CAspER in the Machine

Insights into Character Variety in LLM-Generated Stories

Anneliese Brei[1], Abhisheik Sharma[2], Nicholas Sanaie[1], Lu Wang[3], Snigdha Chaturvedi[1]

Contact: abrei@cs.unc.edu

(1) University of North Carolina at Chapel Hill, Department of Computer Science; (2) Georgia Institute of Technology, School of Computer Science;
(3) University of Michigan, Computer Science and Engineering Division

Overview

- Adapt narratological methods to automatically understand characters at level of portrayal;
- Construct high-quality labeled dataset of human-written & LLM-generated short stories;
- Compare and analyze LLM-generated & human-written characters.

The Task

- CAspER: method for analyzing 8 axes of character portrayal;
- Using CAspER, address:
  - Do LLM-generated stories and human-written stories have similar characters?
  - Do LLMs generate stories with a variety of characters?

Experiments

1. **Corpus creation:**
   - 200 human-written from *r/shortstories*
   - 4400 LLM-generated from 8 models (*Llama, Phi, Mistral, GPT4o-mini*)
2. **Character classification:** using LLM-judge evaluated on hand-labeled testset of 100 stories with average F1-macro score=84.64.

CAspER Framework

Human-Written Story
"The cafe was busy, but not overwhelmingly so. The before-work crowd was still streaming in, corporate-looking men and women hurriedly ordering coffees and sandwiches…"

Genre: Domestic

Writing-prompt
"Write a story about a character struggling to come to terms with a significant loss from their past, and how the conversation with a friend in a mundane setting triggers a flood of memories and emotions."

Genre: Domestic

LLM-Generated Story
"In the quiet humdrum of her small kitchen, Emma sliced through the last of the cucumbers, the small thud of each piece hitting the chopping board like a metronome keeping time with her thoughts. She had been coming to terms with the loss of her mother…"

- *Stylization vs. Naturalism*
- *Coherence vs. Incoherence*
- *Wholeness vs. Fragmentariness*
- *Literalness vs. Symbolism*
- *Complexity vs. Simplicity*
- *Transparency vs. Opacity*
- *Dynamism vs. Statism*
- *Closure vs. Openness*

Category Classification


Stylized — *Natural*

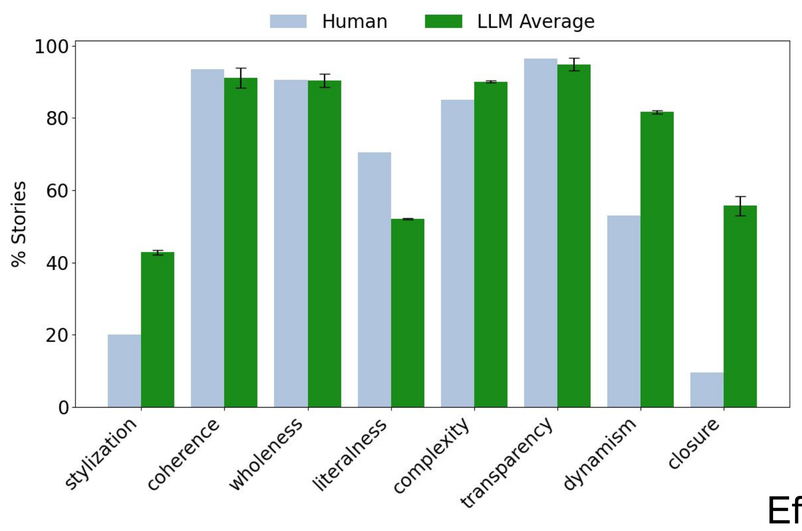
Categories of Human-Written Story

Analysis of Similarities & Differences

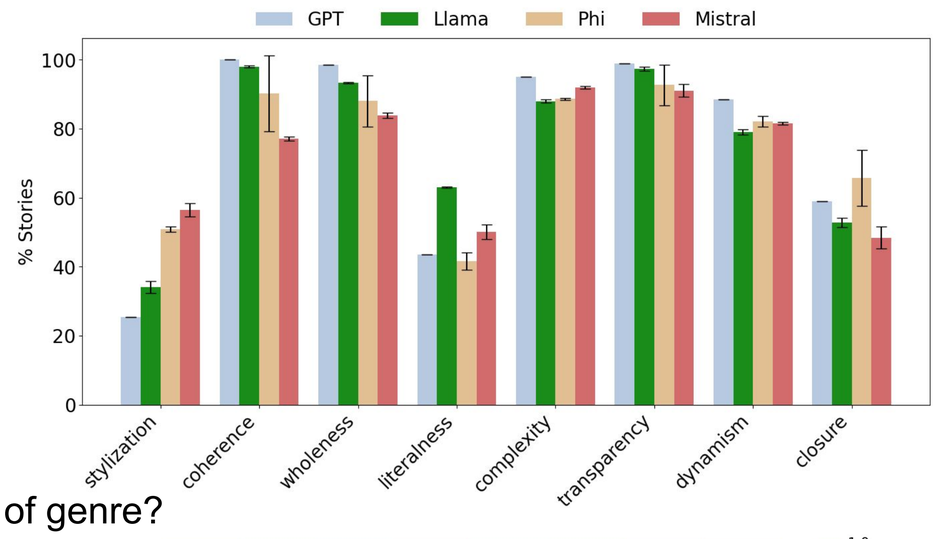
Categories of LLM-Generated Story

Analysis (Highlights)

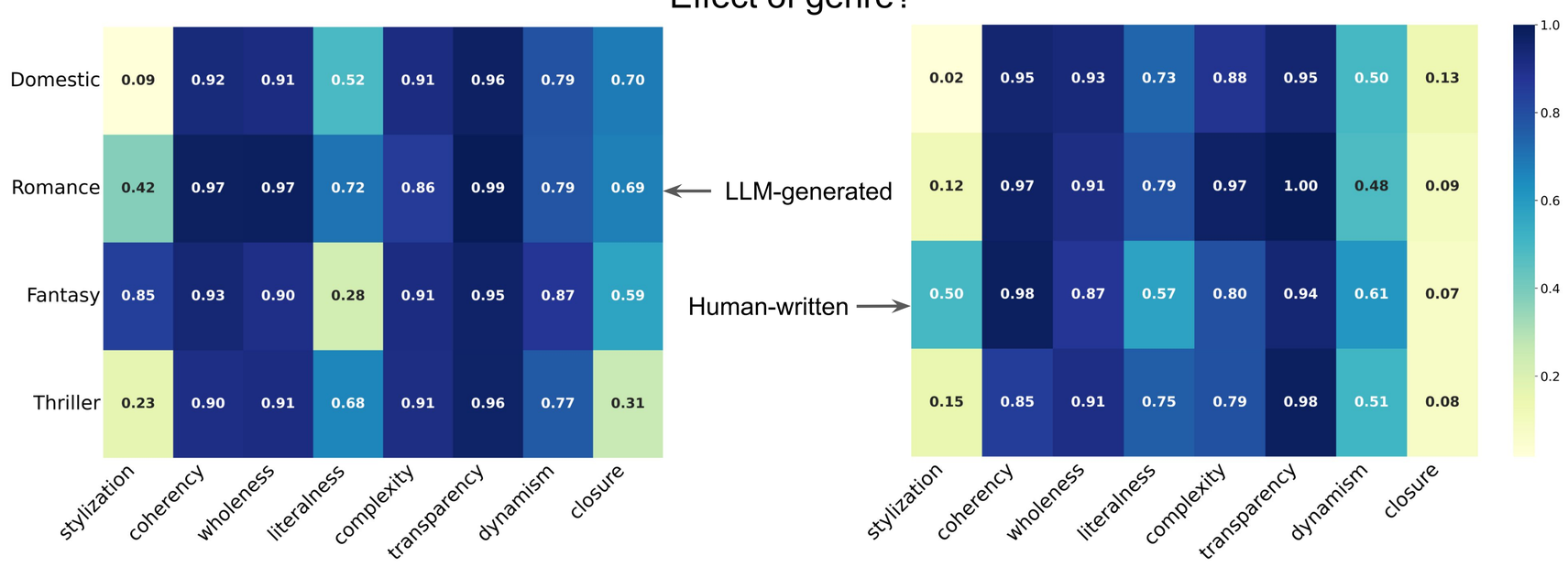
LLM vs. human-written characters?



Characters by different LLM families?



Effect of genre?



Key Takeaways

- LLMs-generated stories are more likely to "play it safe."
- Within family, size does not change types of characters generated.
- Phi family generates most diverse; Llama family generates least diverse characters.
- Genre plays major role in types of character
- When re-generating stories from a prompt, *literalness* & *closure* vary most.

Dumbledore: https://www.imdb.com/gallery/rg2936118016/mediaviewer/rm1061797632?ref_=nm_ecw_g_margolis_i Potter: https://www.imdb.com/name/nm0705356/mediaviewer/rm3471437313/