# NARRABENCH:
# A Comprehensive Framework for Narrative Benchmarking

**Sil Hamilton**
Cornell University

**Matthew Wilkens**
Cornell University

**Andrew Piper**
McGill University

## Abstract

We present NARRABENCH, a theory-informed taxonomy of narrative-understanding tasks, as well as an associated survey of 78 existing benchmarks in the area. We find significant need for new evaluations covering aspects of narrative understanding that are either overlooked in current work or are poorly aligned with existing metrics. Specifically, we estimate that only 27% of narrative tasks are well captured by existing benchmarks, and we note that some areas – including narrative events, style, perspective, and revelation – are nearly absent from current evaluations. We also note the need for increased development of benchmarks capable of assessing constitutively subjective and perspectival aspects of narrative, that is, aspects for which there is generally no single correct answer. Our taxonomy, survey, and methodology are of value to NLP researchers seeking to test LLM narrative understanding.

## 1 Introduction

Narratives are omnipresent in daily life. They can be used to entertain, inform, persuade, and maintain shared beliefs in communities across generations. Designing large language models (LLMs) that can both understand and generate narrative communication is of utmost importance to building artificially intelligent systems useful for humans.

Using narratives to test long context understanding and models' theory of mind has come into vogue of late (Kim et al., 2023; Karpinska et al., 2024; Chen et al., 2024b; Wang et al., 2025a), and non-narrative-specific benchmarks can test for narrative understanding in unexpected ways: "Needle in a haystack" tests show models can discern what is relevant to the present contextual frame (Kamradt, 2025), and summarization tasks require models to integrate information across long contexts to identify important through-lines (Lin, 2004).

Several large-scale benchmarks contain a significant amount of narrative content in this way. For
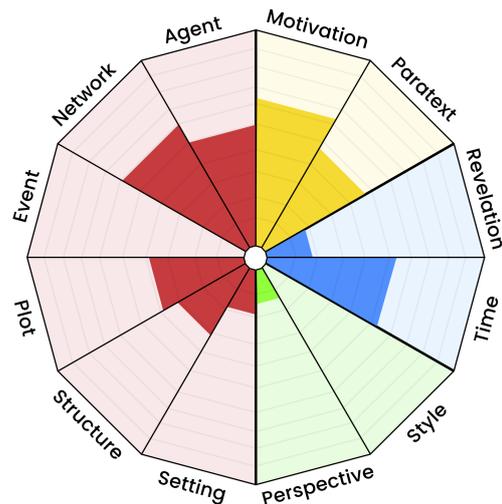


Figure 1: The twelve primary narrative features of the NARRABENCH taxonomy coloured according to the Big-4 narrative dimensions (story, narration, discourse, and situatedness) and shaded by how well existing benchmarks match as determined by our survey.

example, we estimate approximately 1 in 7 BIG-bench questions contains a story (Srivastava et al., 2023), 1 in 4 in HellaSwag (Zellers et al., 2019), and 1 in 3 in MMLU (Hendrycks et al., 2021).[1] Correctly answering these questions indicates some level of narrative understanding.

Nevertheless, benchmarks that *explicitly* and *systematically* test for narrative understanding are rare and often lack theoretical consistency in their principal goals and methodological foundations. Existing benchmarks often deploy single-answer frameworks, overlooking the way narrative understanding is a fundamentally interpretive (i.e., perspectival, not deterministic) task. Benchmarks likewise tend to focus on story-level content, missing more complex dimensions of narrative communication (e.g. discourse structure and narrative perspective).

In this paper, we propose a novel comprehen-

---

[1] All datasets were classified on a per-document basis with the BERT-based classifier from Antoniak et al. (2024).

sive benchmarking framework designed to assess LLMs' capacity for narrative understanding. Our framework incorporates relevant existing benchmarks and identifies areas for future development under a single, theoretically coherent umbrella. Our core contributions are:

**Survey of existing benchmarks for narrative understanding.** We review 78 existing benchmarks for their relevance for the task of narrative understanding. Of these, we find that 39 lack available data, while the remaining 39 offer reasonable or good fit with key aspects of narrative understanding. Of the 39 plausible and available benchmarks, 32 still fail to align exactly with narrative theory. As can be seen in Figure 1, existing benchmarks cover approximately 27% of our taxonomy. We highlight the consistent focus on deterministic, story-level understanding tasks as a major limitation to existing benchmark design. We finally describe a method for implementing benchmarks in a unified testing harness to serve as a reference implementation of the NARRABENCH framework.

**A novel taxonomy of narrative understanding tasks.** We develop a taxonomy of fifty distinct narrative understanding tasks derived from well-established theoretical frameworks in narratology and state of the art benchmarking theory. This taxonomy provides the first systematic integration of narrative understanding benchmarks within a unified theoretical framework that allows for future expansion. It also foregrounds the importance of perspectival alignment in benchmark development.

**Charting a path forward.** We highlight where existing benchmarks fit within this new taxonomy and where there are areas for future work. Our taxonomy provides a usable, expandable road-map for the development of new benchmarks and benchmarking data to create a robust resource for assessing LLM performance on a central aspect of human communication and cultural behaviour.

## 2 Related Work

**Narrative Understanding.** Early computational work modelled narratives as stochastic processes (Kahn, 1973) or scripts known as *Fillmorean frames* (Schank and Abelson, 1975) containing chronological chains of events composed of entities and their actions (Chambers and Jurafsky, 2008, 2009, 2010; Balasubramanian et al., 2013; Reiter et al., 2014). Advancements in NLP led

to increasingly sophisticated narrative extraction methods that made use of topic models, manually derived features, and associated rule mining (Antoniak et al., 2019; Belyy and Van Durme, 2020; Lyu et al., 2021; Zhang et al., 2023). This approach was extended to plot arc extraction, newly reinterpreted as sentiment over time (Hogan, 2011; Reagan et al., 2016; Somasundaran et al., 2020; Elkins, 2022; Knight et al., 2024), and social network analysis, where entities and their actions were rendered as graphs (Roemmele, 2019; Sims and Bamman, 2020; Tangherlini et al., 2020). Further work has focused on higher-level understanding including narrativity detection (Antoniak et al., 2025; Piper and Bagga, 2025) and narrative intent understanding (Zhu et al., 2023). As LLMs are increasingly used to assess and generate narratives, now is an optimal time to codify the diverse array of tasks that fall under the heading of narrative understanding.

**Advances and limitations in current benchmarking.** Benchmarks are the principal infrastructure of empirical NLP, enabling standardized comparison and shared progress measures across models. Of particular interest has been benchmarks for evaluating long-context understanding (Chang et al., 2024; Karpinska et al., 2024; Wang et al., 2025a; Thai and Iyyer, 2025) and summarization performance (Fabbri et al., 2021; Zhao et al., 2022; Subbiah et al., 2025). A smaller set of benchmarks focus on explicitly narrative phenomena— including dialogue (Vishnubhotla et al., 2022), plot arcs (Chun, 2021), or genre classification (Worsham and Kalita, 2018). General benchmarks for narrative understanding remain scarce. Evaluation methods also remain limited; most rely on multiple-choice questions for efficiency — increasingly seen as unreliable estimates of model comprehension (Sap et al., 2019; Wang et al., 2024; Balepur et al., 2024; Alzahrani et al., 2024; Wang et al., 2025b). Recent alternatives such as LLM-as-a-Judge evaluation (Li et al., 2024; Tan et al., 2025; Wang et al., 2025c) and distributional or multi-annotator scoring (Ni et al., 2024; Meister et al., 2024) point toward richer, less rigid evaluation pipelines, but have not yet been applied to narrative domains.

## 3 Theoretical Foundations

Narrative theory has developed, over more than a century, increasingly sophisticated qualitative descriptions of narratives across different mediums and contextual settings.

## 3.1 The Elements of Narrative

Our base definition of narrative communication is drawn from Piper et al. (2021). At its most elementary level, a narrative can be said to occur when all of the following criteria are met:

| | |
|---|---|
| A | Someone |
| B | tells |
| C | someone |
| D | somewhere |
| E | for some reason |
| | *that* |
| F | someone |
| G | did something(s) |
| H | [to/with someone] |
| I | somewhere |
| J | at some time |
| K | for some reason. |

For there to be a narrative, we need (A) a teller, (B) a mode of telling (i.e., medium), (C) a recipient, (D) a social situation, (E) a motivation or intent for telling the story, (F) an agent, (G) at least one action or event, (H) a possible object, (I) a location, (J) a time-frame, and (K) a motivation or cause of the actions involved. Narratologists distinguish between the frame of the storyworld (the elements that come after the double lines above) known as "diegetic" elements, and the frame of telling (those elements that come before the double lines) known as "heterodiegetic" elements, where diegesis refers to a narrative "frame" or "world." These dimensions need not be explicit, but they do need to be implied for a coherent narrative to form.

Narrative theory has additionally foregrounded the centrality of change or conflict as an elementary component of narrative communication (Herman, 2009), through concepts like "change of state" (Prince, 2012), "canonicity/breach" (Bruner, 1991), denouement (Freytag, 1895), or Aristotle's foundational emphasis in the *Poetics* (c. 330 BCE) on beginning, middle, and end structure. At a root level, narratives foreground *temporal difference* with respect to agent-centred events (Genette, 1980; Sternberg, 1992; Ricoeur, 2012).

### 3.2 Higher-Order Narrative Dimensions

To integrate these elements into a broader framework, we combine Genette's narrative triangle with Herman's notion of *situatedness* (Figure 2). The Russian formalists (Tomashevsky, 1965) were the first to distinguish between *story* ("what hap-
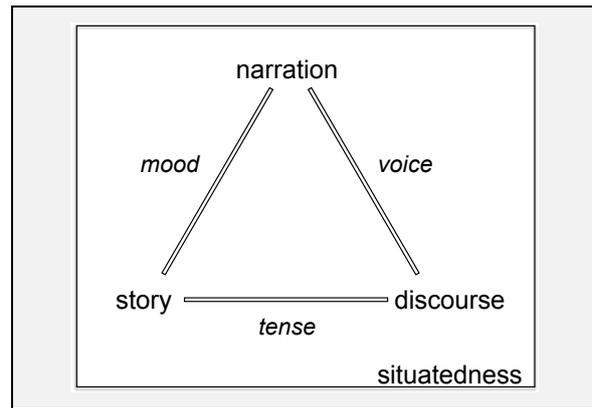


Figure 2: NARRABENCH's primary theoretical foundations from Genette (1980) and Herman (2009).

pened") and *discourse* ("how it was told"), to which Genette (1980) added a third dimension called *narration*, which refers to the perspectival dimensions of narrative communication produced by the narrator's voice ("who speaks").

Genette then introduced three further terms to capture the relationship between these dimensions, *tense*, *mood*, and *voice*. Genette extrapolates from the linguistic meanings of these terms to capture specific narratological features. These include aspects of time and the ordering of events (*tense*); the relationship between eventfulness and description (*mood*); and aspects related to perspective, such as point of view, dialogue, and focalization (*voice*). Herman (2009) importantly adds a fourth dimension, *situatedness*, which captures the social dimensions of narrative (medium, social context, interactive dimensions). We combine these elementary and higher-level dimensions to structure our taxonomy described in the next section.

## 4 The NARRABENCH Taxonomy

Our taxonomy focuses on two core levels of narrative understanding. The first is a hierarchically arranged set of fifty **narrative aspects** that can be mapped to the twelve *narrative features* shown in Figure 1. We group these aspects under the four principal *narrative dimensions* depicted in Figure 2. These dimensions form the root of our taxonomy.

The second level of our taxonomy focuses on **evaluation criteria**. These include textual *scale* (local, global, meso), *mode* (discrete, progressive, holistic judgments), and the expected *variance* of potential answers (deterministic, consensus, perspectival). We present the taxonomy in its entirety in Table 4, and a condensed version in Table 1.

| Dimension | Feature | Aspect | SMV |
|---|---|---|---|
| **Story** | Agent | name | L D D |
| | | | G H C |
| | | role | G H P |
| | | attributes | L D D |
| | | | G H C |
| | | emotions | L D P |
| | | | G H P |
| | | motivation | L D P |
| | | | G P P |
| | Social Net | interaction | L D D |
| | | connections | G H D |
| | | relationship | G H C |
| | Event | event | L D D |
| | | | G D C |
| | | schema | G H C |
| | | causality | G P P |
| | Plot | topic | G H C |
| | | plot | G H P |
| | | plotline | G H C |
| | | moral | G H P |
| | | obstacle | G H P |
| | | conflict | G H P |
| | | archetype | G H C |
| | structure | plot arc | G P C |
| | Setting | setting | L D D |
| | | | G H C |
| | | location | L D D |
| | | | G D D |
| **Discourse** | Time | duration | L D D |
| | | | G P D |
| | | | G H D |
| | | order | G P D |
| | | | G H D |
| | Revelation | suspense | G P P |
| | | curiosity | G P P |
| | | surprise | G P P |
| **Narration** | Perspective | point of view | G D D |
| | | focalization | L D D |
| | | dialogue | L D D |
| | Style | allusion | L D P |
| | | figurative | L D P |
| | | imageability | L H P |
| | | complexity | L H P |
| | | evaluative | L D P |
| **Situatedness** | Paratext | genre | G H C |
| | | author | G D D |
| | | date | G D D |
| | | medium | G D D |
| | | platform | G D D |
| | Motivation | intent | G H P |

Table 1: The complete NARRABENCH taxonomy. We indicate **SMV** (Scale-Mode-Variance) with: L/G=local/global, D/H/P=discrete/holistic/progressive, D/C/P=deterministic/consensus/perspectival.

## 4.1 Narrative Aspects

### 4.1.1 STORY

Story captures all aspects of a narrative that relate to storyworld content. We pay attention to:

**Agents.** Narratives foreground the lived experiences of agentic entities (i.e., characters) (Fludernik, 2002). Agent detection and inventorying *character names* are an essential first step. Higher-level synthesis of characters around their narrative *roles* and *attributes* is necessary for understanding the functional position of characters within narratives (Propp, 1968). We also include theory-of-mind related tasks such as *narrative emotion* labelling and character *motivation* understanding. Interpreting agent intentionality and inner experience is a central component of narrative understanding (Zunshine, 2006; Mar, 2011; Oatley, 2016).

**Social Networks.** Agents in a narrative exist in a particular social configuration. Here we divide social networks into three core aspects: *interactions* (local social events), *connections* (who does each character know), and *relationships* (the social types of connections, including employment, family, civil, romantic, and friendship-based connections) (Labatut and Bost, 2019).

**Events.** Narratives are structured around agents who experience change over time. Such change arises through *action*, where an action and its corresponding effect together form a *causal frame*. While narratologists debate the precise boundaries of a frame within a story, it is broadly accepted that reconstructing the underlying chain of actions and their causal relations is central to recovering the narrative's event structure or *schemas* (Schank and Abelson, 1975; Chambers and Jurafsky, 2009).

**Plot.** According to Kukkonen (2014), "The term *plot* designates the ways in which the events and characters' actions in a story are arranged and how this arrangement in turn facilitates identification of their motivations and consequences." In a technical sense, plot sits between story and discourse, but for our purposes we consider it as part of the story dimension. Plot can thus be subdivided into *topics* (core thematic concerns), *plot lines* (sub-plots), key structural forces like narrative *obstacles* or *conflicts*, central value lessons (i.e., *morals*), as well as larger schematic structures like *archetypes*.

**Structure.** Kukkonen (2014) makes a core distinction between plot as a function of "global structure" (our primary definition above) and "progressive structuration," that is, aspects of plot that depend on the sequencing of events for their meaning. We reserve these progressive aspects for the *discourse* dimension, with the exception of *plot*

*arcs*, which capture structural elements related to events such as fortune (Reagan et al., 2016; Elkins, 2022; McAdams et al., 2001), reversal (Knight et al., 2024), and denouement (Freytag, 1895). Understanding the reliance or inability of LLMs to produce certain plot arc structures is important for shaping reader expectations.

**Setting.** The spatial distribution of perception and movement within narratives is a key aspect of defining audiences' cognitive orientation. Where narrative theory has introduced concepts such as the "chronotope" (Bakhtin, 2010) or "possible worlds" (Ryan et al., 2016), we use the more integrative term of *setting* to capture both the holistic understanding of narrative space and the particular location of a given narrative moment.

### 4.1.2 NARRATION

Narratives require narrators. According to Genette (1980), "who speaks" plays a central role in governing the selection and communication of narrative content, which will in turn impact how audiences interpret the narrative.

**Perspective.** Narrative perspective encompasses classical notions of *point-of-view* (technically referred to as homo- and heterodiegetic narrators, 1P/3P respectively) as well as the subtler category of *focalization*, a local sense of the eyes or mind through which certain events are perceived. We also include *dialogue* detection and speaker attribution here to capture aspects of direct speech that are central to narratives.

**Style.** We use style to capture rhetorical and intertextual aspects of narratives that bear on their meaning. *Allusion* refers to inter-textual references (Kristeva, 1980; Genette, 1997), while *figurative* language and *imageability* are concerned with the symbolic and/or perceptual dimensions of narrative communication. We also include *complexity* as a measure of reading level and social capital and *evaluative* language as a way of capturing narrative meta-awareness (Labov and Waletzky, 1997).

### 4.1.3 DISCOURSE

Where *narration* captures aspects of perspective and style, *discourse* reflects organizational choices made in *how* events are rendered.

**Time.** Narratives take place over time. But time is not necessarily represented in a chronological manner, nor is time represented in an equitable manner. Narrators may decide to dedicate greater fidelity to certain events, or reflect back on prior events or foreshadow future events. Following Genette (1980), we focus on both *duration* and *order* with respect to time.

**Revelation.** How information is revealed to the reader over the course of narrative time fundamentally shapes readers' affective state (Brewer and Lichtenstein, 1982). Here we focus on the classic tripartite scheme of reader *suspense*, *curiosity*, and *surprise* to capture the assessment of narrative revelation over time.

### 4.1.4 SITUATEDNESS

A narrative is always communicated in a real context. This social context inflects both authorial choice and audience expectations.

**Paratext.** What theorists call the *paratext* are those elements of a narrative that reflect its real-world context. This includes aspects like *genre*, *author*, *date*, and *medium*. These dimensions help inform reader interpretations of the material.

**Motivation.** The "why" behind the telling of a narrative is captured by the author's intent. Narratives may be told for a variety of reasons, but being able to reasonably infer authorial *intent* is an essential task of narrative understanding.

## 4.2 Evaluation Criteria

We divide our evaluation framework into three primary dimensions.

**Scale.** When assessing narrative understanding, it is important to identify the textual *scale* of a particular judgment. Some features, such as character interactions, are always *local* in nature, because they are located at a discrete point in the narrative. Others are *global*, such as character roles or plot summaries. These require abstractive reasoning from a narrative unit (the entire narrative or a meaningful sub-portion). We reserve the term *meso* for features that depend on reasoning around temporal sequences, such as duration or suspense.

**Mode.** Answers might take various forms (e.g., a string, an integer, or a boolean) depending on the measurement. We call this the *mode*. The mode of the answer refers to whether the value itself is *discrete* in nature (meaning it is a single, token-based value), *progressive* (a collection of values forming a time series), or *holistic* (a comprehensive

| Quality | Distance | Benchmark Count |
|---|---|---|
| Good | 0 | 10 |
| Decent | 1 | 14 |
| Poor | 2 | 10 |
| Bad (ignored) | 3 | 5 |
| Missing Data | - | 39 |
| *Total* | | 78 |

Table 2: Overall benchmark alignment distribution. Note we consider only those benchmarks with an edit distance of 2 or below.

answer in the form of a string written in natural language). The mode plays a part in determining what form of automatic evaluation is admissible.

**Variance.** This dimension captures assumptions about interpretive freedom. Some features may require single, correct answers, such as a character name. We label these *deterministic*. Other features are more open-ended, such as narrative emotion or narrative intent. We label these features *perspectival* when there is a distribution of possible answers that is also contextually dependent. We finally use *consensus* where there are clear central tendencies in a distribution of answers.

## 4.3 Taxonomy Construction

We identify fifty total permutations of evaluation criteria and narrative attributes that are well-founded in the theoretical literature and critical for narrative understanding tasks (Table 4). As we highlight in our conclusion, a core aspect of this framework is its expandability. We capture the foundation of each permutation with a single question, of possible use for model testing.

## 5 Benchmark Survey

The purpose of the survey is to adjudicate to what degree currently available LLM benchmarks satisfy the aspects of narrativity identified above. For our purposes, we define a benchmark to be an annotated dataset or testing technique made available for testing language model accuracy on some qualitative task. The benchmark should:

- Yield results useful for adjudicating model performance of one or more of our twelve *narrative features*.

- Have an existing repository available for providing code and data.

- Provide convenience functions (e.g., a script or software framework) for testing on arbitrary language models using a standard API.

| **Agent** | role<br>attributes<br>emotions<br><br>motivation | Crab, Ditto<br>AustenAlike, NoLiMa<br>CULEMO, EQ-Bench<br>ToMBench<br>EmotionBench<br>OpenToM<br>FANTOM<br>MOTIVEBENCH<br>CHARACTERBENCH |
|---|---|---|
| **Social Net** | connections<br>relationship | PhantomWiki<br>DialogBench |
| **Plot** | topic<br>plot<br>plotline<br>moral | RepLiQA<br>FlawedFictions<br>StorySumm<br>MoralBench<br>MORABLES |
| **Setting** | location | Dataset-GSS |
| **Revelation** | suspense | ConflictBank |
| **Perspective** | dialogue | SODA-EVAL, CoSER |
| **Time** | order | TRaVelER, ToT,<br>MAVEN-ERE, TRAM |
| **Motivation** | intent | Persuasive Pairs |
| **Paratext** | genre<br>author<br>date | BookCover30<br>VeriDark, REASONS<br>TimeAware |

Figure 3: Aligned open benchmarks for narrative understanding tasks identified in this survey.

- Serve as a metric for comparing the relative performance of different models. It should not expect a particular model family.

Following these guidelines, we surface 78 potential benchmarks published over the last twelve years via relevant keyword searches on Google Scholar. Filtering for open data reduces this count to 39 benchmarks. We then adjudicate how well each benchmark matches the respective evaluation attributes for each hierarchal attribute triple (e.g. local-discrete-deterministic), measuring the edit distance. From this value we assign alignment ratings with our taxonomy of "good," (a distance of 0), "decent," (1), "poor," (2), and finally "bad" (3), resulting in the distribution seen in Table 2. Retaining all benchmarks with a quality rating of "poor" and above leaves us with 34 benchmarks which we present in Figure 3 and with a complete rating breakdown in Table 5. We now highlight trends observed when conducting our survey.

**Increasing popularity.** Of the 39 benchmarks we surveyed, half were published in 2024 or 2025, with the majority in 2025. As LLM use increases, so too do benchmarking efforts in an attempt to provide metrics for questioning whether advancements in ML techniques are translating to performance

increases in downstream capabilities. This increasing popularity carries with it the risk of decreasing the signal-to-noise ratio in the benchmarking space. This risk highlights the need to build unified theoretical frameworks to guide these efforts.

**Moving beyond classification tasks.** Of the 39 benchmarks surveyed, 27 rely on classification or multi-label prediction, while 12 evaluate open-ended text generation. Causal LLMs are generative models. Evaluating their generated text therefore offers a direct window into their underlying biases and proficiencies, as has been increasingly adopted (Lucy and Bamman, 2021; Hicke et al., 2025). Expanding such generative benchmarking approaches would enable the study of model responses in more holistic and perspectival ways.

**An overemphasis on story.** We identified 19 benchmarks that address our story-specific features. Compared to benchmarks identified for narration (2), discourse (5), and situatedness (5), this suggests current benchmarking efforts overemphasize the story as a target. One of the principal aims of our theoretical framework is to foreground the complexity of narrative communication and the importance of narrative perspective, discourse structure, and social context for surfacing narrative meaning.

**Lack of event-specific benchmarking.** While we identified many existing annotated datasets capturing event-related tasks (Baker et al., 1998; Sims et al., 2019; Tang et al., 2021; Wang et al., 2022b), none were expressly equipped for benchmarking as outlined in our requirements. These annotated datasets, while valuable, still need to be converted to the criteria outlined in the beginning of section 5 to be considered benchmarks.

**Lack of style-specific benchmarking.** We did not identify any benchmarks for allusion detection, figurative language production, imageability, complexity, or evaluative language use in language models. Figurative use of language is an important dimension for understanding narrative meaning. Benchmarking how effectively large language models recognize this intentional use of language is an important aspect of understanding their general narratological abilities. We encourage researchers to draw on existing methodologies and datasets in the fields of stylometry and metaphor detection when designing benchmarks for these aspects.

**Lack of subjectivity in responses.** All but two of the benchmarks constructed their ground truth in a deterministic manner, meaning LLMs *must* must correctly select an answer from a set of possible responses. As we note in section 2, benchmark developers are now phasing this mode of judgment out in favour of allowing for a multiplicity of responses. This method is more accurate when considering ground truth is often based on annotations from multiple annotators, meaning ground truth begins as a distribution of responses. Non-deterministic responses likewise more proficiently capture the subjective quality of audience response, in that no two audience members will respond to a narrative in precisely the same way. The same expectation must be afforded to LLMs if their abilities are to be fairly judged (Plank, 2022; Pichler et al., 2025).

**Per-token responses are still needed.** Early statistical NLP techniques were often built on lexical corpora attributing per-token scores for various linguistic phenomena. This enabled algorithms for assessing the phenomena over particular substrings, important for researchers interested in identifying unexpected failure modes over text input Lin (2004). The convenience of passing in an entire text to a LLM and receiving back whole answers has motivated developers to prefer holistic/global assessments over discrete responses. This has the benefit of allowing language models the space to respond as they naturally would to a query of a similar nature, but evaluating per-token responses can elicit a more nuanced understanding of how language models represent complex semantics. We argue both approaches are important for narrative understanding and encourage benchmark developers to continue to value token-level benchmarking.

**Global languages are not well attended.** Four of the 39 benchmarks we shortlisted were multilingual, with the rest implementing their test in English. This gap leaves open the question of whether language models can properly represent narrative dimensions in non-English languages (and especially low-resource languages). We encourage benchmark developers to seek platforms that provide human assessments from global perspectives alongside multilingual narrative resources.

**Lack of open data for reproducibility.** We finally note only 39 of the original 78 narrative-aligned benchmarks we identified made their code and data available in a freely-accessible repository.

| Model | Attributes austenalike | Emotional culemo | Connections phantomwiki | Plotline storysumm | Order tot | Order tram | Order traveler | Mean |
|---|---|---|---|---|---|---|---|---|
| Gemma-3 12B | **3.2** | 41.0 | 17.8 | 40.6 | 3.8 | 67.0 | 54.0 | 32.5 |
| GPT-OSS 20B | 0.0 | **100.0** | 2.6 | **62.5** | 1.2 | 0.0 | 19.0 | 26.5 |
| Llama-3 8B | 2.1 | 28.5 | **23.0** | 55.2 | **31.8** | 59.5 | **61.0** | **37.3** |

Table 3: Results across all well-conforming benchmarks and models. Accuracy is reported as percentages.

While certain papers made their rationale known (e.g. their benchmark depends on in-copyright text), other papers offered links leading to dead webpages. Adhering to best practices in data management, i.e. the use of long-term repositories with stable DOIs, is an essential dimension for future benchmarking initiatives.

**Advancing benchmarks towards multimodal inputs.** Narratives exist in multiple modalities. Narratological research in recent decades has developed theoretical frameworks for describing how narratives behave in music, imagery, paintings, and movies (Fludernik, 2010). Approximately 5% of our surveyed benchmarks were either multimodal or non-textual in their inputs, testing vision language models (VLMs) on their visual understanding abilities. Future benchmarks will want to combine these approaches to work towards a framework for describing multimodal narrative concepts.

## 6 Testing Harness

NARRABENCH is a novel attempt at reconfiguring existing benchmarks to fulfill theoretical goals, and as such can be implemented as an expandable collection of relevant benchmarks. To guide future implementations we have constructed a testing harness containing the most suitable benchmarks identified in our study, showcasing sample scores from a series of popular open-weight LLMs in Table 3.[2] Our results suggest non-uniform performance across models and specific narrative aspects, a result not immediately obvious when making use of a specific benchmark as an indicator for story understanding. We encourage all interested to identify or develop further benchmarks in areas not yet covered as per our initial survey. Distributing a stable implementation is one important way NARRABENCH will remain an evolving and relevant resource moving into the future.

## 7 Conclusion

To help guide and centralize growing efforts in narrative benchmarking, we present NARRABENCH,

the first comprehensive narratological taxonomy for benchmark developers to integrate efforts around a shared, holistic understanding of narrative communication. NARRABENCH distills decades of literary theory down to four fundamental narrative dimensions: *story*, *narration*, *discourse* (or the structure of the narrative), and *situatedness* (the context in which the narrative exists). We further divide these dimensions into twelve primary *features* and fifty overall *aspects*, described in Table 4.

As a first step towards constructing our envisioned framework, we conducted a first-ever survey of existing narrative benchmarks to identify how contemporary benchmarking efforts align with the NARRABENCH taxonomy (Table 5). We identify 39 open-data benchmarks whose qualitative targets roughly match one of the fifty permutations defined in our taxonomy, suggesting existing benchmarks satisfy ≈27% of the NARRABENCH taxonomy as indicated in Figure 1. This gap suggests opportunities for researchers to produce new tasks.

Central to our effort is the notion of an *expandable* framework for narrative benchmarking. Our high-level Big-4 dimensions allow for growth at all three subsidiary levels. Like other large benchmarking efforts, NARRABENCH can be seen as a benchmark of benchmarks organized around a core theoretical foundation. We encourage researchers in the field to both fill in the missing holes identified by our survey (as shown in Table 5) and also propose novel features and assessments. Providing one theoretical description of narrative understanding in the form of NARRABENCH offers benchmark developers the opportunity to consult a central resource to ensure efforts do not replicate prior work and address core methodological shortcomings.

**Next steps.** Researchers are increasingly focusing on the value of narrative understanding and narrative generation. Our testing harness will allow community members to consult the current state of the art, and contribute where possible. Now is an optimal time to begin the work of consolidation to provide LLM researchers with clear assessment criteria around one of the most fundamental ways humans relate to each other: storytelling.

---

[2]We make our harness available here.

## Limitations

While NARRABENCH provides the first comprehensive taxonomy for narrative benchmarking, several limitations should be acknowledged. First, the framework is grounded in a particular lineage of narrative theory—what is known as the classical model—which privileges compositional and communicative dimensions of narrative (story, narration, discourse, situatedness). This excludes alternative conceptions of narrativity, such as those derived from cognitive narratology, rhetorical theory, or postclassical approaches emphasizing affect, ideology, or embodiment. As a result, NARRABENCH captures a specific, though widely applicable, sense of computational narrative understanding.

Second, the benchmark survey is constrained by the availability and accessibility of existing resources. Of the 78 benchmarks identified, only half provided open data or usable repositories. This reliance on public datasets may bias coverage toward English-language and Western-centric corpora, underrepresenting non-Western narrative traditions and low-resource languages. Likewise, many of the included benchmarks are text-based, limiting our ability to generalize conclusions about multimodal or cross-media narratives, which are increasingly central to human storytelling practices.

Third, NARRABENCH currently formalizes narrative understanding in a way that presupposes task modularity—that discrete components such as perspective, style, or revelation can be isolated and measured independently. While this approach enables systematic comparison, it underplays the interdependence and nonlinearity of narrative meaning-making. Future iterations of the framework could incorporate compositional or causal dependencies across dimensions, testing how models integrate multiple narrative features jointly rather than in isolation.

Fourth, because the taxonomy is conceptual rather than empirical, its coverage and weighting of dimensions remain interpretive. Our survey indicates that existing benchmarks cover roughly 27% of the proposed taxonomy, but the relative importance of each category for narrative comprehension has not yet been empirically validated. The framework should therefore be viewed as a scaffolding for community refinement rather than a definitive account of narrative intelligence.

Finally, despite its explicit focus on perspectival and consensus-based evaluation, NARRABENCH inherits many of the broader evaluation challenges facing LLM benchmarking: the instability of LLM-as-a-judge methods, limited reproducibility of generative scoring, and uncertainty about how human interpretive diversity should be represented in gold standards. Expanding community participation, diversifying annotation pipelines, and incorporating non-deterministic scoring schemes will be essential to properly realizing narrative benchmarking.

In sum, NARRABENCH represents a first synthesis—a structured, extensible foundation for assessing narrative understanding—while acknowledging that both its theoretical premises and empirical coverage are partial. Its value lies in providing a coherent starting point for an evolving research community to debate, test, and revise what narrative understanding should mean for NLP and ML.

## Ethical Considerations

Narrative is among the most powerful instruments of human communication. It can promote empathy, understanding, and collective meaning-making, but it can equally serve as a vehicle for misinformation, propaganda, and ideological manipulation. As large language models increasingly generate and interpret narratives at scale, the ethical implications of how these capacities are benchmarked become critical. A benchmark that rewards coherence or emotional resonance without attention to factuality, bias, or intent risks amplifying the very distortions that narrative can produce.

NARRABENCH is designed in part to mitigate these risks by making explicit the multidimensional nature of narrative understanding—its agents, events, perspectives, and social contexts. By foregrounding interpretive variance and perspectival alignment, the framework aims to encourage evaluations that account not only for textual quality but also for the moral and epistemic consequences of narrative production. Nevertheless, benchmarking itself is never neutral: the selection of narratives, the construction of evaluation criteria, and the choice of what counts as "understanding" all carry normative assumptions.

We therefore view narrative benchmarking as an ethical task of representation. Establishing shared standards for how models comprehend and reproduce stories is essential for preventing the misuse of narrative generation in contexts that shape public opinion, cultural memory, personal well-being, and political discourse. Getting narrative benchmark-

ing right means ensuring that models are evaluated not only for their formal proficiency but also for their capacity to reflect the diversity, accountability, and responsibility inherent in human storytelling.

## Acknowledgements

## References

Kabir Ahuja, Melanie Sclar, and Yulia Tsvetkov. 2025. Finding Flawed Fictions: Evaluating Complex Reasoning in Language Models via Plot Hole Detection. *Preprint*, arXiv:2504.11900.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where Do People Tell Stories Online? Story Detection Across Online Communities. *Preprint*, arXiv:2311.09675.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2025. Where Do People Tell Stories Online? Story Detection Across Online Communities. *Preprint*, arXiv:2311.09675.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Mikhail Mikhaĭlovich Bakhtin. 2010. *The dialogic imagination: Four essays*, volume 1. University of texas Press.

Niranjan Balasubramanian, Stephen Soderland, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural Lenses on Emotion – Benchmarking LLMs for Cross-Cultural Emotion Understanding. *Preprint*, arXiv:2503.10688.

Anton Belyy and Benjamin Van Durme. 2020. Script induction as association rule mining. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 55–62.

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

Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry*, 18(1):1–21.

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. *ACL-2008*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09*, volume 2, page 602, Suntec, Singapore. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2010. A Database of Narrative Schemas. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *Preprint*, arXiv:2310.00785.

Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024a. EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models. *Preprint*, arXiv:2409.13359.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. T MBENCH: Benchmarking Theory of Mind in Large Language Models.

Jon Chun. 2021. SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs. *Preprint*, arXiv:2110.09454.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*, 1 edition. Cambridge University Press.

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

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning. *Preprint*, arXiv:2406.09170.

Monika Fludernik. 2002. *Towards a 'Natural' Narratology*. Routledge.

Monika Fludernik, editor. 2010. *Postclassical Narratology: Approaches and Analyses*. Theory and Interpretation of Narrative. Ohio State University Press, Columbus.

Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.

Gérard Genette. 1980. *Narrative Discourse: An Essay in Method*. Cornell University Press, Ithaca, N.Y.

Gérard Genette. 1997. *Palimpsests: Literature in the second degree*, volume 8. U of Nebraska Press.

Albert Gong, Kamilė Stankevičiūtė, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, and Kilian Q. Weinberger. 2025. PhantomWiki: On-Demand Datasets for Reasoning and Retrieval Evaluation. *Preprint*, arXiv:2502.20377.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A Novel Configurable Role-Playing LLM with Assessing Benchmark.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Preprint*, arXiv:2009.03300.

David Herel, Vojtech Bartek, Jiri Jirak, and Tomas Mikolov. 2025. Time Awareness in Large Language Models: Benchmarking Fact Recall Across Time. *Preprint*, arXiv:2409.13338.

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

Rebecca M. M. Hicke, Sil Hamilton, and David Mimno. 2025. The Zero Body Problem: Probing LLM Use of Sensory Language. *Preprint*, arXiv:2504.06393.

Patrick Colm Hogan. 2011. *Affective Narratology: The Emotional Structure of Stories*. U of Nebraska Press.

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2017. Judging a Book By its Cover. *Preprint*, arXiv:1610.09204.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. MoralBench: Moral Evaluation of LLMs.

Edward Kahn. 1973. Finite-state models of plot complexity. *Poetics*, 3(1):5–20.

Greg Kamradt. 2025. Needle In A Haystack - Pressure Testing LLMs.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One Thousand and One Pairs: A "novel" challenge for long-context language models. *Preprint*, arXiv:2406.16264.

Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. 2024. Benchmarking the Ability of Large Language Models to Reason About Event Sets:. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 74–82, Porto, Portugal. SCITEPRESS - Science and Technology Publications.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. *Preprint*, arXiv:2310.15421.

Samsun Knight, Matthew D Rocklage, and Yakov Bart. 2024. Narrative reversals and story success. *Science Advances*, 10(34):eadl2013.

Julia Kristeva. 1980. *Desire in language: A semiotic approach to literature and art*. Columbia University Press.

Karin Kukkonen. 2014. Plot. In Peter Hühn and 1 others, editors, *The Living Handbook of Narratology*. Hamburg University / Interdisciplinary Center for Narratology. Viewed 25 January 2014, revised 24 March 2014.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yu-jia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *Preprint*, arXiv:2412.05579.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025. From Pixels to Places: A Systematic Benchmark for Evaluating Image Geolocalization Ability in Large Language Models. *Preprint*, arXiv:2508.01608.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. *Preprint*, arXiv:2401.12474.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Qing Lyu, Li Zhang, and Chris Callison-Burch. 2021. Goal-Oriented Script Construction. *Preprint*, arXiv:2107.13189.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web. *Preprint*, arXiv:2207.03477.

Raymond A Mar. 2011. The neural bases of social cognition and story comprehension. *Annual review of psychology*, 62(1):103–134.

Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2025. MORABLES: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables. *Preprint*, arXiv:2509.12371.

Dan P McAdams, Jeffrey Reynolds, Martha Lewis, Allison H Patten, and Phillip J Bowman. 2001. When bad things turn good and good things turn bad: Sequences of redemption and contamination in life narrative and their relation to psychosocial adaptation in midlife adults and in students. *Personality and social psychology bulletin*, 27(4):474–485.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking Distributional Alignment of Large Language Models. *Preprint*, arXiv:2411.05403.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024. Soda-Eval: Open-Domain Dialogue Evaluation in the age of LLMs. *Preprint*, arXiv:2408.10902.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. NoLiMa: Long-Context Evaluation Beyond Literal Matching. *Preprint*, arXiv:2502.05167.

João Monteiro, Pierre-André Noël, Étienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vázquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. 2024. REPLIQA: A Question-Answering Dataset for Benchmarking LLMs on Unseen Reference Content.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *Advances in Neural Information Processing Systems*, 37:98180–98212.

Keith Oatley. 2016. Fiction: Simulation of social worlds. *Trends in cognitive sciences*, 20(8):618–628.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as Human-like Dialogue Systems. *Preprint*, arXiv:2311.01677.

Samuel J. Paech. 2024. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. *Preprint*, arXiv:2312.06281.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. *Preprint*, arXiv:2406.17753.

Axel Pichler, Janis Pagel, and Nils Reiter. 2025. Evaluating LLM-prompting for sequence labeling tasks in computational literary studies. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46, Albuquerque, New Mexico. Association for Computational Linguistics.

Andrew Piper and Sunyam Bagga. 2025. NarraDetect: An annotated dataset for the task of narrative detection. In *Proceedings of the The 7th Workshop on Narrative Understanding*, pages 1–7, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Gerald Prince. 2012. *Narratology: The form and functioning of narrative*, volume 108. Walter de Gruyter.

Vladimir Iakovlevich Propp. 1968. *Morphology of the Folktale: Second Edition.* University of Texas Press.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Nils Reiter, Anette Frank, and Oliver Hellwig. 2014. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605.

Paul Ricoeur. 2012. *Time and Narrative, Volume 1.* University of Chicago press.

Melissa Roemmele. 2019. Identifying sensible lexical relations in generated stories. In *Proceedings of the First Workshop on Narrative Understanding*, pages 44–52.

Marie-Laure Ryan, Kenneth Foote, and Maoz Azaryahu. 2016. *Narrating space/spatializing narrative: Where narrative theory and geography meet.* The Ohio State University Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Yash Saxena, Deepa Tilwani, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. 2025. Attribution in Scientific Literature: New Benchmark and Methods. *Preprint*, arXiv:2405.02228.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans and goals. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence. IJCAI*, volume 1.

Matthew Sims and David Bamman. 2020. Measuring Information Propagation in Literary Social Networks. *Preprint*, arXiv:2004.13980.

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.

Swapna Somasundaran, Xianyang Chen, and Michael Flor. 2020. Emotion arcs of student narratives. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 97–107.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Meir Sternberg. 1992. Telling in time (ii): Chronology, teleology, narrativity. *Poetics Today*, 13(3):463–541.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLM.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia B. Chilton, and Kathleen McKeown. 2025. STORYSUMM: Evaluating Faithfulness in Story Summarization. *Preprint*, arXiv:2407.06501.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. JudgeBench: A Benchmark for Evaluating LLM-based Judges. *Preprint*, arXiv:2410.12784.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From Discourse to Narrative: Knowledge Projection for Event Relation Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742, Online. Association for Computational Linguistics.

Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 15(6):e0233879.

Katherine Thai and Mohit Iyyer. 2025. Literary Evidence Retrieval via Long-Context Language Models. *Preprint*, arXiv:2506.03090.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are Large Language Models Capable of Generating Human-Level Narratives? *Preprint*, arXiv:2407.13248.

Boris Tomashevsky. 1965. *Russian Formalist Criticism: Four Essays.* University of Nebraska Press.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts. *Preprint*, arXiv:2204.05836.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue

Zhang. 2025a. NovelQA: Benchmarking Question Answering on Documents Exceeding 200K Tokens. *Preprint*, arXiv:2403.12766.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025b. LLMs May Perform MCQA by Selecting the Least Incorrect Option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862, Abu Dhabi, UAE. Association for Computational Linguistics.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. 2025c. Improving LLM-as-a-Judge Inference with the Judgment Distribution. *Preprint*, arXiv:2503.03064.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022a. MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction. *Preprint*, arXiv:2211.07342.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025d. CoSER: Coordinating LLM-Based Persona Simulation of Established Roles. *Preprint*, arXiv:2502.09082.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking Temporal Reasoning for Large Language Models. *Preprint*, arXiv:2310.00835.

Zhilin Wang, Anna Jafarpour, and Maarten Sap. 2022b. Uncovering surprising event boundaries in narratives. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 1–12.

Joseph Worsham and Jugal Kalita. 2018. Genre Identification and the Compositional Effect of Genre in Literature.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *Preprint*, arXiv:2402.06044.

Funing Yang and Carolyn Jane Anderson. 2024. Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark. *Preprint*, arXiv:2408.16131.

Xixian Yong, Jianxun Lian, Xiaoyuan Yi, Xiao Zhou, and Xing Xie. 2025. MotiveBench: How Far Are We From Human-Like Motivational Reasoning in Large Language Models? *Preprint*, arXiv:2506.13065.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *Preprint*, arXiv:1905.07830.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Liyang Zhou, Hainiu Xu, Li Zhang, Lara J. Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Brown, Reece Suchocki, and Chris Callison-Burch. 2023. Human-in-the-Loop Schema Induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 1–10.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. NarraSum: A Large-Scale Dataset for Abstractive Narrative Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2024. CharacterBench: Benchmarking Character Customization of Large Language Models.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP Models Good at Tracing Thoughts: An Overview of Narrative Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

Lisa Zunshine. 2006. *Why we read fiction: Theory of mind and the novel*. Ohio State University Press.

# A  Complete Taxonomy

We present the complete NARRABENCH taxonomy together with example questions in Table 4.

# B  Survey Results

We present our full survey results in Table 5.

| Dimension | Feature | Aspect | Scale | Mode | Variance | Question |
|---|---|---|---|---|---|---|
| **Story** | | name | local | discrete | deterministic | *Who are the characters in the text?* |
| | | | global | holistic | consensus | *Who are the main characters in the text?* |
| | | role | global | holistic | perspectival | *What is the character's role in the text?* |
| | Agent | attributes | local | discrete | deterministic | *What attributes does this character have?* |
| | | | global | holistic | consensus | *What attributes does this character have?* |
| | | emotions | local | discrete | perspectival | *What is the character feeling right now?* |
| | | | global | holistic | perspectival | *What are the central emotional states?* |
| | | motivation | local | discrete | perspectival | *Why is the character doing this right now?* |
| | | | global | progressive | perspectival | *What motivates this character?* |
| | | interaction | local | discrete | deterministic | *How are these two characters interacting?* |
| | Social Net | connections | global | holistic | deterministic | *Who does the character know?* |
| | | relationship | global | holistic | consensus | *What is the relationship type?* |
| | | event | local | discrete | deterministic | *What is happening?* |
| | Event | | global | discrete | consensus | *What happened?* |
| | | schema | global | holistic | consensus | *What is the narrative schema?* |
| | | causality | global | progressive | perspectival | *What caused this event?* |
| | | topic | global | holistic | consensus | *What are the topics of this story?* |
| | | plot | global | holistic | perspectival | *What is the plot summary?* |
| | | plotline | global | holistic | consensus | *What happened in this plotline?* |
| | Plot | moral | global | holistic | perspectival | *What is the moral of the story?* |
| | | obstacle | global | holistic | perspectival | *What is the central negative force?* |
| | | conflict | global | holistic | perspectival | *What is the central conflict?* |
| | | archetype | global | holistic | consensus | *What is the narrative archetype?* |
| | structure | plot arc | global | progressive | consensus | *What is the plot arc structure?* |
| | | setting | local | discrete | deterministic | *What is the setting?* |
| | Setting | | global | holistic | consensus | *What is the setting?* |
| | | location | local | discrete | deterministic | *Where is this taking place?* |
| | | | global | discrete | deterministic | *What locations has the story visited?* |
| **Discourse** | | duration | local | discrete | deterministic | *How much time is passing?* |
| | | | global | progressive | deterministic | *How much time since the previous scene?* |
| | Time | | global | holistic | deterministic | *How much time has passed?* |
| | | order | global | progressive | deterministic | *Does this scene come before or after?* |
| | | | global | holistic | deterministic | *Does this story tell events out of order?* |
| | | suspense | global | progressive | perspectival | *Is key information being withheld?* |
| | Revelation | curiosity | global | progressive | perspectival | *Are causal antecedents being withheld?* |
| | | surprise | global | progressive | perspectival | *Is key information suddenly revealed?* |
| **Narration** | | point of view | global | discrete | deterministic | *Who is telling? (1P, 2P, 3P)* |
| | Perspective | focalization | local | discrete | deterministic | *From whose POV are we seeing events?* |
| | | dialogue | local | discrete | deterministic | *Who speaks? Identify speakers.* |
| | | allusion | local | discrete | perspectival | *What texts is this alluding to?* |
| | | figurative | local | discrete | perspectival | *Is this using figurative language?* |
| | Style | imageability | local | holistic | perspectival | *How well can you imagine this scene?* |
| | | complexity | local | holistic | perspectival | *How complex is the sentence structure?* |
| | | evaluative | local | discrete | perspectival | *Is this engaging in evaluative discourse?* |
| **Situatedness** | | genre | global | holistic | consensus | *What is the genre?* |
| | | author | global | discrete | deterministic | *Who is/are the author(s)?* |
| | Paratext | date | global | discrete | deterministic | *What is the date of creation/publication?* |
| | | medium | global | discrete | deterministic | *What medium?* |
| | | platform | global | discrete | deterministic | *What platform?* |
| | Motivation | intent | global | holistic | perspectival | *What is the author's intent?* |

Table 4: The complete NARRABENCH taxonomy of narrative understanding tasks, systematically derived from narratological theory. Each task is defined by six attributes: dimension, feature, aspect, scale, mode, and variance.

3800

| Feature | Aspect | Benchmark | Scale | Mode | Variance | Citation |
|---|---|---|---|---|---|---|
| **Agent** | name | — | — | — | — | — |
| | role | Crab | global | holistic | consensus (P) | He et al. (2025) |
| | | Ditto | global | holistic | deterministic (P) | Lu et al. (2024) |
| | attributes | AustenAlike | global | holistic | consensus | Yang and Anderson (2024) |
| | | NoLiMa | global | holistic | deterministic (C) | Modarressi et al. (2025) |
| | emotions | CULEMO | global | holistic | perspectival | Belay et al. (2025) |
| | | EQ-Bench | global | holistic | deterministic (P) | Paech (2024) |
| | | ToMBench | global | holistic | deterministic (P) | Chen et al. (2024b) |
| | | EmotionBench | global | holistic | deterministic (P) | Chen et al. (2024a) |
| | motivation | OpenToM | local | progressive (D) | deterministic (P) | Xu et al. (2024) |
| | | FANTOM | global | holistic (P) | deterministic (P) | Kim et al. (2023) |
| | | MOTIVEBENCH | global | holistic (P) | deterministic (P) | Yong et al. (2025) |
| | | CHARACTERBENCH | global | holistic (P) | deterministic (P) | Zhou et al. (2024) |
| **Social** | interaction type | — | — | — | — | — |
| | connections | PhantomWiki | global | holistic | deterministic | Gong et al. (2025) |
| | relationship | DialogBench | global | holistic | deterministic (P) | Ou et al. (2024) |
| **Event** | causality | — | — | — | — | — |
| | event | — | — | — | — | — |
| | schema | — | — | — | — | — |
| **Plot** | topic | RepLiQA | global | holistic | deterministic (C) | Monteiro et al. (2024) |
| | plot | FlawedFictions | global | discrete (H) | deterministic (P) | Ahuja et al. (2025) |
| | plotline | StorySumm | global | holistic | consensus | Subbiah et al. (2025) |
| | moral | MoralBench | global | holistic | deterministic (P) | Ji et al. (2024) |
| | | MORABLES | global | holistic | deterministic (P) | Marcuzzo et al. (2025) |
| | obstacle | — | — | — | — | — |
| | conflict | — | — | — | — | — |
| | archetype | — | — | — | — | — |
| **Setting** | setting | Dataset-GSS | global | discrete (H) | deterministic (C) | Li et al. (2025) |
| | location | — | — | — | — | — |
| **Structure** | plot arc | EventStoryLine | global | holistic (P) | deterministic (C) | Caselli and Vossen (2017) |
| | | Narrative-Discourse | global | holistic (P) | deterministic (C) | Tian et al. (2024) |
| **Perspective** | dialogue | SODA-EVAL | global (L) | holistic (H) | deterministic | Mendonça et al. (2024) |
| | | CoSER | global (L) | holistic (H) | deterministic | Wang et al. (2025d) |
| | point of view | — | — | — | — | — |
| | focalization | — | — | — | — | — |
| **Style** | allusion | — | — | — | — | — |
| | figurative lang. | — | — | — | — | — |
| | imageability | — | — | — | — | — |
| | complexity | — | — | — | — | — |
| | evaluative lang. | — | — | — | — | — |
| **Time** | duration | — | — | — | — | — |
| | order | TRaVelER | global | holistic | deterministic | Kenneweg et al. (2024) |
| | | ToT | global | holistic | deterministic | Fatemi et al. (2024) |
| | | MAVEN-ERE | global | holistic | deterministic | Wang et al. (2022a) |
| | | TRAM | global | holistic | deterministic | Wang and Zhao (2024) |
| **Revelation** | suspense | ConflictBank | global | holistic (P) | deterministic (P) | Su et al. (2024) |
| | curiosity | — | — | — | — | — |
| | surprise | — | — | — | — | — |
| **Paratext** | genre | BookCover30 | global | discrete (H) | deterministic (C) | Iwana et al. (2017) |
| | author | VeriDark | global | holistic (D) | deterministic | Manolache et al. (2022) |
| | | REASONS | global | discrete | deterministic | Saxena et al. (2025) |
| | date | TimeAware | global | discrete | deterministic | Herel et al. (2025) |
| | medium | — | — | — | — | — |
| | platform | — | — | — | — | — |
| **Motivation** | intent | Persuasive Pairs | global | holistic | deterministic (P) | Pauli et al. (2025) |

Table 5: Benchmark survey results showing existing benchmarks mapped to NARRABENCH taxonomy attributes. Green indicates match with desired criteria, red indicates mismatch, with initials in brackets indicating the preferred property according to our taxonomy.