Classifying Unreliable Narrators with Large Language Models

Anneliese Brei[1*] **Katharine Henry**[1†] **Abhisheik Sharma**[2*]
Shashank Srivastava[1*] **Snigdha Chaturvedi**[1*]
[1]UNC Chapel Hill [2]Virginia Polytechnic Institute and State University
abrei@cs.unc.edu, katharinehenry@alumni.unc.edu, abhisharma@vt.edu,
{ssrivastava, snigdha} @cs.unc.edu

Abstract

Often when we interact with a first-person account of events, we consider whether or not the narrator, the primary speaker of the text, is reliable. In this paper, we propose using computational methods to identify unreliable narrators, i.e. those who unintentionally misrepresent information. Borrowing literary theory from narratology to define different types of unreliable narrators based on a variety of textual phenomena, we present TUNA, a human-annotated dataset of narratives from multiple domains, including blog posts, subreddit posts, hotel reviews, and works of literature. We define classification tasks for intra-narrational, inter-narrational, and inter-textual unreliabilities and analyze the performance of popular open-weight and proprietary LLMs for each. We propose learning from literature to perform unreliable narrator classification on real-world text data. To this end, we experiment with few-shot, fine-tuning, and curriculum learning settings. Our results show that this task is very challenging, and there is potential for using LLMs to identify unreliable narrators. We release our expert-annotated dataset and code at https://github.com/adbrei/unreliable-narrators and invite future research in this area.

1 Introduction

Imagine that you are on social media warning your friends about a recent shopping experience, and before submitting the post, you wonder if the presentation of your writing undermines your credibility. In another window, you are writing a cover letter. You recount a critical learning experience from your past job and wonder if your present voice sounds reliable to the reader. In the next room, your family is discussing the debate transcript between political candidates. Your sister thinks one of the candidates speaks like a villain from a novel she read, and your family differs on how reliable the candidate actually is. For each of these situations, it would be useful to have an automatic tool that identifies unreliability.

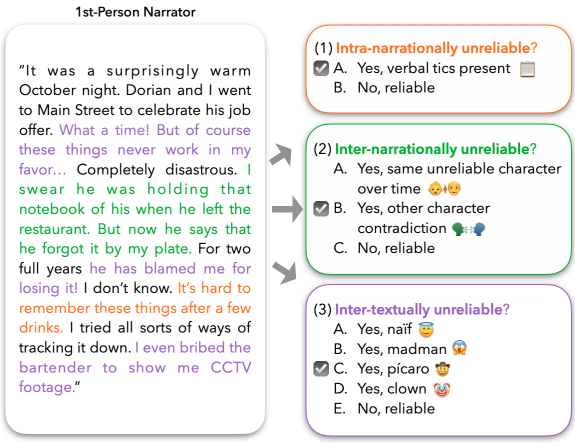

Figure 1: Real-world text with first-person narrators, such as the narrative shown (left), can be analyzed to determine the unreliability of the narrator. We separately classify three types of unreliability (right): intra-narrational, inter-narrational, and inter-textual.

Readers of personal accounts, such as reviews, online comments, cover letters, and college application essays, often implicitly question the reliability of the narrator: *Can I trust how this narrator has perceived and is describing the event?* Meanwhile, writers who wish to defend their points are concerned about how they textualize their ideas: *Am I sharing information in a reliable way?* Answering such questions is critical for the safe transmission of information (Nünning, 2015).

However, answering these questions is not a simple task. That is because unreliability cues are often subtle and context-dependent (Hansen, 2007). They might be scattered across the text or involve a deeper understanding beyond what is explicitly stated. Sometimes it is necessary to draw abstract inferences about the emotional and mental state of the narrator. Also, a text might have many readers,

[*]Department of Computer Science
[†]Department of English and Comparative Literature

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20766–20791
July 27 - August 1, 2025

some of whom focus on different aspects of these cues. From a writer's perspective, it is important to pay attention to all of these cues to ensure the writing sounds reliable to all readers.

Narratology has explored these questions by attempting to define the *unreliable narrator*, **a first-person speaker who *unintentionally* describes situations misleadingly** (Booth, 1961). Hansen (2007) considers leading definitions and observes "the unreliable narrator is a concept covering very diverse textual phenomena" and accordingly proposes a taxonomy containing different forms of unreliable narrators with "conceptual distinction."

These forms include intra-narrational, inter-narrational, and inter-textual unreliability. The first form, ***intra-narrational unreliability*** is the classical definition that focuses on the presence of verbal tics (textual cues that indicate uncertainty). The left half of Figure 1 shows an example: an excerpt from a blog post where the writer narrates an experience with another person, Dorian, at a bar. The text in orange font indicates content that is narrated in an intra-narrationally unreliable manner because the narrator admits having trouble remembering details due to being inebriated. Consequently their narration is possibly unreliable. The second form, ***inter-narrational unreliability*** occurs when a secondary voice presents a contrasting version of events. For example, in Figure 1, highlighted in green, Dorian does not agree with the narrator regarding the whereabouts of a notebook. Such a contradiction indicates that either the narrator or Dorian must be wrong and raises reader's doubts about the reliability of the narrator. The third form, ***inter-textual unreliability*** involves pattern-matching the narrator with established unreliable character tropes (Riggan Jr, 1978). In Figure 1, highlighted in purple, the reader questions the narrator's reliability because they seem cunning as they bribe the bartender (fitting the trope of *pícaro*). More detailed definitions of each type of unreliability are outlined in Section 3.

Identifying these three forms of unreliable narrators requires picking up on subtle cues that range from specific lexical choices (e.g., a direct statement such as "it's hard to remember") to increasingly abstract inferences (e.g., drawing inferences through statements and actions that a narrator has cunning and self-interest). These forms may contain overlapping characteristics; however, they are classified and determined separately. Hence, a narrator might be unreliable in one of these forms but not another. It is valuable to analyze narrators in this way because it provides in-depth views of the narrator from lexical to abstract contextual levels. Determining narrator unreliability ultimately considers all three forms since together they provide a more complete picture.

In this work, we borrow these definitions from narratology and introduce the task of automatically identifying three forms of unreliable narrators. We pose this problem as a set of binary/multi-class classifications corresponding to the three types of unreliability (shown in the right of Figure 1). We propose that these ideas from the theoretical field of narratology can be used more broadly to identify unreliability across diverse real-world domains.

We observe that as of date there has been no work on analyzing narrator unreliability with automatic methods, and there are no available resources or labeled datasets. Hence, we introduce TUNA, a collection of personal anecdotes from blog posts, subreddit posts, online reviews, and fiction. We hire expert annotators obtaining honors undergraduate or graduate degrees in English literature to annotate these accounts for the three forms of unreliability mentioned above.

To identify unreliable narrators automatically, we explore using large language models (LLMs). We conduct experiments with 6 open and closed-source LLMs of a variety of sizes. We try zero/few-shot settings, fine-tuning, and curriculum learning (Bengio et al., 2009). With these methods, we attempt to learn from labeled data from fiction and generalize this knowledge to real-world text. We observe that classifying unreliable narrators is a very difficult task and encourage future research to further explore its nuances and challenges. Our contributions are as follows:

- We introduce the task of automatically identifying unreliable narrators;
- We borrow narratological definitions for unreliable narrator (i.e., we consider three diverse and increasingly abstract forms: intra-narrational, inter-narrational, and inter-textual);
- We introduce TUNA, an expert annotated dataset of unreliable first-person accounts spanning four different text domains;
- We experiment with multiple methods that learn how to identify unreliable narrators in snippets from fiction and transfer this knowledge to common text read in everyday situations.

2 Background and Related Work

The term "unreliable narrator" is originally defined by Booth (1961): "For lack of better terms, I have called a narrator *reliable* when he speaks for or acts in accordance with the norms of the work (which is to say, the implied author's norms), *unreliable* when he does not." The vagueness of this definition has encouraged more recent narratologists to attempt to define the unreliable narrator in more certain terms (Cannings, 2023; Jacke, 2018; Heyd, 2006; Olson, 2003; Fludernik, 2000; Currie, 1995; Riggan Jr, 1978). Culler (1997) states, "Narrators are sometimes termed unreliable when they provide enough information about situations and clues about their own biases to make us doubt their interpretations of events..." Hansen (2007) builds upon the work of Culler and other salient narratologists to propose a taxonomy with definitions of multiple aspects of narrator unreliability. In this work, we adopt these definitions and taxonomy.

To the best of our knowledge, *there is currently no existing literature that explores automated approaches for identifying unreliable narrators.* We note recent efforts to automatically understand other aspects of protagonists, who are sometimes depicted in first-person (Yuan et al., 2024; Jang and Jung, 2024; Brahman et al., 2021; Huang et al., 2021; Bamman et al., 2013). Additionally, some works attempt to analyze the emotions of protagonists (Brahman and Chaturvedi, 2020; Rahimtoroghi et al., 2017) or their relationships with other characters (Vijjini et al., 2022; Kim and Klinger, 2019; Chaturvedi et al., 2017; Iyyer et al., 2016; Srivastava et al., 2016). Such works indicate that using automatic methods is a reasonable approach for addressing our task.

Classifying unreliable narrators is to a limited extent related to tasks such as the automatic identification of misinformation (Saeidnia et al., 2025; Jarrahi and Safari, 2023) and rumors (He et al., 2025; Kwao et al., 2025). It is also distantly related to deception detection, defined by Burgoon and Buller (1994) as the identification of narrators who intend to commit *deception*, "a deliberate act perpetrated by a sender to engender in a receiver beliefs contrary to what the sender believes is true to put the receiver at a disadvantage" (Hazra and Majumder, 2024; Constâncio et al., 2023; Sarzynska-Wawer et al., 2023; Fornaciari et al., 2021; Van der Walt et al., 2018; Eloff et al., 2015; Almela et al., 2013). These tasks are similar because they analyze first-person narrators and consider aspects of narrative believability. However, Booth (1961) draws a clear distinction by determining that conscious lying is not a characteristic of unreliable narrators; instead unreliability is "a matter of [what is called] *inconscience*; the narrator is mistaken, or he believes himself to have qualities which the author denies him." We follow Booth's reasoning and do not consider narrators who deliberately intend to mislead readers but only those who sound unreliable.

3 Definitions of Unreliability

In choosing our definitions of unreliability, we make two assumptions. Firstly, given a text, we assume that it contains explicit or implicit information that can be leveraged to ascertain the narrator's unreliability (Chatman, 1990). By *explicit* information, we refer to statements that directly state that an account may be unreliable (e.g., the narrator admits to being inebriated during the time of the described events, as demonstrated in the narrative in Figure 1). By *implicit* information, we refer to less direct details (e.g., patterns exhibited by the narrator that resemble unreliable character tropes). Secondly, following Wall (1994), we assume a narrator is reliable until the reader notices explicit or implicit information indicating unreliability.

We borrow definitions from the taxonomy for unreliable narration introduced by Hansen (2007). We note this taxonomy, proposed as the culmination of a broad range of prior definitions, provides a diverse set of tools for analyzing narrators from different perspectives and levels of difficulty. We use three forms that analyze traits with increasingly abstract conceptions of unreliability, as described in the next three subsections. Examples for each of these unreliable forms from the different textual domains are given in Appendix A.

3.1 Intra-narrational Unreliability

In intra-narrational unreliability the narrator exhibits verbal tics, "small interjections and comments that hint at an uncertainty in the narrator's relating of the events", such as "I think" or "it was so long ago, it's hard to remember." (Hansen, 2007). Table 1 shows various types of verbal tics and corresponding examples. If at least one type of verbal tic is present in a text, its narrator is considered intra-narrationally unreliable.

Type	Example
Admission of fault or bias: Explicit admission of mistakes, biases, missing details, or reporting details from another likely unreliable character.	"I tend to see things from a unique point of view.", "Like others of my generation..."
Defensive tone: Multiple phrases in protestation.	"I feel I should explain"
Digressions: Statement that veers off-topic.	"I will do that in a minute. By the way..."
Hedging language: Multiple phrases that indicate uncertainty or vagueness.	"it seems that", "it appears to be", "I think", "maybe", "sort of"
Inconsistencies: Two or more contradicting statements or events that do not add up.	"I am a nobody. But look! There is a plane drawing my name in the sky."
Selective memory: Explicit admission that narrator may have forgotten details.	"It was so long ago, it's hard to remember", "My memory is not what it used to be"
Statement of potential disbelief: Explicit admission that narrative sounds unlikely.	"You might not believe me, but...", "what happened next might seem strange"

Table 1: Examples of verbal tics exhibited by intra-narrationally unreliable narrators.

3.2 Inter-narrational Unreliability

In this form, the narrator is unreliable from a secondary point of view as in the following two cases:

Same-unreliable-character-over-time 👴👴: The narrator is reflecting on events in the distant past when he/she exhibits traits of unreliability *and* the present-day narrator does not indicate change within the narrative snippet (i.e., the current voice of the narrator has traits of unreliability). For example: *"I used to be a crazy man. I'd wait in line each day, desperately hoping that they would let me in. Weee, those were good times."* In this snippet, the narrator describes his distant past as unreliable with "I used to be a crazy man..." His last statement, "Weee, those were good times", indicates his perspective has not changed over time.

Other-character-contradiction 🥦🗣️: Another character contradicts the narrator, typically in the form of direct dialogue. For example: *"I thought the offer from Henry's was incredible. As I picked up a pen to sign, I heard the judge's voice: he had entered the room through the far door and was talking to two well-dressed men. "What scammers these men from Henry's have become," he was saying."* In this snippet, the narrator believes he has received a good offer, but another character, a judge, has a contradicting perspective that the offer is a scam. The reader does not know which character understands the situation best, leaving the narrator's reliability in doubt.

3.3 Inter-textual Unreliability

In this form, if the narrator fits the description of one of the following unreliable character tropes, as defined by Riggan Jr (1978), the narrator is considered inter-textually unreliable:

Naïf 😇: *Blind to wrongs.* Naive observer who lacks the social savvy, maturity, or awareness to understand the complexity of their environment. For example: *"I accepted the assignment willingly. Dimly, I heard the people around me muttering – talking about some danger? I ignored them and went to the other room."* In this snippet, the narrator acts blindly without understanding the situation.

Madman 😱: *Highly emotional.* Narrator, often with a frantic voice, who feels deep positive or negative emotions toward others and is maddened by perceived torture or alienation. For example: *"My heart beat wildly. It took my greatest strength to turn and walk away. How could he? My best friend, a betrayer?!"* In this snippet, the narrator reveals deep negative feelings, perceived alienation, and a frantic tone revealed through stylistic choices.

Pícaro 🤠: *Tries to be cunning.* Socially aware rogue or antihero who experiences the rise and fall of fortune while attempting to improve their prospects and cleverly justifying their chaotic worldview. For example: *"The school teacher scolded me and took away the paper airplane. As soon as her back was turned, I whipped out a fresh sheet of paper, determined to be more stealthy this time. All the while, I kept one eye on the girl who had reported me."* This narrator experiences a fall of fortune when his paper airplane is taken away. He tries to improve his prospects by making a new airplane and shows cunning when he stealthily tries to avoid being caught again.

Clown 🤡: *Flips the narrative.* Narrator who offers reinterpretations that repackage internal and/or external conflict in a new light, potentially from

Corpus	# Samples	Avg	Min	Max
Fiction	499	194.31	24	924
Train/Valid	373	194.74	24	514
Test	126	193.06	48	924
Blog posts	106	315.31	114	1050
Subreddit	112	396.88	73	858
Reviews	100	157.43	53	460

Table 2: TUNA statistics, including the total number of samples and the average, minimum, and maximum number of tokens in each sample per domain. The first row of Fiction is the combination of *Train/Valid* and *Test* subsets (rows 2 and 3 respectively).

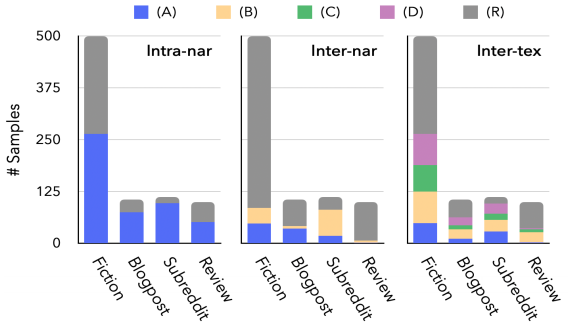

Figure 2: Distribution of resolved labels. For intra-nar (left): # narratives with **(A)** verbal tics or **(R)** none (reliable). For inter-nar (middle): # narratives with **(A)** "same unreliable character over time", **(B)** "other character contradiction", or **(R)** none. For inter-tex (right): # narratives with **(A)** naïf, **(B)** madman, **(C)** pícaro, **(D)** clown, or **(R)** none.

behind a facade that allows them to say whatever they want. For example: *"They called me a coward. What ho! I saw myself rather as my own liberator."* This narrator describes a societal view (that they are a coward) and makes it clear they have a different, reinterpreted view (that they are a liberator).

3.4 The TUNA Dataset

Since there are no currently available resources for classifying unreliable narrators, we build an expert-annotated dataset, **T**exts with **U**nreliable **Na**rrators (TUNA), containing texts with labeled intra-narrationally, inter-narrationally, and inter-textually unreliable narrators. We collect short text samples containing first-person narrators from multiple textual domains, including personal accounts from PersonaBank (*Blog post*) (Lukin et al., 2016), posts from r/AITA[1] (*Subreddit*) (Vijjini et al., 2024), and hotel reviews from Deceptive Opinion[2] (*Reviews*) (Ott et al., 2011, 2013). We intend to first learn how to classify narrators from a fictional domain and then generalize this knowledge to other textual domains. To this end, we additionally collect about 500 narrative snippets from stories from Project Gutenberg (*Fiction*).[3]

All text samples are written in first-person (hand-verified) and range from 24 to 1050 tokens. Samples from Blog post, Subreddit, and Review contain the entire original written text and are arguably complete narratives. We note that Fiction samples are narrative snippets and do not necessarily contain complete stories with fully developed beginnings, middles, and endings. Table 2 shows corpora statistics. Additional details, such as how snippets are selected from the source corpora are given in Appendix D.1.

We design an annotation study, determined exempt by the Institutional Review Board, and ask 10 human annotators with undergraduate or graduate degrees in English literature to read and determine the intra-narrational, inter-narrational, and inter-textual unreliabilities of each narrative. We note that this is a time-consuming task: each sample takes annotators roughly 5 minutes each to read, analyze, and annotate. Because the 3 tasks focus on different aspects of the narrative, annotators report having to re-evalutate the narrative for each task. For 817 narratives, each annotated at least twice, we estimate the study took 172 hours. See Appendix B for additional details.

Annotators are given the definitions and examples of the three forms of unreliability as described in Section 3. For each form, they are tasked with choosing the most relevant unreliable label. If none fit, they may decide that the narrator is reliable for that form. For example, for inter-textual unreliability, the annotator is asked to choose one label from "naïf", "madman", "pícaro", "clown", or "none: reliable". See Appendix B.1 for an outline of the instructions given to the annotators.

Each text sample is annotated by a minimum of two expert annotators. For these pairs of initial results, we calculate inter-annotator agreement with Cohen Kappa's score and observe substantial agreement (Landis and Koch, 1977) across all samples: intra-narrational $\kappa = 0.75$, inter-narrational

[1]Posts are scraped between April 2020 and October 2021 from https://www.reddit.com/r/AmItheAsshole/

[2]The Review dataset contains real and fake (deceptive) hotel reviews and is intended for the task of identifying deceptive reviews. Since the Reviews task differs from identifying unreliable narrators, we only collect real reviews for our dataset.

[3]https://www.gutenberg.org/

$\kappa = 0.71$, inter-textual $\kappa = 0.73$. We improve label consistency by resolving disagreeing labels: annotators participate in robust conversations[4] regarding differing labels and choose the best one. Statistics for the distribution of resolved labels are given in Figure 2 and additional information, including a numerical breakdown of counts, is given in Appendix C.

To encourage thoughtful choices, annotators write short descriptions listing observations and brief explanations for why they demonstrate unreliability. All resolved labels have corresponding descriptions; hence, each narrative has three descriptions (1 per unreliability). We calculate across all descriptions an average of 21.2 tokens, with a maximum of 299 tokens in a given description. See Appendix A for examples.

4 Identifying Unreliable Narrators

4.1 Task Definition

Given n, a text narrated by a first-person narrator, we classify narrators for intra-narrational, inter-narrational, and inter-textual unreliability as follows. For intra-narrational unreliability, we want to determine $n \in \{A, R\}$ where A corresponds to n having verbal tics and R corresponds to n not having verbal tics (intra-narrationally reliable). For inter-narrational unreliability, we want to determine $n \in \{A, B, R\}$ where A corresponds to n having a "same reliable character over time", B corresponds to n having an "other character contradiction", and R corresponds to $n \notin \{A, B\}$ (inter-narrationally reliable). For inter-textual unreliability, we want to determine $n \in \{A, B, C, D, R\}$ where A, B, C, D corresponds to n having a naïf, madman, pícaro, or clown, respectively, and R corresponds to $n \notin \{A, B, C, D\}$ (inter-textually reliable). See Figure 1 for an example with the list of classes.

4.2 Methods

We seek methods that deal with the complexities of classifying unreliable narrators by learning from snippets from Fiction and testing in an out-of-domain manner on real-world domains. For this purpose, we try zero-shot and few-shot settings, fine-tuning using Parameter-Efficient Fine-Tuning with Low-Rank Adaptation (LoRA) (Hu et al., 2022), and curriculum learning (CL) which trains models first on easy and then harder samples.

For CL, the training dataset is divided into easy samples (*Subset-Easy*) and difficult samples (*Subset-Difficult*). We define difficulty of a sample based on how ambiguous it is. Specifically, for each type of unreliability (i.e., intra-narrational, inter-narrational, inter-textual), we observe some samples might contain traits of more than one label. For example, in difficult samples, a narrator who is predominantly a madman might also exhibit some pícaro-like or clown-like traits. Hence, for this sample, in addition to madman, pícaro or clown are also incorrect but reasonable *candidates* for the label. We hypothesize that samples with fewer candidates are easier to classify because there are fewer potential choices for the final label. Samples with multiple candidates are more challenging because each candidate has an arguable, albeit potentially weak, claim to being chosen as the final label.

Based on this motivation, we create (*Subset-Easy*) and (*Subset-Difficult*). For this, the LLM is queried to produce a list of counts for the number of traits for each label. For example, for inter-textual unreliability the LLM generates a list such as, `[A:<NUM>, B:<NUM>, C:<NUM>, D:<NUM>]` where `A`, `B`, `C`, `D` respectively correspond to naïf, madman, pícaro, and clown, and `NUM` is the total number of traits present in the narrative for the given label. Candidates are labels with a `NUM` value > 0. The training samples are ranked accordingly in order of the least to the most number of candidates. The reordered set is divided in half into *Subset-Easy* and *Subset-Difficult*. An LLM is first fine-tuned on *Subset-Easy* and then on *Subset-Difficult* using LoRA adapters with 8-bit quantization for 3 epochs and default PEFT configuration.

5 Experiments

Experiments are performed on Instruct models for Llama3.1-8B, Llama3.3-70B, Mistral-7B, Phi3-medium, GPT-4o mini, and o3-mini (reasoning model). We also compare results with smaller LM classifiers, BERT and ModernBERT. Setup and prompts are described in Appendix D. We use Fiction training/validation samples for model training and development and the (remaining) narratives from Fiction, Blog posts, Subreddit, and Reviews as testing samples. In this way, we test on Fic-

[4] Annotators either meet via video-call or exchange detailed messages. For disagreeing labels, they discuss their choices and select a final resolved label. If they are unable to agree, a third annotator decides the resolved label, given their arguments. Time spent per discussion: simple texts ≈ 2 minutes, samples with very complicated narrators ≈ 15-20 minutes.

tion in an in-domain manner and on the remaining datasets in an out-of-domain manner.

5.1 Results

Table 3 presents performances of CL, fine-tuned, zero-shot, and few-shot methods where macro-averaged F1 scores are provided for each domain (using Llama3.1-8B). Table 4 presents the performance of LLMs averaged across domains, and Table 5 shows the performance of LM classifiers.

We notice six key takeaways. First, generally speaking, all methods and models perform better for the intra-narrational task than for the other two tasks. Similarly, they perform better for the inter-narrational task than for inter-textual. This finding indicates the intra-narrational task is easiest, and the inter-textual task (requiring more abstract inferences) is most difficult for LLMs. Appendix E.2 shows an example demonstrating how the inter-textual task requires a deeper understanding of the narrator's state of mind, making it more difficult.

Second, methods using training samples (i.e., CL, fine-tuning, few-shot) outperform the zero-shot method, indicating that training data does improve LLM performance. Appendix E.3 shows samples where incorrect labels are predicted in zero-shot and correct labels are predicted in few-shot because the model learns from the shots.

Third, for most cases, CL outperforms fine-tuning, indicating that more sophisticated ways of leveraging the training data is promising for better performance.

Fourth, in Table 3, we observe that out-of-domain performances, especially those whose methods use more training data (i.e., CL and fine-tuning), are not better but good compared to in-domain performances. This result indicates that it is possible to learn from the Fiction text domain and apply that knowledge to other real-world text domains. We make similar observations for other models (not shown here due to space constraints).

Fifth, Table 4 shows CL improves performance of smaller models but not larger ones. E.g., Llama3.3-70B few-shot performs competitively with CL and fine-tuning, indicating that as model size increases, learning from fewer samples yields comparable predictive capabilities to learning with more samples.

Finally, for experiments with LM classifiers, we observe average values across all test sets are less than average values for CL and fine-tuned methods. These results indicate that LMs do not outperform

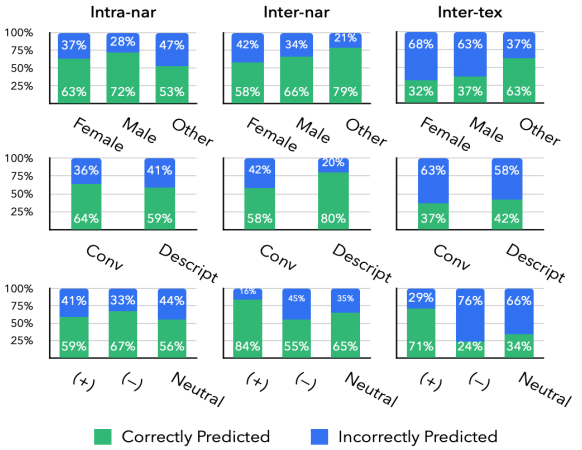

Figure 3: Breakdown of correctly predicted (green) vs. incorrectly predicted (blue) unreliable narrators. Top row: with respect to the narrator's gender ∈ {female, male, other}. Middle row: with respect to narrative style ∈ {conversational, descriptive}. Bottom row: with respect to narrative sentiment tone ∈ {positive, negative, neutral}. Results are from Llama3.1-8B experiments.

LLMs. To understand these observations, we note that for Fiction (in-domain experiment), LMs give results comparable to our zero-shot method; however, for all the other test sets (out-of-domain experiment), the performance of the LMs drastically drops. Hence, we determine that LMs are less capable than LLMs of generalizing knowledge learned from one domain to other domains.

We provide individual performance breakdowns of the remaining LLMs for each domain in Appendix E (including a breakdown of class-wise scores in Table 9) and an error analysis of incorrectly classified narrators in Appendix E.1.

6 Analysis

In this section we analyze unreliability classification with respect to various narrative properties: narrator's gender, number of characters, narration style, and overall narrative sentiment. For these experiments, we use CL outputs for one model from each open-source LLM family (3 total models). We use Llama3.3-70B to automatically infer these narrative properties (the complete prompts and an error analysis are given in Appendix F.1 and F.2).

RQ1: Does the gender of the narrator affect the prediction? Across all testing samples, we count 125 female, 215 male, and 43 other/ambiguous narrators. The first row of Figure 3 shows the percentages of female, male, and other narrators

		CL	Fine-tuned	Zero-Shot	One-Shot	Three-Shot
Intra-nar	*Fiction*	58.51±1.93	50.09±1.96	45.17±1.83	52.67±2.00	51.72±2.12
	Blog post	53.94±2.22	50.63±2.27	45.56±1.80	29.33±4.48	40.54±0.73
	Subreddit	50.04±2.21	49.00±2.05	47.41±1.32	52.03±2.38	48.87±1.86
	Review	67.17±2.16	55.85±2.35	58.46±2.29	60.22±2.20	52.81±2.25
Inter-nar	*Fiction*	34.59±1.82	34.63±2.26	16.20±2.19	15.97±1.19	17.09±1.26
	Blog post	35.92±2.47	28.73±1.80	23.15±2.92	22.19±1.40	27.46±1.47
	Subreddit	30.91±1.80	25.59±1.90	30.97±1.77	22.65±1.35	21.68±1.37
	Review	35.29±1.66	36.59±2.18	25.85±1.79	25.67±3.11	25.37±3.10
Inter-tex	*Fiction*	27.42±1.87	28.59±1.87	18.22±2.38	24.00±1.55	23.54±1.69
	Blog post	19.58±1.78	18.99±1.34	24.23±2.79	28.59±1.75	24.35±1.56
	Subreddit	13.49±1.55	10.85±1.31	12.95±1.21	12.01±1.11	10.71±1.14
	Review	16.72±0.67	17.54±1.35	15.75±1.31	20.32±1.08	19.30±2.08

Table 3: Breakdown of unreliability F1 (macro) scores for each domain for Llama3.1-8B. Improvements on left are statistically significant compared to results on right row-wise with $p < 0.05$ (Dror et al., 2018).

		CL	Fine-tuned	Zero-Shot	One-Shot	Three-shot
Intra-nar	*Llama3.1-8B*	57.42±2.13	51.39±2.16	49.15±1.81	48.56±2.76	48.48±1.74
	Llama3.3-70B	51.26±2.12	51.28±2.09	54.20±1.65	63.89±2.28	61.41±1.91
	Mistral-7B	55.76±1.70	56.46±2.11	56.79±2.05	50.87±1.96	52.99±2.24
	Phi3-medium	53.75±2.14	52.18±2.36	60.00±2.22	44.70±1.69	44.86±1.49
	GPT-4o mini	—	—	47.88±2.05	50.51±1.67	51.77±2.25
	o3-mini	—	—	42.22±1.97	43.47±2.00	44.32±2.04
Inter-nar	*Llama3.1-8B*	34.18±1.94	31.39±2.03	24.04±2.17	21.62±1.76	22.90±1.80
	Llama3.3-70B	33.49±2.31	30.32±1.29	29.11±1.63	31.23±1.72	34.02±2.23
	Mistral-7B	31.15±1.45	25.75±0.44	19.49±1.36	33.07±1.92	31.29±1.86
	Phi3-medium	22.32±1.49	35.76±1.81	25.23±1.88	23.42±1.71	24.66±1.73
	GPT-4o mini	—	—	28.15±1.49	31.48±1.70	26.00±1.52
	o3-mini	—	—	32.18±1.90	28.79±0.91	27.40±1.59
Inter-tex	*Llama3.1-8B*	19.30±1.47	18.99±1.47	17.79±1.92	21.23±1.37	19.48±1.62
	Llama3.3-70B	21.04±1.69	21.02±1.64	28.52±1.96	30.80±1.81	28.23±1.89
	Mistral-7B	29.68±2.01	24.38±1.29	20.23±1.51	18.35±1.43	17.12±1.35
	Phi3-medium	25.00±1.51	26.24±1.82	27.56±1.70	18.84±1.84	16.41±1.38
	GPT-4o mini	—	—	17.84±1.41	20.66±1.42	19.98±1.38
	o3-mini	—	—	16.65±1.14	15.44±0.33	15.84±1.54

Table 4: Unreliability F1 (macro) scores for combined domains for all model families and sizes. Results on left are statistically significant compared to results on right row-wise.

		BERT	ModernBERT
Intra-nar	*Fiction*	48.42	49.48
	Avg	17.77	39.94
Inter-nar	*Fiction*	31.37	38.46
	Avg	25.76	27.07
Inter-tex	*Fiction*	12.46	14.71
	Avg	11.12	16.98

Table 5: Unreliability F1 (macro) scores for Fiction and combined domains for smaller LM classifiers.

classified w.r.t. unreliable narrators correctly vs. incorrectly by Llama3.1-8B for 5 runs across all testing samples. Figure 5 in Appendix F.3 shows results from other models. We observe across all model families that male narrators are predicted correctly more frequently than female narrators. For inter-narrational and inter-textual tasks, other/ambiguous characters are predicted more correctly than either female or male narrators, indicating that performance improves when the narrator is not specified as female or male.

RQ2: How does the narration style change the difficulty of the prediction? The middle row of Figure 3 shows that narratives written in a conversational style tend to perform slightly better than those written in a descriptive style for intra-narrational unreliability. This could be because it might be easier to detect the verbal tics within a conversational tone. However, for inter-narrational and inter-textual tasks, narratives written in a descriptive tone perform better. Figure 5 in Appendix F.3 shows results from other models.

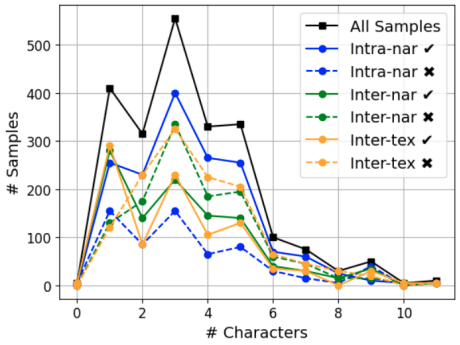

Figure 4: Number of characters vs. number of samples. All Samples (solid black) is the distribution of all narratives with respect to the number of characters. Blue, green, and orange solid lines show correct predictions, and corresponding dashed lines show incorrect predictions. Results are from Mistral experiments.

RQ3: How does the overall narration sentiment affect the prediction? The last row of Figure 3 demonstrates that narratives written in a negative tone perform better than narratives written in a positive tone for intra-narrational unreliability. This result is likely a consequence of negative tones often harboring multiple verbal tics, resulting in an easier prediction. For inter-narrational and inter-textual unreliabilities, narratives written in a positive tone result in significantly better predictions than narratives written in a negative tone. See Figure 7 in Appendix F.3 for results from other models.

RQ4: Are narratives with multiple characters trickier to predict? Figure 4 shows the majority of narratives contain 1-5 characters. Within this range, correct predictions for unreliabilities (solid blue, green, yellow) peak at narratives with 1 and 3 characters. For intra-narrational classification, there are consistently more correct (solid blue) than incorrect (dotted blue) predictions, indicating the number of characters does not change the difficulty of the narratives to classify. For inter-narrational and inter-textual classification, the number of incorrectly predicted narratives (dotted green and yellow) surpasses the number of correctly predicted narratives (solid green and yellow) when the number of characters ≥ 2, suggesting that narratives with multiple characters are tricker to predict than narratives with only the narrator. See Figure 8 and Figure 9 in Appendix F.3 for other model results.

Additional details regarding our methods of performing analysis are given in Appendix F.

7 Conclusion

We propose using automatic methods to classify intra-narrationally, inter-narrationally, and inter-textually unreliable narrators. Borrowing definitions from narratology we define binary and multi-class classification tasks, annotate narratives from a diverse domain of texts, and evaluate the ability of LLMs to perform these classification tasks in zero-shot, few-shot, fine-tuned, and curriculum learning settings. We observe that these tasks are very tricky for LLMs to solve and offer our findings as a call for future work to further investigate the use of NLP methods to identify unreliable narrators.

8 Limitations

Firstly, this work focuses on short texts (no longer than 1050 tokens each), some of which do not contain complete beginnings, middles, and endings. We encourage future work to consider this task for longer-length texts, such as full-length short stories or books. Secondly, we note that all samples in our datasets are written in English. As the definitions of unreliability are applicable to works of other languages, we recommend future work exploring this task on other languages. Thirdly, for *RQ1* in Section 6, we limit our analysis to only female, male, and other/ambiguous genders. Finally, we observe that the size of the dataset is relatively small due to the high cost of high-quality annotations.

Acknowledgments

We are grateful for the suggestions from our anonymous reviewers, and we thank Haoyuan Li, Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Amartya Banerjee for their discussions and valuable insights. This work was supported in part by NSF grant IIS2047232.

References

Angela Almela, Rafael Valencia-García, and Pascual Cantos. 2013. Seeing through deception: A computational approach to deceit detection in spanish written communication. Linguistic Evidence in Security, Law and Intelligence, 1(1):3–12.

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