Trustworthiness of Children Stories Generated by LLMs

Prabin Bhandari & Hannah Marie Brennan
George Mason University

Research Question

Are the children's stories generated by large language models (LLMs) and instruction following models <u>trustworthy</u>?

- Minimizes Error
- Minimizes Biases
- Minimizes potentially harmful content
- Produce age-appropriate content

Methodology

Dataset:
- 122 old stories collected from project Gutenberg.
- 10 modern stories collected from the web.

Models:
- Open Pre-trained Transformer (OPT) [6.7 Billion parameters]
- LLaMA [7 Billion parameters]
- Alpaca (LLaMA tuned for instruction-following)

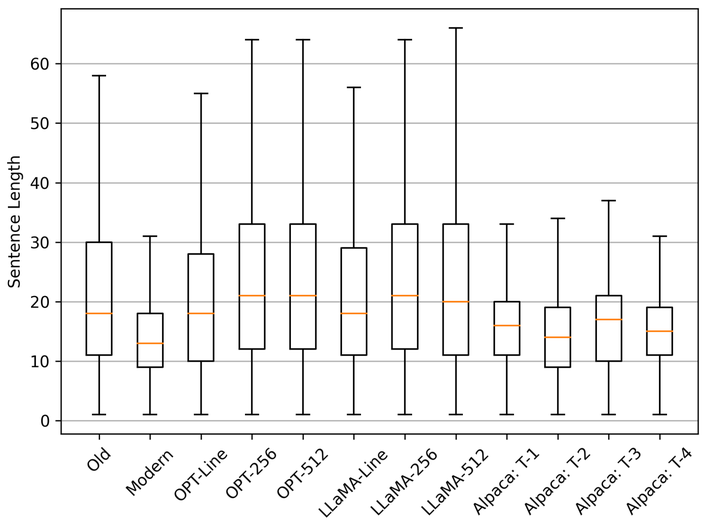
Story generation:

- **LLM**:
  - Provide a part of old story as the context for the LM to generate on
    - First sentence
    - First 256 tokens
    - First 512 tokens

- **Instruction-following Model**:
  - Use four different instruction templates based on old stories

S.N	Template
1	Instruction : Write a short children's story given the title. Input: {TITLE}
2	Instruction : Write a short children's story.
3	Instruction : Write a children's story given the title. Input: {TITLE}
4	Instruction : Write a children's story.

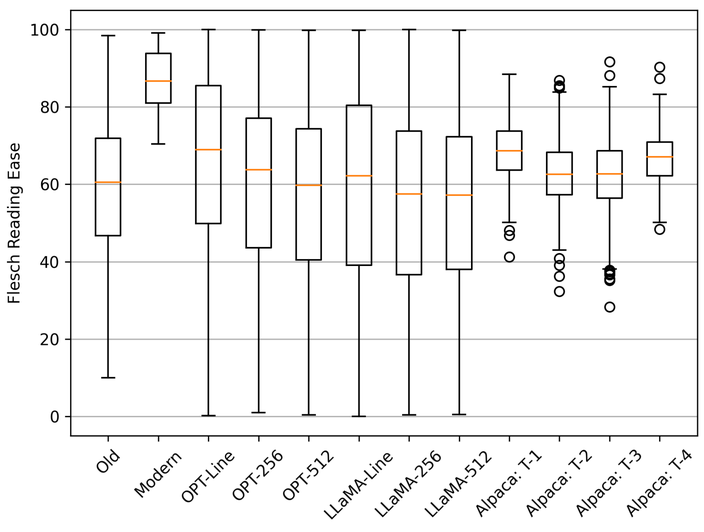
Generated stories follow modern trends but struggle with nuances



- Modern children's stories use shorter sentences.
- Longer prompts result in LLM stories with longer sentences.
- Alpaca's sentences resemble modern stories.

$$FRES = 206.835 - 1.015\left(\frac{total\ words}{total\ sentence}\right) - 84.6\left(\frac{total\ syllabels}{total\ words}\right)$$

- FRES assesses text readability.
- Higher value, easier to read.
- Modern children's stories have higher FRES
- LLMs prompted with older stories tend to mimic the context
- Alpaca generates stories that are easier to read



Generated stories may contain toxic text

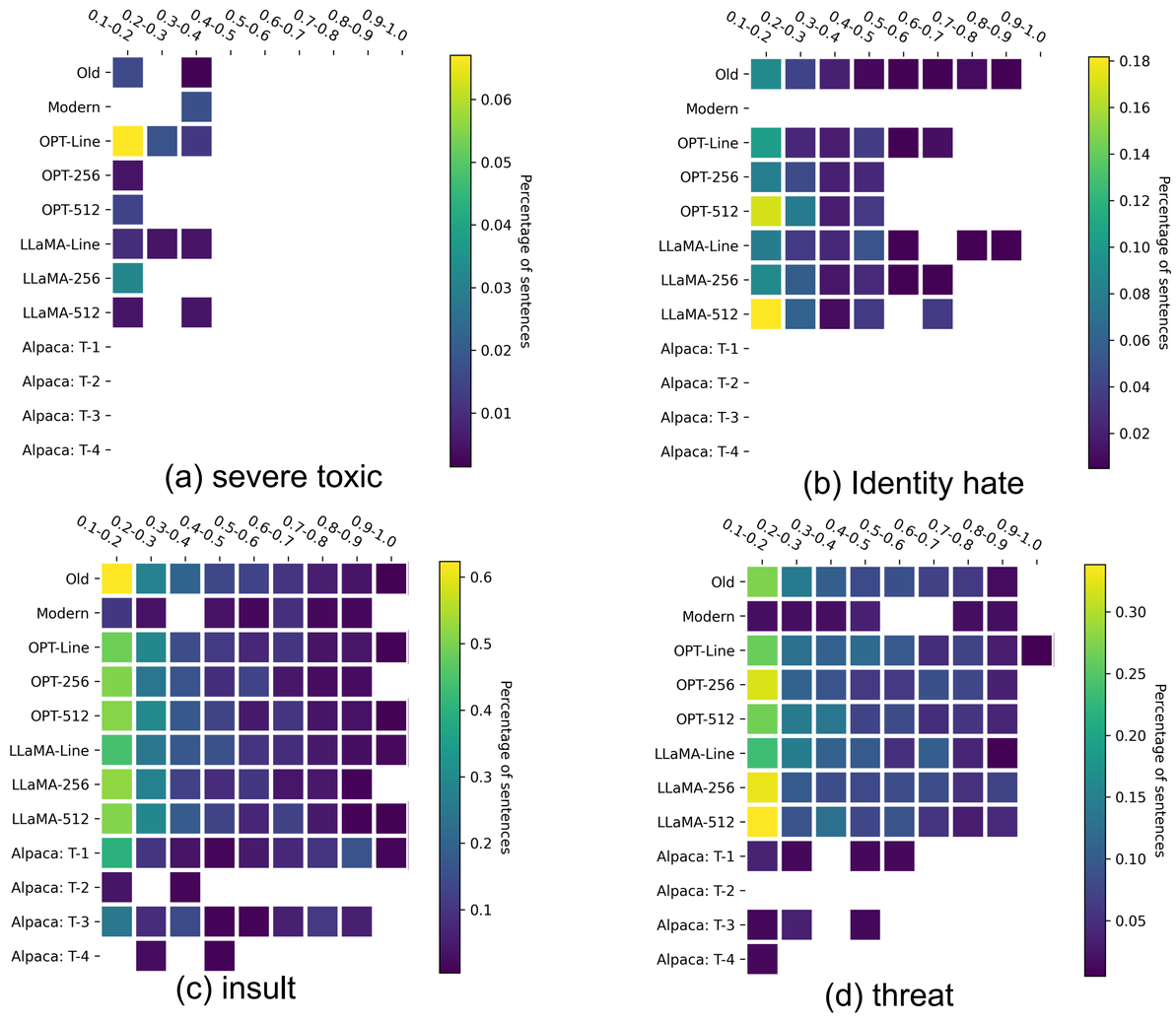
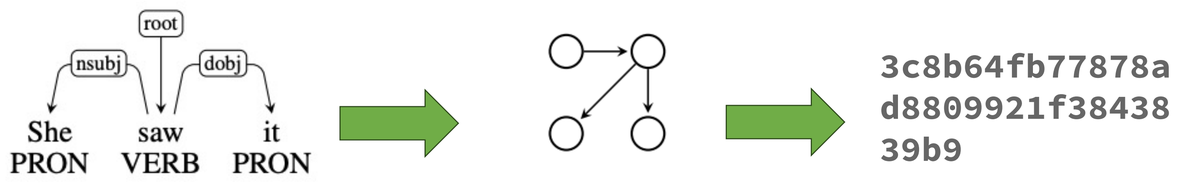


Fig: Various toxicity measures for the actual and generated stories. Each cell in a subplot represents the percentage of sentences rated on a toxicity scale, with x-axis values indicating the toxicity level.

- Older stories tend to be more toxic than modern ones.
- LLMs can learn toxic patterns from context leading to the generation of toxic text.
- LLMs can even generate toxic text from a very innocuous prompt.

Generated stories do not have similar sentence structure to original stories



Dependency Tree → Unlabeled Directed Graph → Weisfeiler Lehman graph hash

Model	Percentage Overlap with	
	Old Stories	Modern Stories
OPT-Line	34.82	34.21
OPT-256	31.37	28.88
OPT-512	32.49	29.89
LLaMA-Line	34.23	33.64
LLaMA-256	32.14	29.82
LLaMA-512	32.27	30.73
Alpaca: T-1	17.31	20.37
Alpaca: T-2	14.67	17.52
Alpaca: T-3	15.20	16.92
Alpaca: T-4	15.41	17.84

Table: Overlap of the hashes of the dependency tree graph of the sentences in generated stories against old and modern actual stories.

Conclusion and Future Work

- Generated stories resemble real ones but lack nuances and may contain inappropriate content.
- LLMs are not yet appropriate for generating high-quality children's literature.
- Future plans: Use reinforcement learning with feedback to improve LLM-generated children's stories.

Scan for Code & Dataset:

Scan to email me: