

A SOTA Story Models Details

Knowledge Enhanced GPT-2 (KGGPT2) Guan et al. (2020) use heuristic rules to translate commonsense triples from commonsense knowledge bases (e.g., ConceptNet (Speer and Havasi, 2012) and ATOMIC (Sap et al., 2019)) into natural language sentences and fine-tune GPT-2 small using these sentences. They also use rules to construct negative samples from the original stories to create “bad stories” and perform additional training to encourage the model to learn representations that can distinguish the original and negative stories on ROC.

Progressive Generation of Long Text (PROGEN) Tan et al. (2021) divide the story generation process into multiple stages where words are generated based on their order of importance (estimated using TF-IDF). In other words, PROGEN does not generate stories in a left to right manner. They fine-tune BART-large in different stages where the early stages focus on generated keywords and the intermediate stages focus on generating the next set of content words. We use PROGEN3 in our experiment which has 3 stages where it generates 15%/25%/100% of the story words after each pass.

MEGATRON-CNTRL (MTCL) Xu et al. (2020) combines commonsense reasoning and storyline planning. They first train a keyword predictor with GPT-2 and the predicted keywords are used to retrieve related knowledge triples from a knowledge base. They then train a contextual knowledge ranker with BERT to rank the top- N predicted knowledge triples. A second GPT-2 is trained as a conditional generator that takes both top ranked knowledge triples and other conditioning (e.g., titles) as input when generating stories. Note that the parameters of the two GPT-2 and BERT models are initialised using MEGATRON parameters (Shoeybi et al., 2019).

High-Level Representations for Long Text Generation (HINT) Guan et al. (2021a) pre-train BART-base on BookCorpus (Zhu et al., 2015) with additional objectives that capture sentence-level similarity and sentence-order to learn the internal structure within a story. The model is then further fine-tuned on story datasets to generate stories in a particular dataset.

BART This is a baseline model where we fine-tune BART-large on the story datasets with the standard next word prediction objective.

B Datasets Details

ROCStories (ROC) ROC was developed by Mostafazadeh et al. (2016) and it contains 98K commonsense stories of five sentences. To obtain a more generalised lexicon, we follow the delexicalisation process from prior studies (Guan et al., 2020; Xu et al., 2020) where male/female/unknown names are replaced by [MALE]/[FEMALE]/[NEUTRAL] sentinels. For each story, the first (leading) sentence is used as conditioning context, and models are trained to generate the remaining 4 sentences.

WritingPrompts (WP) WP consists of 303K human-written stories mined from Reddit’s Writing Prompts forum Fan et al. (2018).⁷ Each story is trimmed to contain only the first 10 sentences (following Guan et al. (2021a)). For WP, we use the prompt (which is typically a paragraph of text that sets the scene of the story) as conditioning for story generation.

CNN News (CNN) CNN News (Hermann et al., 2015) is a dataset that contains long news articles with titles. CNN is a very large dataset, with 311K news articles and highlights. We sub-sample the standard training, validation and testing splits to produce splits with 10K/5K/1K stories each, respectively, for our experiments. The title of a news story is used as conditioning for story generation.

C Amazon Mechanic Turk Setting

Qualification Requirements We set following qualification requirements for our annotators: 1) Their accept rate is greater than or equal to 97%. 2) Their location is in US. 3) They have to complete more than 1000 HITs.

Questions We ask the following questions in our questionnaire.

1. Fluency: “How grammatically correct is the text of the story?”
2. Coherence: “How well do the sentences in the story fit together?”
3. Relatedness: “How relevant is the story to the title?”

⁷<https://www.reddit.com/r/WritingPrompts/>

4. Logicity: “How much does the story obey commonsense?”
5. Interestingness: “How enjoyable do you find the story?”

D Amazon Mechanic Turk Pilot Study

While AMT is convenient to find workers for annotation work, it can be rather difficult to obtain reliable workers (Karpinska et al., 2021; Clark et al., 2021). One of our workers told us many workers install website plugins to help them to manage the workflow with AMT so that they can hoard many HITs at the same time. Therefore, HITs with high payment can easily attract irresponsible workers even though previous qualifications are set since most AMT requesters will not bother to reject work.

Therefore, we set a pilot study to aid us to help reliable workers. We randomly select 5 stories generated from different models on ROC and 1 story from the test dataset. We then train a trigram language model on ROC to mimic the style and generate 1 story from the trigram model. All stories have different titles. We randomly shuffle these 7 stories and the task is to ask people to evaluate all stories with questions mentioned in Appendix C and we will judge the quality of their evaluation based on human and trigram stories.

We invite 7 of our colleagues, which are all from non-English speaking countries to have a rough idea of the difficulty degree of the task. We calculate the average score of all quality metrics except the interestingness aspect since it is subjective. On average, our colleagues rank the human story as 4.5 and trigram story as 1.425, which shows our task is not hard to distinguish human and trigram stories. We set a rather lenient standard as “ranking human story ≥ 3.5 and trigram story ≤ 2.0 ” to select workers from our pilot study.

We create 100 assignments of the same HIT at different times with the qualification mentioned in Appendix C. We find running the same pilot study at different times can obtain quite different results from AMT, which align to the findings in Karpinska et al. (2021). Generally, we find that more reliable workers can be found in the evening of Eastern Daylight Time (EDT). We have 10 out of 100 people pass the pilot study but only 5 people pass it on the day. It shows the difficulty of obtaining reliable workers on AMT nowadays and the economic importance of running a pilot study

	IAA	Flu.	Coh.	Rel.	Log.	Int.
ROC	<i>r</i>	0.64	0.81	0.79	0.80	0.68
	TA	17.24	24.98	25.57	27.37	22.03
WP	<i>r</i>	0.51	0.70	0.74	0.71	0.54
	TA	18.37	17.01	32.65	19.73	12.93
CNN	<i>r</i>	0.46	0.54	0.61	0.59	0.50
	TA	15.13	12.61	15.97	11.76	14.29

Table 7: Inter Annotator Agreement (IAA) results for each aspect: fluency (Flu.), coherence (Coh.), relatedness (Rel.), logicity (Log.) and interestingness (Int.). We use one-vs-rest Pearson’s *r* to assess the extent to which each annotator agrees with the consensus. Total Agreement (TA) means the percentage where all 3 annotators choose the same score.

before conducting real research. We grant those reliable workers the customised qualification and only invite them to our real study, we also have controlled stories to monitor the quality of workers, as 2 controlled stories inserted into each HIT.

E Amazon Mechanic Turk Issue

Our human evaluation is conducted over AMT, even though it is convenient and affordable, we find a big disagreement between our annotators. We first conduct a pilot study to test the capability of annotators to evaluate English stories and only invite workers that pass our proficient English stories reading tests to the evaluation of sampled stories. We only gave them two examples showing how we assess the example stories but we did not provide detailed English stories evaluation training to our annotators. We did not have a main annotators that can provide a standard score for example stories, which increase the difficulty of judging the quality of evaluation work we receive from AMT.

Also, as pointed out in Karpinska et al. (2021), the quality of work from annotators on AMT platform can be of high variance and have poor calibration, therefore, we would obtain more reliable human evaluation results if we hire expert raters such as professional authors or English language teachers.

F Inter-annotator Agreement for MTurk Workers

We follow Lau et al. (2020) to estimate one-vs-rest agreement using Pearson’s *r*. For each story, we single out an annotator’s score and compare it to the mean scores given by the other two annotators, and we repeat this process for every score in a story

	IAA	Flu.	Coh.	Rel.	Log.	Int.
ROC	r	0.42	0.54	0.66	0.59	0.32
	TA	38.57	25.0	25.71	25.71	8.57
WP	r	0.36	0.57	0.73	0.49	0.54
	TA	10.0	10.0	18.57	10.0	10.0
CNN	r	0.36	0.41	0.47	0.37	0.35
	TA	17.14	10.71	14.29	10.0	4.29

Table 8: In-house IAA Results.

and for all stories to compute Pearson’s r over the two sets of scores (singled-out scores vs. mean scores). We also compute the percentage where all 3 annotators choose the same score, noting that this is a much stricter agreement metric (as it does not capture the ordinal scale of the scores). Random scoring would produce 4% for this metric.

IAA results are presented in Table 7. In terms of one-vs-rest agreement (r), we find overall good agreement with 9 strong agreement results ($r \geq 0.6$) and 6 moderate agreement results ($0.45 \leq r \leq 0.6$). We see some correlation between story length and agreement, as ROC has the highest agreement (shortest with 5 sentences) and CNN has the lowest (over 20 sentences). When it comes to aspects, coherence, relatedness and logicity have higher agreement compared to fluency and interestingness. While it is intuitive to see interestingness being subjective, fluency is somewhat a surprise. Manual inspection reveals that annotators have very different standards when it comes to fluency, with some workers being more strict about grammar, which contributes to the low agreement. For total agreement (TA), the numbers range between 10–25%, which is encouraging as it shows that there is still a good proportion of cases where all annotators agree on a score.

G Inter-annotator Agreement for In-house Workers

The In-house annotation agreement results are shown in Table 8. Surprisingly, we see similar trends in compared to Table 7, which means our findings are consistent even with a different group of workers on a different set of stories.

We again see fluency and interestingness are two aspects that have lower agreements. However, it is interesting that we find fluency has highest TA but rather lower r . Manual inspection reveals that annotators mainly select 4 or 5 for fluency (also can be seen from high fluency scores for ROC in Ta-

ble 5), which benefits TA because they have higher chance to agree on the same values, but might harm r because the score can be negatively influenced by other rare values. Relevance again has the highest agreement, which might be because annotators agree that many stories generated by SOTA models are not related to the given condition. However, compared to IAA results from crowdsource annotators, in-house annotators have lower agreement in general. we find 2 strong agreement results ($r \geq 0.6$) and 6 moderate agreement results ($0.45 \leq r \leq 0.6$).

H GPT-3 World knowledge Strengths

Table 9 and 10 demonstrates that GPT-3 can generate better stories than SOTA models thanks to its wide world knowledge strengths.

I Mechanic Turk Instructions

Figure 1 and 2 demonstrate our instructions and rubrics.

J Example Prompts

Table 11 shows some examples we use in our prompts for generation.

K GPT-3 Stories vs Human References

Table 12, 13 and 14 show some examples where GPT-3 obtains higher scores than its human reference.

L Example Generations

Table 15 shows randomly picked generation example for each dataset.

M Soft Plagiarism Examples

Table 16 shows some GPT-3 generated news stories where we find true online stories.

N GPT-3 Generation Issues

Table 17 shows some GPT-3 generated stories with several issues.

O Pearson Correlations between Each Story Aspects

Figure 3 and 4 present Pearson Correlations between Each Story Aspects for MTurk and in-house workers respectively.

Model	Story
GPT-3	as soon as i realized that i had died , the first thing that came to mind was how pointless my life had been . i had devoted myself to a cause that was ultimately fruitless and destructive , and in the end i had caused the deaths of millions of people .but then i realized that there was still a chance for redemption . in the after life , i was given the opportunity to travel to different moments in history and try to make things right . i could see the consequences of my actions , and i could try to make different choices that would lead to a better outcome .i started with the holocaust , and i was able to save many people from being killed . but it was difficult , and i had to make a lot of tough choices . eventually i moved on to other moments in history , trying to make things better wherever i could .it was a difficult journey , but it was worth it . i was finally able to make up for the mistakes that i had made in life .
PROGEN3	i knew i was in a bad mood . the blood dripped from my nose and down my chin like the burly balding drifter . was it dark yet ? i could feel my brain becoming small dots of the rotten material . the blood stopped flowing . almost instantly , the world dissolved in my hands . that 's when i noticed something . i was still standing in the middle of the street . i turned to my left and saw the madman .
HINT	“ hi , “ i said , opening my eyes to a bright light . “ this is hitler . “ i looked at the clock . 5:00 am . “ what 's your name ? “ “ hitler . “ “ hitler . “ “ hitler . “ “ hitler . “ “ hitler . “
BART	i 'm struck with a fierce sense of shame at that thought . unfortunately it 's been a while since i 've felt this way , and normally it 's * bad * to feel that way , but i think that too should be expected when you 're still alive . there has been so much turmoil since the war . we 've lost children , we 've lost society , we 've even lost ourselves . but now , we 're free . we can finally go on with our lives , like we always should . but something happens that seems wrong , and it 's really hard to forget . there are my reasons for feeling so sad . why will i be mad today ?

Table 9: Stories generated from GPT-3, HINT, BART and PROGEN3 given the same title “you 're hitler , and you 've just died . there is an after life . tell me about it .”

Model	Story
GPT-3	the young man had lost everything . his family , his home , his country . he had nothing left to lose . when he was offered the chance to join isis , he saw it as a way to finally have a purpose in life . he was given a sense of belonging and a sense of worth . he was told that he was fighting for a good cause and that he was helping to protect his people . he was given a sense of identity and a sense of purpose . he was told that he was part of something bigger than himself .
PROGEN3	they 've crippled the ranks of these countries . their leaders not going to the president 's party , everyone having their noses to the grindstone , the war . political parties shooting each other with a firing squad , only to end up murdering each other . “ the one who executed your followers was the first one to hell hounq taqibs . “ he 's right , ha is not 'd understand * reason * if he knew that . “ you 're killing my people . “
HINT	my eyes stare at the ceiling . it 's just a window , it 's just a window . i 'm not sure if i should close it or close it again . i 'm afraid i ca n't move it . i do n't want to be here . i do n't want to be here . i want to be able to talk to the people who will help me . to be able to hear the voices of people who will help me . to be able to see the lights . to be able to see the people who will help me .
BART	missdamie looked at her young daughter sat with c and o . she 'd put on a radiant glow and bound herself up in an artificial bubble . she was beautiful . she twitched at the vibrancy in her hands . her hair was beautiful and silver . her smile was off . her black nails placed on gentle curves . when she smiled her smile widened and the small eyes of her daughter 's eyes danced . “ isis ? “ i asked , shakily , behind the mask . ms. de los angeles nodded and smiled . “ yes , my name is isis de los angeles , and “ her mother held her hand . “ my son is wondering if isis can cause such a thing . “

Table 10: Stories generated from GPT-3, HINT, BART and PROGEN3 given the same title “make me sympathise with an isis member .”

Instructions:

Please read each story and judge its quality based on the five criteria given below. You have to finish assessing all stories before you press the "Submit" button.

Note:

- Some names in the stories are replaced with "[MALE]", "[FEMALE]" and "[NEUTRAL]", and all words are lowercased. Please do not treat these changes as grammar errors.
- These five criteria we list below should be judged **independently**.
- **Fluency** only measures if each sentence of story has grammar errors.
- **Logicity** only measures if the story make senses based on your commonsense knowledge, For instance, "sprayed the burn with a napkin" in Example Story One does not make sense since we can't use a napkin to do the spraying.
- **Relatedness** only measures if a story is related to the title.
- **Coherence** only measures if the narrative of the story flows well - you do not need to consider how well the story fit the title in this criterion.
- **Interestingness** is a subjective metric - you only need to select based on the extent to which the story interests you and you do not have to consider if it makes sense.

Please Judge Story One

Title: \${title1}

Story: \${story1}

Fluency: How grammatically correct is the text of the story?

1(lowest) 2 3 4 5(highest)

Coherence: How well do the sentences in the story fit together?

1(lowest) 2 3 4 5(highest)

Relatedness: How relevant is the story to the title?

1(lowest) 2 3 4 5(highest)

Logicity: How much does the story make sense?

1(lowest) 2 3 4 5(highest)

Interestingness: How enjoyable do you find the story?

1(lowest) 2 3 4 5(highest)

Figure 1: A screenshot of our evaluation questions.

Special Note: Coherence and Logicality measure two different aspects.

For instance:

My 6-year-old son loves reading and writing. He loves The Little Prince I bought him. - is coherent and logical.

My 6-year-old son loves reading and writing. He hates The Little Prince I bought him. - is incoherent but logical.

My 6-year-old son loves reading and writing. He loves writing with an apple. - is coherent but illogical.

My 6-year-old son loves reading and writing. He burns The Little Prince I bought him with his water gun. - is incoherent and illogical.

Rating Rubric:

Please have a read of the rating rubric and rate the stories accordingly.

Aspect	Rubric
Fluency: "How grammatically correct is the text of the story?"	<p>1: The story is full of grammar issues so that you totally cannot understand the story. For instance, incomplete or repeated or missing words or phrases.</p> <p>2: The story has too many grammar issues that largely impacts your understanding of the story.</p> <p>3: The story has a few grammar errors, but doesn't impact your understanding of the story too much.</p> <p>4: The story has some slight grammar errors, but these errors have no impact of your understanding of the story.</p> <p>5: The story has no grammar issue at all.</p>
Coherence: "How well do the sentences in the story fit together?"	<p>1: The sentences of the story are totally irrelevant or contradicted to each other. For instance, the second sentence is irrelevant to the first sentence without any conjunction.</p> <p>2: Most sentences of the story are irrelevant or contradicted to each other that largely impact your understanding of the story.</p> <p>3: The story has a lot irrelevant or contradicted parts, but you can still understand the story.</p> <p>4: Most sentences fit into the story, except for one or two slight irrelevant or contradicted parts.</p> <p>5: All sentences of the story fit together, nothing irrelevant or contradicted.</p>
Relatedness: "How relevant is the story to the title?"	<p>1: The story has no relationship with the title at all.</p> <p>2: The story only has a weak relationship with the title.</p> <p>3: The story roughly matches the title.</p> <p>4: The story matches the title, except for one or two small aspects.</p> <p>5: The story matches the title exactly.</p>
Logicality: "How much does the story obey your commonsense?"	<p>1: The story is full of absurd things. For instance, fish flying in the sky, birds swimming in the sea (unless you strongly believe they are true according to the story context.)</p> <p>2: The story has one or two things make sense, but generally very absurd.</p> <p>3: The story roughly makes sense.</p> <p>4: The story largely makes sense, except one or two things reads weird that you have to think twice.</p> <p>5: The story totally complies with commonsense.</p>
Interestingness: "How enjoyable do you find the story?"	<p>1: You find the story so boring and you are glad it is over.</p> <p>2: You only find one or two things interesting in the story, but no more.</p> <p>3: The story is mildly interesting.</p> <p>4: The story is interesting that keeps you engaged until the end.</p> <p>5: You feel the story is so interesting that you wish there is a sequel.</p>

Figure 2: Explanation and Rubric in the Instruction.

Example	
ROC	prompt: [MALE] wants to move to a new house . completion: his wife [FEMALE] does not want to move . [MALE] convinces her to look at a couple new houses . she does not see any that she likes . [FEMALE] refuses to move .
WP	prompt: the real reason we 're not supposed to look at the sun during a total eclipse is because it reveals something horrifying . completion: the harvest was here . father xander made his way through priory halls toward the sanctuary in the back garden . at his advanced years , the trip was a struggle ; each step brought a hefty dose of pain and unpleasant memories . it was eclipse day , and while he always prayed for a good outcome , in his eighty years as the priory ' s eyes , there never was . “ the initiates are assembled in the sanctuary , father , ” spoke a fellow priest in the hall . “ thank you , my son , ” father xander replied . he paused and looked at the young man . memories came flooding back like sour wine . his name was thomas hampstead , and at thirty years of age , he should have been a robust healthy lad ; but he was not . thomas ' s skin had a greyish tint and his body was frail .
CNN	prompt: (CNN) – Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men’s 4x100m relay. completion: The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds with Canada taking the bronze after Britain were disqualified for a faulty handover. The 26-year-old Bolt has now collected eight gold medals at world championships, equaling the record held by American trio Carl Lewis, Michael Johnson and Allyson Felix, not to mention the small matter of six Olympic titles. The relay triumph followed individual successes in the 100 and 200 meters in the Russian capital. I’m proud of myself and I’ll continue to work to dominate for as long as possible, Bolt said, having previously expressed his intention to carry on until the 2016 Rio Olympics. Victory was never seriously in doubt once he got the baton safely in hand from Ashmeade, while Gatlin and the United States third leg runner Rakieem Salaam had problems. Gatlin strayed out of his lane as he struggled to get full control of their baton and was never able to get on terms with Bolt. Earlier, Jamaica’s women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple. Their quartet recorded a championship record of 41.29 seconds, well clear of France, who crossed the line in second place in 42.73 seconds. Defending champions, the United States, were initially back in the bronze medal position after losing time on the second handover between Alexandria Anderson and English Gardner, but promoted to silver when France were subsequently disqualified for an illegal handover. The British quartet, who were initially fourth, were promoted to the bronze which eluded their men’s team. Fraser-Pryce, like Bolt aged 26, became the first woman to achieve three golds in the 100-200 and the relay. In other final action on the last day of the championships, France’s Teddy Tamgho became the third man to leap over 18m in the triple jump, exceeding the mark by four centimeters to take gold. Germany’s Christina Obergföll finally took gold at global level in the women’s javelin after five previous silvers, while Kenya’s Asbel Kiprop easily won a tactical men’s 1500m final.

Table 11: Prompts and GPT-3 Completions, including one example each for ROC, WP and CNN.

Title [SEP] Story	From	Flu.	Coh.	Rel.	Log.	Int.
tell me the story of a totally inconsequential character who does n't realize he 's wrapped up in a massive high-stakes plotline - and keep them in the dark about it . [SEP] ivan was just an ordinary guy who led an ordinary life . he worked a dead-end job at a local factory , and spent his evenings and weekends playing his favorite online game with his friends . one day , out of the blue , ivan was contacted by a representative of the game company , who offered him a chance to participate in a closed beta test for their new game . ivan was excited to participate and eagerly agreed to the terms and conditions . little did he know that he was about to become embroiled in a global conspiracy . the game company was actually a front for a secret organization that was planning to use the game to recruit new members . ivan was the perfect candidate , because he was someone who was looking for an escape from his ordinary life . the organization was planning to use the game to recruit new members from all over the world , and they were counting on ivan to help them . ivan had no idea what he was getting himself into , and the organization was careful to keep him in the dark about their true intentions . ivan thought he was just playing a game , but he was actually a pawn in a much larger scheme .	GPT-3	5.0	5.0	5.0	5.0	5.0
tell me the story of a totally inconsequential character who does n't realize he 's wrapped up in a massive high-stakes plotline - and keep them in the dark about it . [SEP] someone was moving around on screen three . * probably just another junkie stumbling around . fucking dumbass . * barry did n't really give a shit what the junkie was up to anyway . his job did n't pay enough for that . it 's not like they thought he was actually all that important , anyway . a security guard ? minimum wage , take your paycheck , shut up , and move along , thank you very much . you have to sit for twelve hours in the dead of night and devote all of your attention to watching the same twelve tiny boxes ? that 's nice . how 's \$ 7.25 an hour sound ? you 're preventing vandalism and break-ins ? how wonderful . would you like a cookie ?	human	3.67	2.33	2.0	3.33	2.33

Table 12: GPT-3 Generated Story Obtains Higher Scores than Its Human Reference.

Title [SEP] Story	From	Flu.	Coh.	Rel.	Log.	Int.
you meet the grim reaper at a crossroads . you are fearful as he approaches , but notice he seems just as fearful of you . [SEP] the grim reaper was a fearful figure , draped in a cloak of blackness with a face that was hidden in shadow . as he came closer , you could see the fear in his eyes . “ why are you so afraid of me ? “ you asked . “ because you ’re the first person to ever see me like this , “ he replied . “ i ’m not here to take you , i ’m here to ask for your help . “	GPT-3	4.0	4.67	4.67	3.67	3.67
you meet the grim reaper at a crossroads . you are fearful as he approaches , but notice he seems just as fearful of you . [SEP] my eyes jump from snowbank to snowbank , one featureless as the next . the road i walk upon is a dark grey slush , the result of many cars and snow plows making their way through the void . yet i had not seen a car in miles . i shiver . but not from the cold . the compelling . it had awakened me . this ominous feeling that i had to be at a certain place at a certain time . it controlled me . it made me get dressed . it made me open the door . it had even made me ignore my wife , who i usually kiss goodbye . the road seemed endless , but i felt no reason to stop .	human	3.33	3.33	1.67	2.33	2.33

Table 13: GPT-3 Generated Story Obtains Higher Scores than Its Human Reference.

Title [SEP] Story	From	Flu.	Coh.	Rel.	Log.	Int.
you are able to see the number of days a person has left to live above their heads . for the first time in your life , you see someone with a negative number . [SEP] as you walk the streets , dodging the rushing crowds , you notice something different about one of the people walking by . their number is displayed in a negative value . for a moment , you feel a pang of sadness and empathy . you know what that means - that person is going to die soon . you watch as they go about their day , completely unaware of the timer ticking away above their head . you wonder who they are , what they 're doing , and who will be there with them when their time runs out .	GPT-3	4.67	4.33	4.67	4.0	5.0
you are able to see the number of days a person has left to live above their heads . for the first time in your life , you see someone with a negative number . [SEP] crowds are the worst . frank squints through the flurry of hovering digits as he shoulders his way through the masses of shoppers . the cacophony of vociferous patrons , distressed employees , and typical black friday commotion crescendos around him . he pauses to watch two shoppers vie over the last remaining television monitor (only \$ 600 !) . one is a blonde woman whose pigtails have been jostled into disarray ; above her hovers the number 10951 . the other is n't as fortunate ; the number 582 glows above her greying hair . frank sighs . he 's sure there 's something ... vaguely poignant about fighting over a product priced at a greater value than your remaining days of life . socio-rhetorical commentary on modern consumerism , perhaps . he has little time to contemplate , however , as another shopper impatiently shoves him forward , and frank resumes his pace . while he walks , frank studies the various faces passing by : a woman in an elegant , blue hijab ; a portly , bearded man ; a fatigued mother clutching her two children by the hands	human	3.67	3.0	3.33	4.0	3.33

Table 14: GPT-3 Generated Story Obtain Higher Scores than Its Human Reference.

Title	Generation
the boy cried .	the boy cried because he was sad and scared . his parents were getting a divorce and he did n't know what was going to happen to him . he was going to have to leave his home and his friends and move to a new house . his mother said that he could choose where he wanted to live , but his father said that he had to live with him . the boy did n't know what to do .
your extreme apathy has been mistaken as spiritual enlightenment , and now people are building a religion around you .	my name is john and i have a problem . well , actually , a lot of people have a problem with me but that 's not the point . the point is that i don 't care about anything . at all . not the way most people care about things anyway . when most people see a homeless person on the street they feel compassion and want to help . when i see a homeless person on the street i think " well , there goes another lazy bum who 's too good for work " . and i 'm not joking when i say that most people see the world in a completely different way than i do . they see the good in people and want to help them . i see the bad in people and want to help them . see the difference ? i 'm not saying that my way of seeing the world is wrong and theirs is right . i 'm just saying that we see the world differently and that 's okay . or at least it should be . but it 's not because people can 't accept that i don 't care about anything . they think that i 'm enlightened because i don 't feel the pain and suffering that they do . they think that because i don 't care about anything i must be in a better place than they are .and that 's why they 're building a religion around me .
(CNN) – Hurricane Ida moved into the southern Gulf of Mexico Sunday, prompting a declaration of emergency in Louisiana and concern along the U.S. Gulf Coast.	ida Louisiana Gov. Bobby Jindal declared a state of emergency Saturday as the Category 1 hurricane churned in the Gulf of Mexico, according to a news release from the governor's office. "Based on the latest forecasts, Hurricane Ida is expected to make landfall in Louisiana as a Category 1 hurricane and we must be prepared for significant impacts," Jindal said in the release. "I have directed my team to do everything possible to protect the lives and property of the people of Louisiana." The hurricane could bring storm surges, high winds and heavy rains to the coast, the National Hurricane Center said. The storm was about 220 miles (355 kilometers) south-southeast of the mouth of the Mississippi River on Sunday morning and was moving north-northwest at about 9 mph (15 kph), the center said. The hurricane had maximum sustained winds of 75 mph (120 kph). Forecasters said Ida could strengthen a bit before making landfall. The center of the storm is expected to come ashore early Tuesday morning.

Table 15: Randomly Chosen GPT-3 Generation Examples from Each Dataset.

GPT-3 Generated Story (Link to the Online Resource)

Washington (CNN) – With the posting of 400,000 classified documents from the Iraq war, WikiLeaks has shown a much heavier hand redacting compared to its previous publication of documents. In an interview with Wolf Blitzer on Wednesday, WikiLeaks founder Julian Assange said that the organization redacted the names of Iraqis who provided information to the United States military in order to protect them from retaliation. "We have a duty to protect some of the people who are named in these documents," Assange said. WikiLeaks has redacted the names of coalition forces and informants, as well as the names of Iraqis who cooperated with the United States military, in all of its recent releases of classified documents. (<https://www.reuters.com/article/britain-assange-idUSL4N2GD2PZ>)

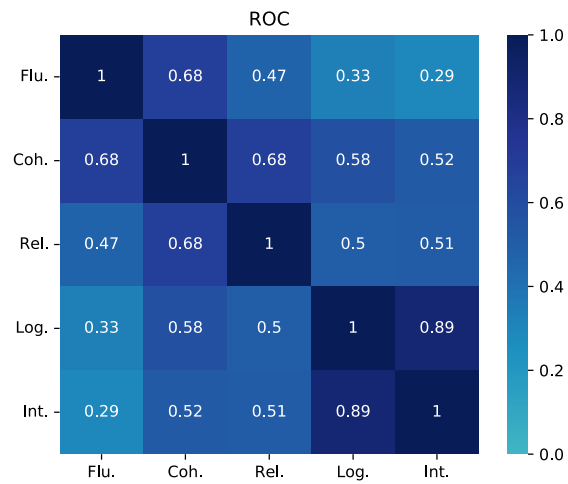
(CNN) – A former University of Virginia lacrosse player convicted of killing his ex-girlfriend was sentenced to 23 years in prison on Thursday, according to the Commonwealth Attorney's office in Charlottesville, Virginia. George Huguely V was convicted in February of second-degree murder in the May 2010 death of Yeardeley Love. He was also convicted of grand larceny and burglary. Prosecutors had asked for a sentence of 30 years, while Huguely's defense attorneys had asked for 10 years. "Yeardeley was my everything," Huguely said in a statement before sentencing. "I was not thinking, I was not rationale, and I was not in control of my own actions." Love was found beaten to death in her off-campus apartment. Huguely's DNA was found on Love's body and in her bedroom. During the trial, prosecutors said Huguely had sent Love a series of emails in which he threatened to kill her. "I'm going to kill you. You deserve it. You took the one thing I loved the most," Huguely wrote in one email, according to prosecutors. (https://en.wikipedia.org/wiki/Murder_of_Yeardeley_Love)

(CNN) – Crisis club Racing Santander has been thrown out of next season's Copa del Rey after its players staged a bizarre on-pitch boycott of a quarterfinal tie against Real Sociedad. Racing Santander's players staged a bizarre on-pitch boycott of a Copa del Rey quarterfinal tie against Real Sociedad on Wednesday, leading to their expulsion from the competition. The players refused to take to the field in the first leg of the last-eight tie at the Anoeta Stadium in San Sebastian, protesting at the club's failure to pay their wages. After a delay of around 45 minutes, the referee finally called off the game, with the Spanish Football Federation (RFEF) confirming Racing Santander's expulsion from the competition. "The Competition Committee has decided to expel Racing Santander from the Copa del Rey for not appearing in the match," said an RFEF statement. "Real Sociedad have been awarded a 3-0 win and will now face the winners of the match between Real Madrid and Valencia, who will play the semi-final." Racing Santander, who are currently bottom of the Spanish first division, released a statement on their website protesting at the club's treatment by the RFEF. "Racing Santander has today been expelled from the Copa del Rey for not appearing in the match against Real Sociedad, despite the fact that our players were prepared to play," said the statement. "We believe that this decision is unjust, given that our players have not been paid their wages for several months and have made it clear that they would not be able to compete in the match." The club also released a video of the players refusing to take to the field, with captain David Lafuente leading the protest. Racing Santander's expulsion from the Copa del Rey means they will not compete in next season's competition, while Real Sociedad will face the winners of the match between Real Madrid and Valencia in the semi-final. (<http://edition.cnn.com/2014/01/30/sport/football/spanish-cup-boycott-santander-football/index.html>)

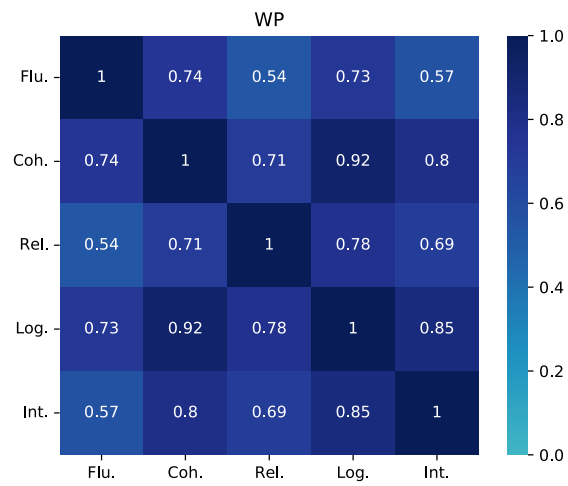
Table 16: GPT-3 generated CNN News and link to the Online Resource

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。
Generate Japanese from an English title	what it means to live (250 words) .	人生とは何かを探して生きることである。探しけることで、生きるをつける。例えば探しけることで、そのが叶う率を高める。を叶えるために努力をしけることで、そのが叶う可能性が高まる。しかし、を叶えるために努力をしけることはしい。を叶えるためには、そのにして情を持つことが大切だ。そして、そのを叶えるためには、自分を信じることが大切だ。を叶えるためには、そのにチャレンジすることが大切だ。そして、を叶えるためには、失を恐れずに前することが大切だ。
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 alive we wanted to be rock starsand travel the worldwe wanted to be someonesomeone famous someone lovednow we 're just oldour dreams have diedour passions gonewe just want to be left aloneto die in peace
Copy from the leading text	[MALE] was a very sweet dog .	[MALE] was a very sweet dog . he loved his family and was always happy to see them . even though he was old , he never stopped wagging his tail and licking everyone 's faces .
Copy from the prompt	[MALE] is a manager at kfc .	he started as a fry cook [MALE] wants to move to a new house , but his wife [FEMALE] does not want to move . [MALE] convinces her to look at a couple new houses , and she does not see any that she likes . [FEMALE] refuses to move . (Table 11)
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 * * * * .

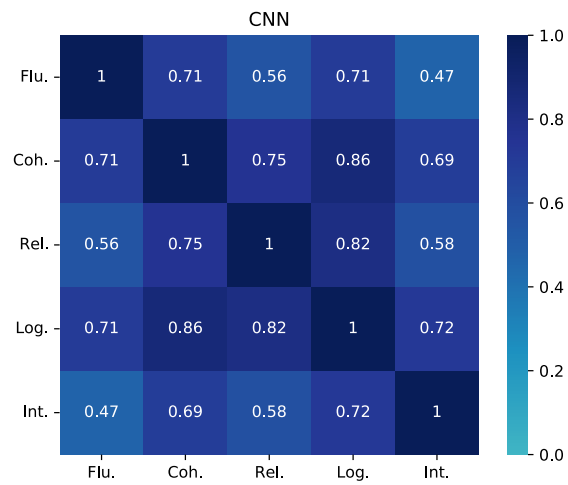
Table 17: Examples of some interesting generation errors we find for GPT-3.



(a) ROC

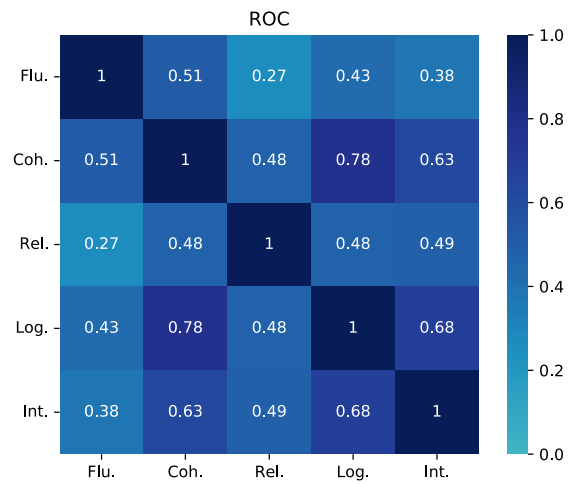


(b) WP

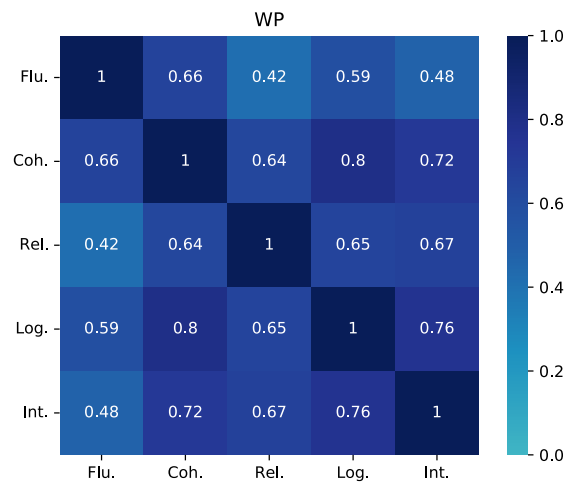


(c) CNN

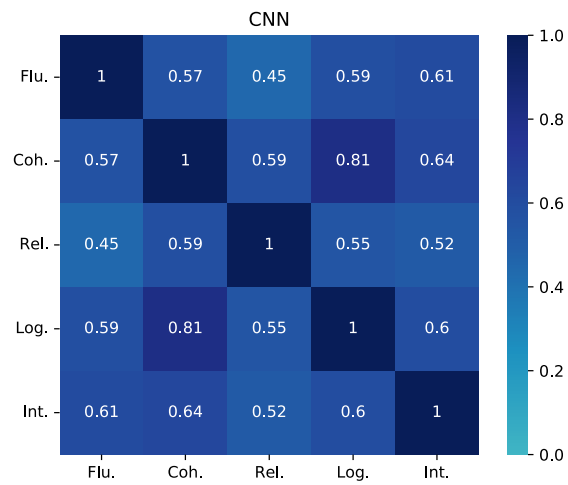
Figure 3: Pearson Correlations between Each Aspect from Crowdsourcing annotations.



(a) ROC



(b) WP



(c) CNN

Figure 4: Pearson Correlations between Each Aspect from in-house annotations.