The Next Chapter: A Study of Large Language Models in Storytelling

Zhuohan Xie Trevor Cohn* Jey Han Lau

School of Computing and Information Systems, The University of Melbourne (*Now at Google DeepMind)

Research Questions?

- **Question 1:** How do LLMs perform in storytelling compared to previous SOTA PLM approaches?
- **Question 2:** What strengths and drawbacks do LLMs have in storytelling?
- **Question 3:** Do LLMs plagiarize from the training datasets?

Experimental Setup

Datasets

- Short commonsense stories: ROCStories (**ROC**)
- Long fictional stories: WritingPrompts (**WP**)
- Long news stories: CNN News (**CNN**)

Models

- LLM Representative: **GPT-3**
- Fine-tuned PLM + Knowledge Enhancement: **KGGPT2** and **HINT**
- Fine-tuned PLM + Storyline Planning: **PROGEN3**
- Fine-tuned PLM + Knowledge Enhancement + Storyline Planning: **MTCL**
- Fine-tuned PLM: **BART**

Quantitative Results (Q1)

Human Evaluation

- **Aspect:** fluency (Flu.), coherence (Coh.), relatedness (Rel.), logicality (Log.), interestingness (Int.).
- **Finding 1:** GPT-3 outperforms other models by a large margin and can even rival human writers.
- **Finding 2:** Most story generation models excel regarding fluency and coherence but face challenges when it comes to more complex aspects like relatedness, logicality, and interestingness.

Table 1. Crowdsource Human Evaluation Results.

	Model	Flu.	Coh.	Rel.	Log.	Int.
ROC	GPT-3	**4.40**	4.43	4.37	4.37	3.57
	KGGPT2	3.90*	3.48*	3.53*	3.00*	2.62*
	PROGEN3	3.88*	3.45*	3.37*	2.95*	2.57*
	MTCL	3.55*	3.12*	3.18*	2.73*	2.42*
	HINT	3.90*	3.27*	3.33*	3.12*	2.58*
	BART	3.92*	3.38*	3.48*	3.03*	2.60*
	human	4.22	**4.58**	**4.42**	**4.48**	**3.77**
WP	GPT-3	**4.37**	**4.67**	**4.28**	**4.48**	**3.47**
	PROGEN3	3.45*	3.08*	2.35*	2.57*	1.98*
	HINT	3.32*	2.63*	2.02*	2.25*	1.77*
	BART	3.42*	2.73*	2.08*	2.27*	1.87*
	human	4.13*	4.22*	3.05*	3.75*	2.97*
CNN	GPT-3	**4.22**	**4.52**	**4.58**	**4.60**	3.20
	PROGEN3	3.63*	3.32*	3.30*	3.22*	2.28*
	BART	3.58*	3.37*	3.30*	3.27*	2.17*
	human	4.10	4.10*	4.23*	4.18*	**3.72***

Automatic Evaluation

- **Finding 1:** In general, there is no clear winner among the current automatic evaluation metrics, indicating a need for improved automatic evaluation methods.
- **Finding 2:** Recent model based evaluation metrics BERTScore (BES), BLEURT (BRT), and BARTScore (BAS) correlates better to human evaluation than lexcical-based ones Corpus BLEU (CBL) and MS-Jaccard (MSJ).

Table 2. Automatic Evaluation Metrics Results.

	Model	Flu./Coh.				Rel.
		CBL(↑)	MSJ(↑)	BES(↑)	BRT(↓)	BAS(↓)
ROC	GPT-3	27.2	11.6	86.6	**8.6**	**4.2**
	KGGPT2	33.5	**15.0**	**87.0**	9.5	4.6
	PROGEN3	26.6	14.6	86.7	9.7	5.0
	MTCL	31.4	14.2	86.9	9.7	4.9
	HINT	**39.6**	13.7	**87.0**	**8.6**	4.3
	BART	27.5	14.7	86.8	9.5	5.0
WP	GPT-3	28.6	12.3	**81.6**	**11.7**	4.3
	PROGEN3	32.3	**16.4**	81.4	13.3	5.4
	HINT	**45.5**	12.8	80.8	12.1	**4.1**
	BART	32.6	16.2	81.4	13.0	5.4
CNN	GPT-3	**33.2**	11.0	**83.5**	**7.5**	**4.4**
	PROGEN3	29.6	**14.8**	82.2	9.3	5.2
	BART	29.1	14.7	82.2	9.8	5.2

Qualitative Analysis (Q2)

Strengths

- GPT-3 owns extensive knowledge coverage through pre-training, and therefore, able to produce a coherent story given by a title *cthulhu calls your cell, he needs to be picked up*, that does not provide information about the fictional cosmic entity *cthulhu*.

Table 3. Stories generated by GPT-3 and BART given the same title "cthulhu calls your cell , he needs to be picked up ."

Model	Story
GPT-3	" yes , lord cthulhu ? " you answer the phone . " i need you to pick me up from the airport . " " of course , my lord . when will you be arriving ? " " in about an hour . i 'll text you the flight number . " you hang up and start getting ready . you know cthulhu is a demanding deity , but you do n't mind . you enjoy serving him .
BART	"ah it ' s here . i thought you wanted to leave , a phone call . * " my coworkers looked at me confused . "i wasn ' t in the mood to go running around scared . but it was important . you need to pick up on your cravings . the kid-ta-bye , gabriel , payliver , cilia de los angeles . and four people ' s fault. " he was right about the fifth , at least my attention was fixed on it . three girls were outside screaming in the street , and a family of four was picking up one of them . the two men yelled and approached as they walked to the mailbox , visibly distressed .

Drawbacks

- GPT-3 contains generation drawbacks such as (1) Multilingual, (2) Tokenization issue, and (3) Expletives. Table 4 shows some examples generated by GPT-3.

Table 4. Examples of some interesting generation errors we find for GPT-3.

Issue	Title	Generation
Generate Chinese from an English title	you are an alien body snatcher , and this is your first day in your new host body . you have no idea what his life is .	**你不是人，你是外星人。这是你第一天在新的主人身上。你不知道他的生活是什么。你只是偶然得到他的身体，他的生活已经开始了，而你还不知道这是什么。你的主人是一个普通人，他没有你的外星人能力。你不知道你的主人在做什么，也不知道他有没有你的朋友。你只是在他的身体里，完全不知道他的生活。**
Stick word together to create "new" words	write a short story based off the lyrics to any song .	there was a timewhen we were youngand our dreams were alivewe wanted to be rock starsand travel the worldwe wanted to be someonesomeone famoussomeone lovednow we 're just oldour dreams have diedour passions gonewe just want to be left aloneto die in peace
Generate profanity words	[FEMALE] was eating lunch at school .	a boy came up to her and asked her to go out with him . she said no and he called her a stuck up b * * * * .

Please be aware that the generated text may contain offensive content.

Plagiarism Study (Q3)

- **Manual Check:** 10 GPT-3 generated stories are randomly sampled for WP and CNN respectively and manually checked whether they copy from web articles by searching for keywords, named entities and long phrases.
- **Result:** For CNN, 7 instances are based on real stories from news articles while no instance is found for WP. One example of plagiarism shown in Table 5.

Table 5. One example of "soft" plagiarism of GPT-3 when generating news.

Generated CNN Story	Wikipedia Article
(CNN) – A former University of Virginia lacrosse player convicted of killing his ex-girlfriend was sentenced to 23 years in prison on Thursday, according to the Commonwealth Attorney's office in Charlottesville, Virginia. George Huguely V was convicted in February of second-degree murder in the May 2010 death of Yeardley Love ...	The murder of Yeardley Love took place on May 3, 2010, in Charlottesville, Virginia, United States. Love, a University of Virginia (UVA) women's lacrosse student-athlete, was found unresponsive in her Charlottesville apartment and later that day, UVA men's lacrosse player George Wesley Huguely V was arrested by Charlottesville police...

Conclusions

- Stories generated by GPT-3 are substantially better than SOTA models on multiple aspects and even rival human authors. Therefore, storytelling has entered **The Next Chapter** with LLMs.
- GPT-3 demonstrates extensive knowledge coverage, leading to outstanding performance in creative tasks like storytelling. However, it occasionally has decoding issues.
- GPT-3 has a tendency to reproduce details or plots from its memories, raising foundational questions about its generation creativity.

More Information