ask, request, seeking, inquiry	A question, request, or appeal.
explanation	explanation, explain, theorize, educate	Statements that contextualize or clarify a situation, concept, opinion, etc.
description_expression	description, describe, expression, convey, communicate or share what something feels like, tell, talk about, information, communicate information, provision, provide, manual, observation	Detailed representation of something. Note: if the author is described as 'troubleshooting', that is considered implicit description of a troubleshooting procedure.
argument_suggestion_rant	argument, argue, rebuttal, rebut, proposal, propose, recommendation, recommend, advertisement, advertise, warning, warn, advise, advice, rant, editorial, guide	Statements that aim to influence a reader. Ranges from solicited advice and recommendations to logical arguments to illogical, fiery arguments and unwelcome advertisements. Distinguished from the 'education_documentation' category, which is concerned with relatively dispassionate forms of influence (e.g. instruction-sharing, education).
review	review; analysis, analyze, evaluation, evaluate, discussion, discuss	An assessment of an artifact (e.g. game review) or set of arguments or opinions. Typically discusses multiple perspectives in good faith before making an evidence-based judgment. Distinguished from argument_suggestion and opinion by the method of arriving at the conclusion. A discussion may contain arguments and opinions.
dialogue	dialogue, conversation, back-and-forth, forum post, blog post, letter, email	References to conversation between characters_persons, or references to the dialogical nature of the communication medium itself.
entertainment	entertainment, entertain, funny, joke, humor, comedy	Artistic text, intended to be funny, enjoyable, challenging, etc for its audience.
sense-making	processing, making sense of, working through, reflection, introspection	The use of language as a means to understand something, such as a memory or complex concept, in a better or new way.
specialized_domain	legal, scientific, poetry, math, diary, speech, c.v., presentation, essay, etc	Catch-all for other types of discourses that may span multiple different categories (e.g. essays) or have their own specialized forms (math).

Table 5: Taxonomy discourse categories.

Short Name	Keywords	Comment
character_person	I, character, protagonist, he, she, they	A person or anthropomorphic agent. Includes the author if text is written from a first-person perspective.
event_experience	event, experience, action, happening, interaction	An event is "a singular occurrence at a particular place and time" (Sims et al., 2019). Distinguished from general, repeating behavior and continuous states.
plot_sequence	event sequence; structured progression; arc; beginning, middle, end; plot; storyline; flow of events or experiences	A structured progression of events involving characters.
problem_conflict	problem, issue, conflict, dilemma.	An issue or conflict.
setting_background	setting, background, context, sets the stage, situation, world-building	The context—environmental or social—in which persons may find themselves or events may transpire. Cursory reference to an individual detail related to setting is not sufficient, unless it significantly affects how the broader story or discourse should be interpreted.
literary_device	literary device	Figurative language (e.g. metaphor, simile, personification).

Continued on next page

Table 6 – continued from previous page

Short Name	Keywords	Comment
theme_moral	theme, moral, point, message	A core idea or takeaway from the text. Can be intended by the author or constructed by a reader's interpretation. Distinguished from "concept_definition" which is a more general category, not necessarily associated with stories.
fictional_hypothetical	fiction, made up, imaginary, hypothetical, hasn't happened yet (!: non-fiction, biography, fact, actually happened, real, true, personal history)	Reader classifies text as fictional or hypothetical. The negative version of this code stands for nonfictional_factual, defined as explicit appeals to as facts, or events described as occuring in real-life. Note that we require explicit references to nonfictuality or factuality to apply the negative code (otherwise we would apply this code to virtually all responses). Note: reference to a "personal experience" isn't enough to justify assigning the negative version of the code.
evocative_transporting	evocative, transporting, paints a picture, takes you on a journey	Reader expresses feeling pulled into and immersed in a constructed world, e.g. visualizing imagery after reading vivid language.
cohesive_interpretable	cohesive, coherent, complete, meaningfully interconnected, flow (e.g. like a story), whole, resolution, interpretable, clear, understandable	Reader reports that all the parts of the text fit together well and/or having a resolution, creating a highly readable and satisfying whole. Reader states that the text was understandable or was well-written in such a way that it lends itself to interpretation and consideration.
suspenseful	suspense, suspenseful, attention-grabbing, edge-of-your-set	Reader reports structured sequence of emotions or tensions while progressing through the story. Otherwise acknowledges that a story commanded their attention.
creative	creative, original	Reader acknowledges the unique artistic choices that individuate the text.
feels_like_story	feels like a story	Reader asserts that the text feels like a story, or draws attention to personalized definition of storytelling that informs judgment that text is or is not a story. Note: when a reader points to a lack of focalization on story-like parts of a text, then the negated version of this code should be assigned.
implicitly_revealing	reveals, tells us something about author	Reader suggests that the text implicitly reveals something about the author, beyond that which can be inferred from a surface-level reading of the text.
opinion	opinion, theory, belief, complaint	An idea held by a person or group that is unproven or not widely accepted as true. Distinguished from argument_suggestion_rant by not being focused on the content of the opinion and not necessarily an attempt to persuade someone that the opinion is correct.
behavior_strategy	activity, behavior, process, troubleshooting, plan, approach, method, options, future plans, choices, instructions	Types of actions, habitual actions, or a defined sequence of events, discussed abstractly. Distinguished from event_experience by either being general or repeating. Often associated with a personality or supply chain. A plan of action. Note that advice is considered as argument_suggestion_rant. If an activity, behavior, process, or troubleshooting procedure is depicted as having been executed or performed once in a concrete setting, then also code event_experience.
concept_definition	concept, definition, idea, state of the world, how it is, how it works, what something is	An abstract idea, or a statement about what something means. While this information may come from individuals with biases, the information itself is relatively stable, and is not in and of itself meant to aggressively influence or persuade.

Continued on next page

Table 6 – continued from previous page

Short Name	Keywords	Comment
artifact	artifact, object, game, video, show, movie, book, car, medication	A physical object (excluding persons), typically created by a human. Distinguished from setting_background by a lack of broader context surrounding the reference to the artifact.
emotion_sensation	emotion, sensation, perception, impression, happiness, elation.	Embodied sensation, or emotion. A category of feeling. Could apply to character_person, the imagined reader (from the author's perspective), or the actual reader.
time	past, future, present, chronological, time, timeframe	Must explicitly mention time or the passage of time

Table 6: Taxonomy feature categories.

D.3 Annotation Rules

We assign a code to a free-text response if the response references the keywords or concepts associated with the code. To be clear, we do not require an exact keyword match to assign a code; we also consider synonyms and basic logical implications (e.g., a reference to a 'personal story' implies that there is at least one 'character_person' – the narrator themself).

If a free-text response refers to a code multiple times, we assign the code only once. This has no effect on the relative prevalence metric we use throughout the paper, which is based on empirical probabilities that a code is referenced (at least once) in a free-text response for a given question type.

D.4 Examples

D.4.1 Example 1

Paraphrased StorySeeker text

> I tip coins for info, funny stuff, or when I see someone on the forum asking who genuinely needs it. There's always an incentive to trip. I've used 1/4 of my coins on this thread alone. The reason I don't tip more is I want to save up to be able to give a larger amount of the future (Halfway through writing this I just tipped 40 coins for a forum I saw in the forum). People are kind, and I know that some time in the future I will have to be, and people will help me out. That's the spirit of this crypto coin! I tipped you to prove a point and for the article that was just an example of what can go wrong if people are stingy. The coin was built upon tipping and keeping coins flowing, and it's all thanks to the community! Tips are sometimes small, but if you pay attention there's always a positive reason for that.

Crowd worker response
story `label`: not story
label `rationale`: "I think this is explanatory writing. No plot, fiction, literary devices, characters, etc. It is a very informally, and frankly, confusingly, written explanation about a person's interactions with crypto."
label `confidence`: 4/5
paraphrased story `span`: `N/A`
perceived `goal` of author: "explain their position regarding tipping Dogecoin."
`alternative` classification: "explanation of personal behavior"
text topic `familiarity`: 1/5

Our meta-annotations
goal `rationale` codes: EXPLANATION, OPINION, BEHAVIOR_STRATEGY, ARTIFACT
label `rationale` codes: EXPLANATION, NOT_PLOT_SEQUENCE, NOT_FICTIONAL_HYPOTHETICAL NOT_LITERARY_DEVICE NOT_CHARACTER_PERSON, NOT_COHESIVE_INTERPRETABLE, CHARACTER_PERSON, ARTIFACT
`alternative` classification codes: EXPLANATION, CHARACTER_PERSON, BEHAVIOR_STRATEGY

D.4.2 Example 2

Paraphrased StorySeeker text

> Thriving in medicine is exactly like doing those things in other professions. The most important thing is learning about, yourself, your habits, your relationship with sleep, motivations, annoyances, capabilities. Then you simply match these things to your options. Medicine is better than other options partly because there are so many different options, and many of us would probably do well in several of them. It is a mistake to pick a specialty only based on pay or theoretical interest in the concepts. You have to actually like the day to day work. A nephrologist who makes 315k a year and loves thinking about the physiology of the tubule but legitimately hates the tedium of activities like correcting fluid balance or electrolyte disequilibriums made a big mistake by becoming a nephrologist. Be the job that was easiest to get out of bed for during your rotational training. Do the field that didn't have you looking at your watch every 10 minutes after 2:30 PM. What you hate, someone else might love.

Crowd worker response

story `label`: story
label `rationale`: "i think the part where the writer shares a scenario of someone who has chosen the wrong specialty counts as a story, just the scenario part, because i imagined the nephrologist getting out of bed, going to work, and doing these tasks in some vague way. Does that make a story? I'm not sure but it feels like one"
label `confidence`: 3/5
paraphrased story `span`: "A nephrologist who makes 315k a year and loves thinking about the physiology of the tubule but legitimately hates the tedium of activities like correcting fluid balance or electrolyte disequilibriums made a big mistake by becoming a nephrologist"
perceived `goal` of author: "to convince people in medicine to go into a field for passion"
`alternative` classification: N/A
text topic `familiarity`: 3/5

Our meta-annotations

label `rationale` codes: CHARACTER_PERSON. PLOT_SEQUENCE, EVOCATIVE_TRANSPORTING. EVENT_EXPERIENCE, FICTIONAL_HYPOTHETICAL. THEME_MORAL
goal `rationale` codes: ARGUMENT_SUGGESTION_RANT, BEHAVIOR_STRATEGY
`alternative` classification codes: `N/A`

E LLM Prompts

We design our prompts to mirror the descriptive annotation task presented to crowd workers in our survey. Crowd workers were presented a text and the question "Does this text contain a story?" with the option to select 'YES' or 'NO'. Because LLM outputs are known to be more sensitive to minor changes in prompts than humans are sensitive to paraphrases, we use 5 paraphrases of the original question and collect independent predictions using each variant. We use the per-text majority vote among the five predictions as the final label.

The question variants include:

1. Does this text contain a story?
2. Is there a story in this text?
3. Is a story present in this text?
4. Does this text include a story?
5. Is there a story embedded in this text?

The full prompt template is shown below, where `MODEL` is replaced with one of "gpt4", "gpt4t", "gpt4o", or "llama3".

[QUESTION VARIANT] Respond with a "yes" or "no" decision, then provide a brief rationale.

Text: [TEXT]

Respond with JSON in the following format. Do not output anything except valid JSON.
{"MODEL_descriptive_label_[QUESTION_INDEX]": "",
"MODEL_descriptive_label_rationale_[QUESTION_INDEX]": ""}

F Additional Results

F.1 Feature Prevalence In Story Rationales

As a complement to the *relative* feature prevalence metrics depicted in Fig. 2, here we show the *independent* feature prevalence metrics for crowd story `rationales` (Fig. 6).

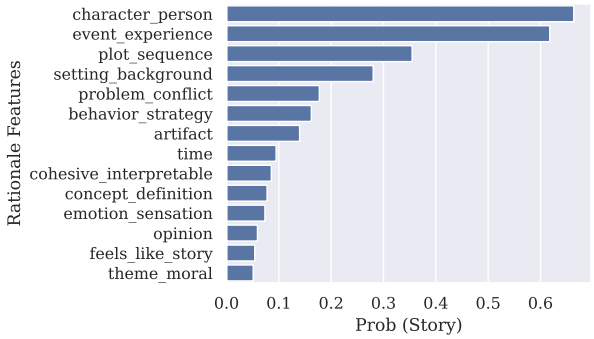

Figure 6: The prevalence of features in story `rationales`. We exclude features mentioned fewer than 50 times.

F.2 Relative Feature Prevalence in Minority-Voted Story (vs. Majority-Voted Non-Story) Rationales

Fig. 7 plots the relative prevalence of feature codes in minority story `rationales` versus majority non-story `rationales`, complementing the converse result presented in Section 5.1. We similarly find that perceptions of plot (and its absence) vary among crowd workers.

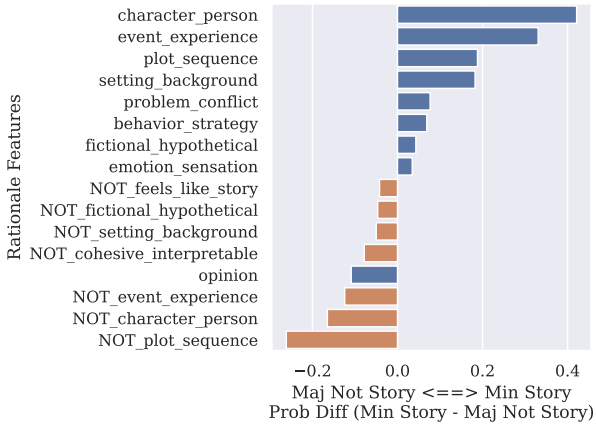

Figure 7: Relative feature code prevalence in minority story (vs. majority non-story) `rationales`.

F.3 Relative Feature Prevalence in Unanimously-Voted Non-Story (vs. Divided Vote Non-Story) Rationales

Complementing the result presented in Section 5.2, Fig. 8 shows the relative prevalence of feature codes in rationales for unanimously-voted non-story posts versus non-story posts with substantial division.

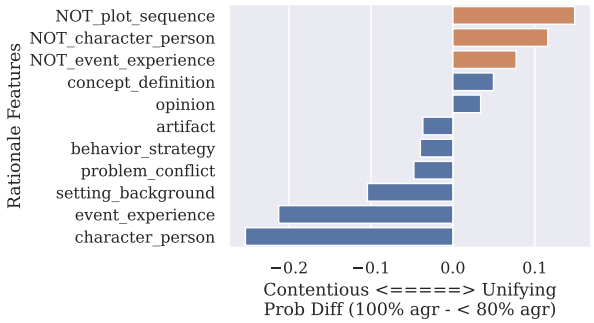

Figure 8: Relative feature prevalence unanimously-voted non-stories versus substantially divided vote non-stories.

CHARACTER_PERSON's and EVENT_EXPERIENCE's relative prevalence in divided non-story votes aligns with our understanding that stories typically include characters and events, which may lead some crowd workers to assign the story label even if most others believe that the text lacks other features required to earn the story label.

F.4 Additional Feature Co-Occurrence Results

We present the 10 most and least co-occurring feature codes in story `rationales` in Table 7. In contrast to Table 1, we do not show only those pairs that consist of one textual and one extra-textual feature. Moreover, the full heatmap of feature co-occurrence scores in story `rationales` is shown in Fig. 9.

Most Co-occurring Feature Pairs		*Least Co-occurring Feature Pairs*	
Feature Pair	NPMI	Feature Pair	NPMI
NOT_character_person & NOT_plot_sequence	0.5	NOT_plot_sequence & event_experience	-0.31
character_person & event_experience	0.43	NOT_plot_sequence & character_person	-0.28
cohesive_interpretable & plot_sequence	0.4	NOT_plot_sequence & setting_background	-0.22
NOT_character_person & NOT_setting_background	0.38	event_experience & opinion	-0.18
event_experience & plot_sequence	0.37	concept_definition & event_experience	-0.18
NOT_character_person & NOT_event_experience	0.35	NOT_cohesive_interpretable & character_person	-0.17
NOT_plot_sequence & NOT_problem_conflict	0.32	NOT_plot_sequence & behavior_strategy	-0.15
NOT_plot_sequence & NOT_setting_background	0.29	NOT_character_person & behavior_strategy	-0.14
NOT_cohesive_interpretable & NOT_plot_sequence	0.27	behavior_strategy & plot_sequence	-0.14
character_person & plot_sequence	0.26	NOT_cohesive_interpretable & event_experience	-0.14

Table 7: Most (left) and least (right) co-occuring pairs of features, irrespective of the types of features in the pair. Scores can range from -1 (features never co-occur) to 0 (features are independent) to 1 (features always co-occur). We filter out pairs that occur less than 20 times.

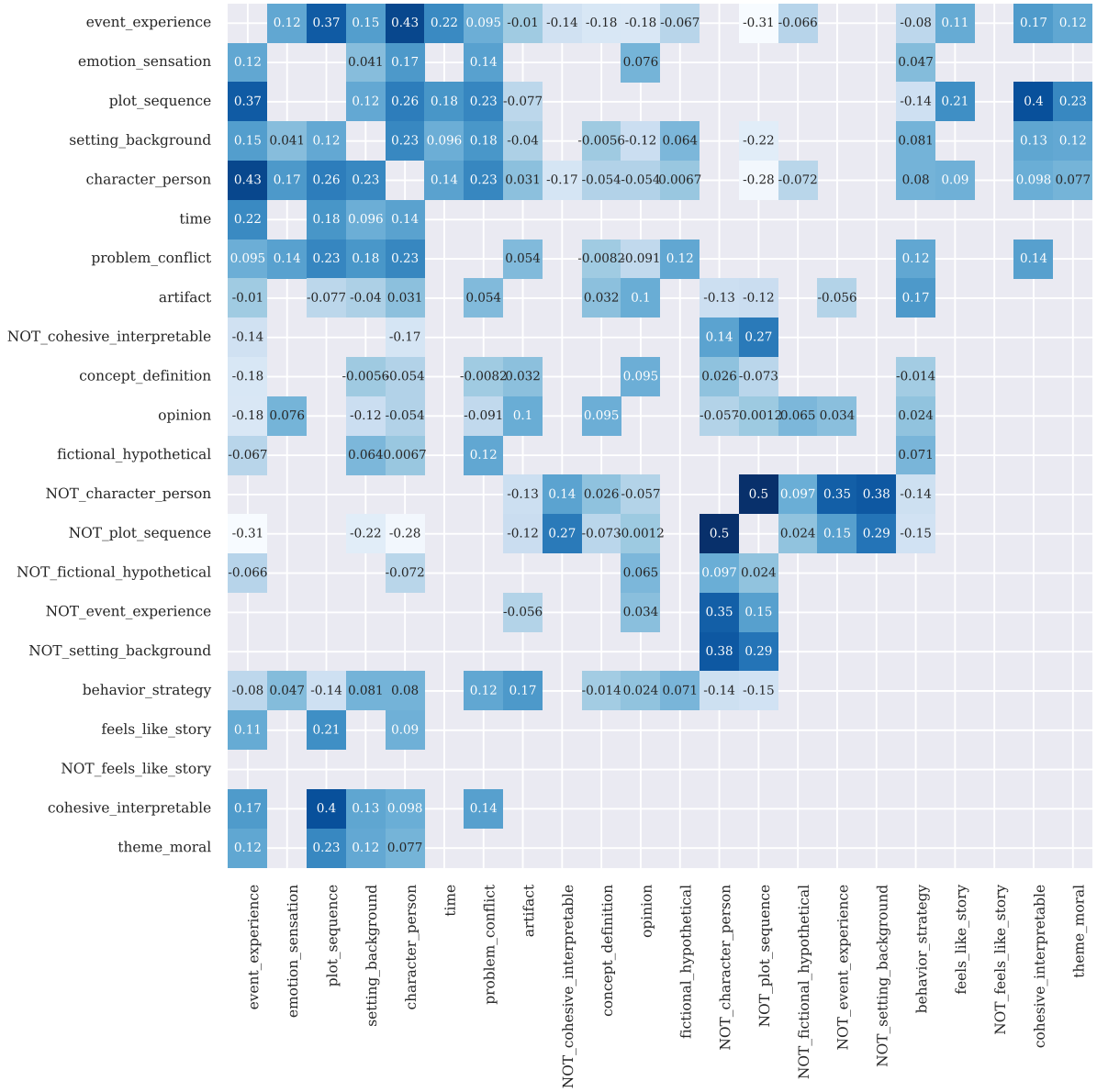

Figure 9: Feature co-occurrence metrics in story `rationales`, using normalized pointwise mutual information (NPMI). Scores can range from -1 (never co-occur) to 0 (independent) to 1 (always co-occur). We consider only those features that appear at least 40 times, and we display co-occurrence ratings for feature pairs that occur at least 20 times.

F.5 Classifier Variation

The primary purpose for story-annotated datasets is to train and evaluate story detection systems. We therefore compare the ranks and rates of story predictions by finetuning RoBERTa (Liu et al., 2019) models using either the crowd majority labels or the prescriptive consensus labels from researchers Antoniak et al. (2024).[15]

Examining Fig. 10, which shows the predicted story rates per subreddit, we observe that the researcher-finetuned and crowd-finetuned models are correlated (Pearson $r = 0.88$, $p < 0.05$), and that the researcher-finetuned model consistently predicts higher rates of storytelling.

Notably, we observe that story prediction rates across models (and, by extension, annotators) are not uniformly distributed across topics. Predicted storytelling rates are quite aligned for subreddits that have

[15]We use the RobertaForSequenceClassification pre-trained model with the 125M parameter roberta-base model from Hugging Face. Our hyperparameter settings are as follows: 3 epochs, a batch size of 16, a learning rate of 5e-5, 20 warmup steps for the learning rate scheduler, and a weight decay of 0.01.

extremely low storytelling rates (news and politics), as well as for subreddits that have high storytelling rates (stories, relationships). In contrast, in the 0.2 to 0.8 range, there is greater divergence in predicted storytelling rates. In particular, the researcher-finetuned model predicts much higher storytelling rates for the "tech" and "fandom" subreddits.

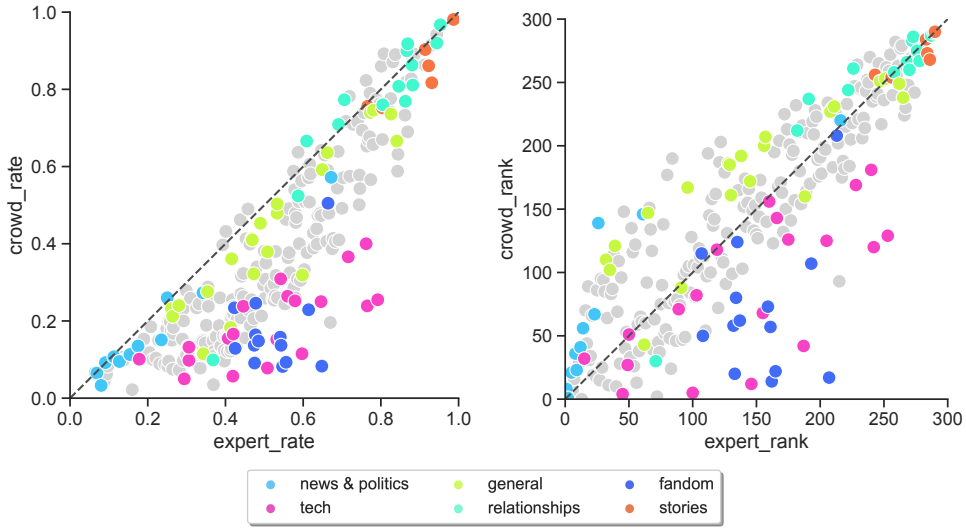

Figure 10: Comparison of story prediction rates between RoBERTa models fine-tuned on prescriptive labels from researchers (experts) vs. descriptive crowd majority vote labels across many subreddits, colored by the assigned subreddit category from StorySeeker.

F.6 Correlation of Automatic Story Features Across Annotators

To shed light onto textual features associated with labeling of stories vs. not, we replicate the feature analyses of Antoniak et al. (2024) using prescriptive labels from researchers, descriptive labels from crowd workers, and descriptive GPT-4 (gpt-4-0613) predictions.

In the StorySeeker dataset that consists of the same texts used in the StoryPerceptions, each text is scored for a number of textual features that are either prominent in prior work and/or are relevant to the prescriptive annotation codebook: entity and pronouns (Eisenberg and Finlayson, 2017; Piper and Bagga, 2022), events (Hühn, 2009; Gius and Vauth, 2022; Sims et al., 2019), verb tense, and concreteness (Piper and Bagga, 2022; Brysbaert et al., 2014). Excepting the event rate metrics, which are based on annotations and a BERT-based event tagging model (Sims et al., 2019; Sap et al., 2022a), most of the other metrics in the StorySeeker corpus are derived from the spaCy NER and POS taggers or lexica.

Antoniak et al. (2024) split texts into story and non-story groups based on the prescriptive consensus label from researchers, then run t-tests to identify which features are significantly positively or negatively associated with stories. For comparison purposes, we run t-tests on the features based on the crowd majority vote and (no-codebook) GPT-4 labels. Table 8 reports these results alongside Antoniak et al.'s (2024) original results for their labels.

Generally, the automatic features from the StorySeeker paper correlate with story labels, regardless of the annotation context. However, the effect sizes point to some notable differences. For instance, prescriptive story labels from researchers are associated with more events; crowd and GPT-4 labels are associated with features indicative of characters (entity rate, third-person singular pronouns) and concreteness relative to the prescriptive labels. Increased dependence on concreteness may point to a more constrained notion of action/events that does not consider certain cognitive/emotional activity or shifts as constituting events or plot in the same way that grounded physical action is perceived as eventful.

Overall, the shared trends across these automatic features strengthen our confidence in these features while also highlighting that teasing apart feature-level insights across annotation contexts requires studying different sets of features, which we address through our fine-grained coding of free-text responses from crowd workers.

Feature	d	Dir	p-val	d	Dir	p-val	d	Dir	p-val
	Researchers (Prescriptive)			*Crowd Majority (Descriptive)*			*GPT-4 (Descriptive)*		
first_person_singular	1.009***	story	0.0	0.799***	story	0.0	0.874***	story	0.0
first_person_plural	0.147	non-story	0.106	0.07	story	0.461	0.118	story	0.291
second_person	0.444***	non-story	0.0	0.555***	non-story	0.0	0.482***	non-story	0.0
third_singular	0.397***	story	0.0	0.544***	story	0.0	0.629***	story	0.0
entity_rate	0.285**	story	0.006	0.345***	story	0.001	0.467***	story	0.0
realis_event_rate	1.429***	story	0.0	1.225***	story	0.0	1.2***	story	0.0
union_event_rate	1.899***	story	0.0	1.507***	story	0.0	1.416***	story	0.0
past_tense_verb_rate	1.408***	story	0.0	1.343***	story	0.0	1.149***	story	0.0
not_past_tense_verb_rate	0.947***	non-story	0.0	0.88***	non-story	0.0	0.579***	non-story	0.0
concreteness	0.439***	story	0.0	0.595***	story	0.0	0.504***	story	0.0
is_comment	0.612***	non-story	0.0	0.332**	non-story	0.002	0.5***	non-story	0.0
text_length	0.174	story	0.106	0.257*	story	0.018	0.131	story	0.291
avg_sentence_length	0.259*	non-story	0.012	0.139	non-story	0.268	0.309**	non-story	0.002

Table 8: Results of t-tests comparing features between texts labeled as containing stories vs. not containing stories according to multiple different annotator contexts (prescriptive labels from researchers, descriptive labels from crowd workers, descriptive predictions from GPT-4). We control for multiple comparisons per annotator type using the Holm method (***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).