UNC DEPARTMENT OF COMPUTER SCIENCE

ACL 2025 VIENNA JULY 27 - AUGUST 1

Classifying Unreliable Narrators with Large Language Models

Anneliese Brei
UNC Chapel Hill

Katharine Henry
UNC Chapel Hill

Abhisheik Sharma
Virginia Polytechnic Institute and State University

Shashank Srivastava
UNC Chapel Hill

Snigdha Chaturvedi
UNC Chapel Hill

Introduction

- We read narratives everyday
- *Can we determine if a narrator is unreliable?*
- Gives us a better understanding:
    - about the narrator
    - whether or not we can trust everything written

Who is an Unreliable Narrator?

1st-person speaker who *unintentionally* describes situations misleadingly

(Wayne C. Booth, 1961)

Unreliable Narrators in Fiction

"Dorian and I went to Main Street to celebrate his job offer. What a time! But of course these things never work in my favor... Completely disastrous. I swear he was holding that notebook of his when he left the restaurant. But now he says that he forgot it by my plate. For two full years he has blamed me for losing it! I don't know. It's hard to remember these things after a few drinks. I tried all sorts of ways of tracking it down. I even bribed the bartender to show me CCTV footage."

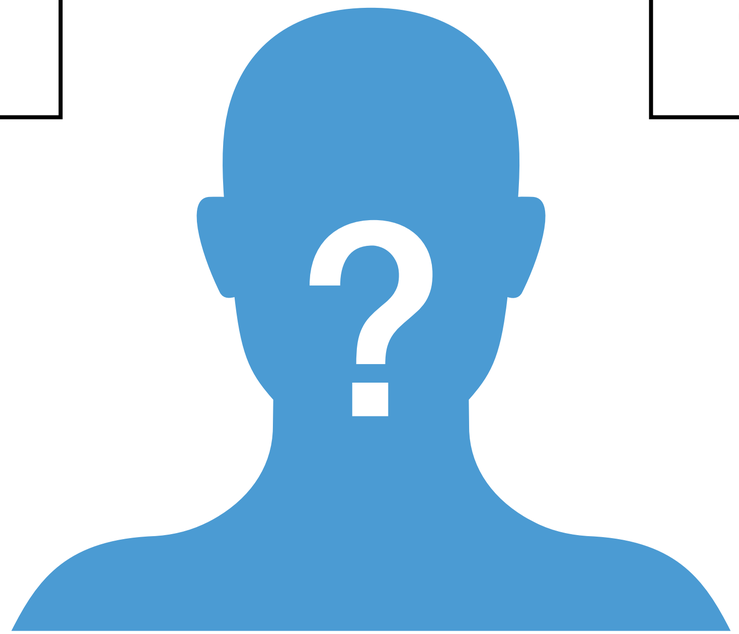
Unreliable Narrators in Fiction

"Dorian and I went to Main Street to celebrate his job offer. What a time! But of course these things never work in my favor... Completely disastrous. I swear he was holding that notebook of his when he left the restaurant. But now he says that he forgot it by my plate. For two full years he has blamed me for losing it! I don't know. It's hard to remember these things after a few drinks. I tried all sorts of ways of tracking it down. I even bribed the bartender to show me CCTV footage."

Unreliable Narrators in Real-World Text

Blog Posts

Opinion Articles



Debate Transcripts

Subreddit

Hotel Reviews

Applications

Why is this Task Challenging?

- Cues often subtle, context-dependent, scattered across text
- Involves deeper understanding of text
- Considers mental state of narrator
- Readers may focus on different aspects of cues

Our Contributions

- Introduce task of automatically identifying unreliable narrators
- Introduce TUNA🐟 dataset
- Test methods of classifying unreliable narrators
- Learn from fiction to classify real-world text

Types of Unreliability

Taxonomy from Per Krogh Hansen, 2007

- Intra-narrational
- Inter-narrational
- Inter-textual

Intra-narrational Unreliability

Narrator exhibits verbal tics

- Admission of fault or bias, defensive tone, digressions, hedging language, inconsistencies, selective memory, statement of potential disbelief

"Dorian and I went to Main Street to celebrate his job offer. What a time! But of course these things never work in my favor... Completely disastrous. I swear he was holding that notebook of his when he left the restaurant. But now he says that he forgot it by my plate. For two full years he has blamed me for losing it! I don't know. **It's hard to remember these things after a few drinks.** I tried all sorts of ways of tracking it down. I even bribed the bartender to show me CCTV footage."

Inter-narrational Unreliability

- Same Unreliable Narrator Over Time
- Other Character Contradiction

"Dorian and I went to Main Street to celebrate his job offer. What a time! But of course these things never work in my favor... Completely disastrous. **I swear he was holding that notebook of his when he left the restaurant. But now he says that he forgot it by my plate.** For two full years he has blamed me for losing it! I don't know. It's hard to remember these things after a few drinks. I tried all sorts of ways of tracking it down. I even bribed the bartender to show me CCTV footage."

Inter-textual Unreliability

- 😇 Naïf
- 😱 Madman
- 🤠 Pícaro
- 🤡 Clown

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.

"Dorian and I went to Main Street to celebrate his job offer. **What a time! But of course these things never work in my favor...** Completely disastrous. I swear he was holding that notebook of his when he left the restaurant. But now he says that he forgot it by my plate. For two full years **he has blamed me for losing it!** I don't know. It's hard to remember these things after a few drinks. I tried all sorts of ways of tracking it down. **I even bribed the bartender to show me CCTV footage.**"

Introducing TUNA🐟

- Expert-annotated dataset
- 817 narratives
- Texts from diverse domains:
    - Fiction, Blogpost, Subreddit, Review

Experiments

Approach problem as binary, multi-class classification tasks

- Zero-shot and Few-shot
- Fine-tuning with LoRA adaptors
- Curriculum learning

Goal: Learn from fiction and test in out-of-domain manner on real-world domains

Curriculum Learning

- Split training set into easy samples and hard samples
  - Easy samples: non-ambiguous, contain cues of only 1 label
  - Hard samples: ambiguous, contain cues of multiple labels
  - Split using counting method with LLMs
- Learn first from easy, then from hard samples

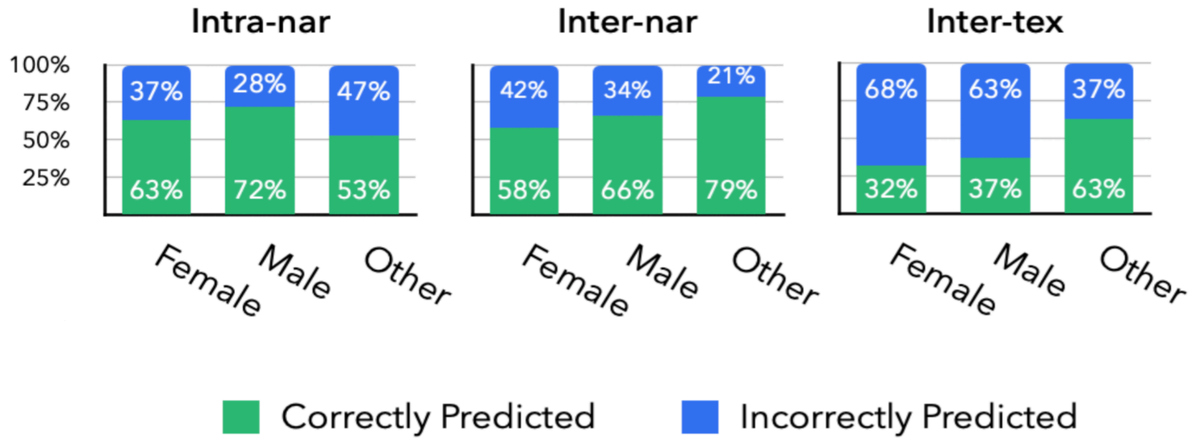
Results

- Most of the time, CL and fine-tuning improves performance
- Inter-textual Unreliability is most challenging task
- Task is very tricky for LLMs

Macro F1-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

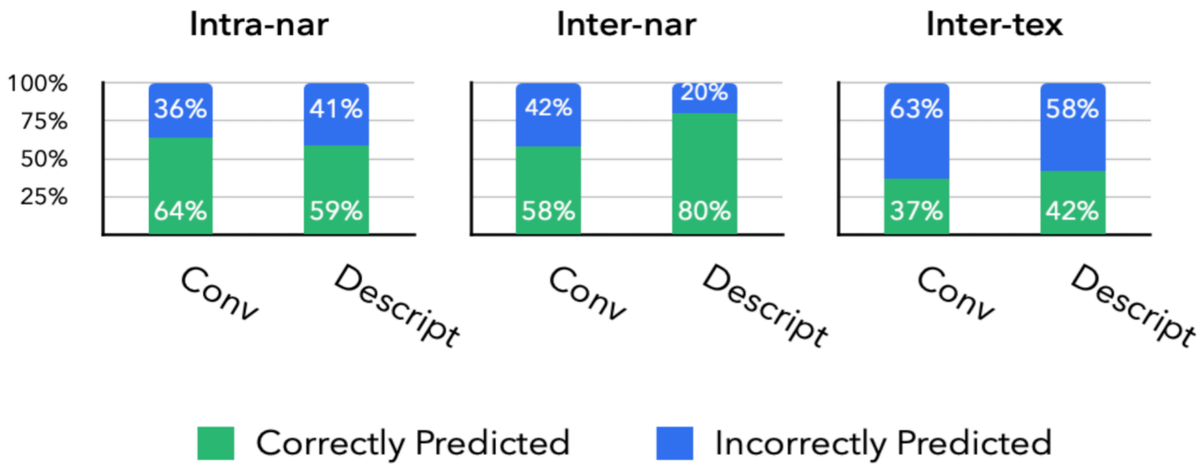
Analysis w.r.t. Narrator Gender

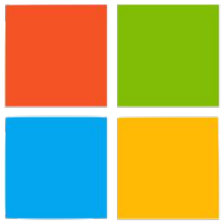
1. *Male* narrators predicted correctly more often than female
2. *Other* narrators predicted correctly most often for inter-nar/inter-tex



Analysis w.r.t. Narration Tone

1. *Conversational tones* demonstrates best performance for intra-narrational unreliability
2. *Descriptive tones* demonstrates best performance for others.





More experiments and results using 6 LLMs from 4 model families given in paper:

Key Takeaways

- Propose task of classifying unreliable narrators with LLMs
- Introduce TUNA🐟 dataset
- Test multiple methods that learn from fiction to classify real-world text
- Analyze LLM capability to solve task –> Difficult task with room to improve