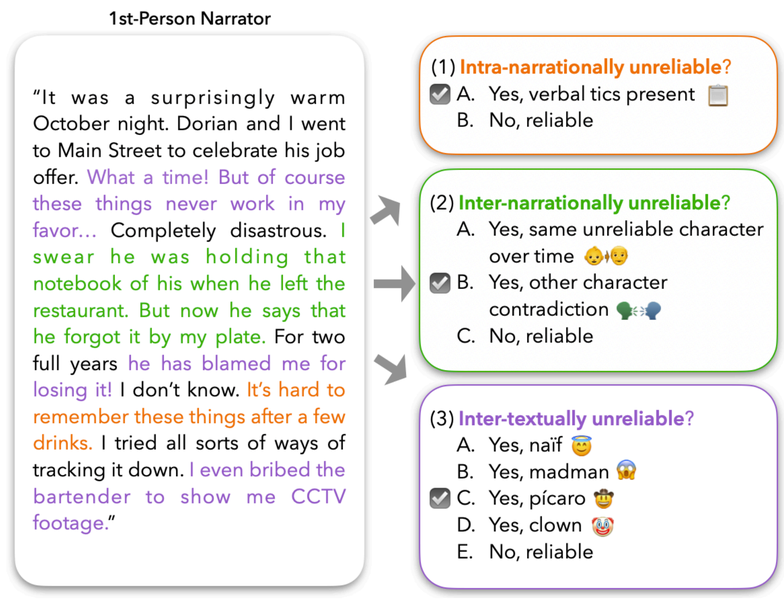
Classifying Unreliable Narrators with Large Language Models

Anneliese Brei, Katharine Henry, Abhisheik Sharma, Shashank Srivastava, Snigdha Chaturvedi
abrei@cs.unc.edu

Overview

- We introduce the task of ***automatically identifying unreliable narrators***;
- We borrow definitions from Narratology for three diverse and increasingly abstract forms of unreliable narrator;
- We introduce **TUNA**🐟, an expert-annotated dataset of unreliable first-person accounts spanning four different text domains;
- We try multiple methods using LLMs and LMs to learn how to identify unreliable narrators in snippets from fiction and transfer this knowledge to everyday text.



Who Is An Unreliable Narrator?

1st-person speaker who unintentionally describes situations misleadingly[1]

Intra-narrational Unreliability[2]

📋 *Contains Verbal Tics*

Contains instances of admission of fault/bias, defensive tone, digressions, hedging language, inconsistencies, selective memory, statement of potential disbelief

Inter-narrational Unreliability[2]

👶➡️👴 *Same Unreliable Narrator Over Time*

Narrator reflects on unreliable self in distant past and in the present does not indicate change within narrative snippet.

E.g., "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."

🥦🗣️ *Other Character Contradiction*

Another character contradicts the narrator, typically in direct dialogue.

E.g., "I thought Henry's offer was incredible. Then I heard the judge say, "Henry is a terrible scammer."

Inter-textual Unreliability[2]

😇 *Naïf*

Blind to wrongs. Naive observer who lacks the social savvy, maturity, or awareness to understand the complexity of their environment.

E.g., "I accepted the assignment willingly. The people around me muttered about some danger? I ignored them and went to the other room."

😱 *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.

E.g., "My heart beat wildly. It took my greatest strength to turn and walk away. How could he? My best friend, a betrayer?!"

🤠 *Pícaro*

Tries to be cunning. Socially aware rogue or antihero who experiences the rise and fall of fortune while attempting to improve prospects and cleverly justify a chaotic worldview.

E.g., "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."

🤡 *Clown*

Flips the narrative. Narrator who offers reinterpretations that repackage internal and/or external conflict in a new light, potentially from behind a facade that allows them to say whatever they want.

E.g., "They called me a coward. What ho! I saw myself rather as my own liberator."

[1] Wayne C Booth. 1961. The rhetoric of fiction. University of Chicago Press.
[2] Per Krogh Hansen. 2007. Reconsidering the unreliable narrator. Semiotica, 165(1/4):227–246.

Texts with Unreliable Narrators (TUNA🐟)

- 817 narratives from 4 diverse domains: fiction, blogpost, subreddit, customer reviews
- Annotated by 10 human experts through rigorous survey

Corpus	# Samples	# Tokens 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

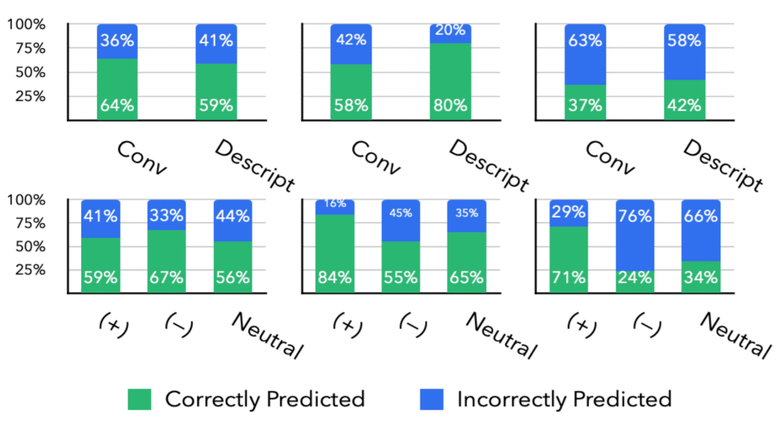
Experiments

- Evaluate 6 instruction-tuned LLMs from Llama, Mistral, Phi, GPT
- Methods: Curriculum Learning (CL), Fine-tuning, Zero-shot, Few-shot
  - For CL: Tune LoRA-adaptor first on easy, then on difficult samples, where difficulty is determined by ambiguity of unreliability
- A Few Takeaways:
  - CL and Fine-tuning improve performance most of the time;
  - Inter-textual unreliability is trickiest task;
  - Subreddit is trickiest domain.

F1-Macro Score		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

Analysis

1. *Conversational tones* show better performance (↑) for intra-narrational unreliability, & *descriptive tones* ↑ for others.
2. *Negative overall sentiment* ↑ for intra-narrational unreliability, & *positive sentiment* ↑ for others.



See paper for more experiments, i.e., analyzing narrator gender and # characters.

https://github.com/adbrei/unreliable-narrators