Correspondence Between Perplexity Scores and Human Evaluation of Generated TV-Show Scripts

Pia Keukeleire[1] , **Tom Viering**[1] , **Stavros Makrodimitris**[1] , **Arman Naseri Jahfari**[1] , **Marco Loog**[1] , **David Tax**[1]

[1]TU Delft

P.Keukeleire@student.tudelft.nl, {T.J.Viering, S.Makrodimitris, A.Naserijahfari, M.Loog, D.M.J.Tax}@tudelft.nl

Abstract

In recent years many new text generation models have been developed while evaluation of text generation remains a considerable challenge. Currently, the only metric that is able to fully capture the quality of a generated text is human evaluation, which is expensive and time consuming. One of the most used intrinsic evaluation metrics is perplexity. This paper researched the correspondence between perplexity scores and human evaluation of scripts for the TV-show *Friends* generated using OpenAI's GPT-2 model. This was done by conducting a survey taken by 226 participants that evaluated selected scripts on creativity, realism and coherence. The survey results revealed that generations with a perplexity value close to that of an actual *Friends* script perform best on creativity, but score low on realism and coherence. The most realistic and coherent generations were those with a lower perplexity value, while the worst in all fields were the generations with the highest perplexity value. The research shows that perplexity is not an adequate measure for the quality of generated TV-show scripts.

1 Introduction

Natural Language Generation (NLG) is a quickly advancing field that is finding more and more new applications. NLG consists of technologies that produce new and coherent text when given an input. Examples of popular applications of text generation are chatbots [6], automated summarization [1] and automatic image description [3]. This research focussed on the generation of TV-show scripts, and more specifically on the generation of lines for the sitcom *Friends*.

At present, models exist that are successful at generating coherent text that appears to be written by a human. The best known example of such model is OpenAI's GPT-2 [11]. This model has been used to create poetry [4], generate movie scripts [10], university papers [12] and much more. This is the model that will be used in this research to generate TV scripts for *Friends*. GPT-2, which stands for "Generative Pre-trained Transformer 2" [2] is a transformer-based language model that is trained using a very large dataset obtained from webpages.

The main challenge in text generation is finding a fitting intrinsic evaluation metric. The main reason for this is that the quality of a text is defined by many aspects. The only way to currently guarantee the quality of each of these aspects is by performing a manual evaluation [14]. Finding a well-fitted evaluation metric is very important; to assess a model, improve optimization of the model, and to enable a comparison between models.

A suitable evaluation metric for generated TV-shows would mean that models would gradually improve at generating TV-shows. This would enrich and renew the world of TV entertainment and offer a new tool for increasing a writer's creativity. One metric that is often used for evaluation is perplexity, which measures a model's certainty of its predictions. Since most models, like GPT-2, are trained to fit the next word probability distribution to that of the training data, perplexity is a logical choice for measuring how well this was done. The question then still remains whether optimizing the perplexity also results in generations of better quality. Ziegler et al. [15] finetuned language models by training them using reinforcement learning based only on a human judgement of the quality of the text. This strategy, although interesting and well-performing, is application specific and time consuming. They finetuned the model to specifically work well for the summarization of CNN/Daily Mail news articles, therefore it is hard to repeat this process for each application. There are multiple advantages to using perplexity; calculating perplexity doesn't require human interference and is easy and straightforward. Understanding perplexity values and where they come from is relatively uncomplicated and, additionally, it's easy to optimize a model for an improved perplexity score.

Although perplexity is used in many papers to assess a model's performance, according to Huyen [8] the relationship between the perplexity and how well the model performs on downstream tasks is rarely published, and the correlation between both is often not researched.

This research investigated how the perplexity scores of dialogues for the TV-show *Friends* generated by the GPT-2 model correspond to human ratings of these scripts.
In order to do this, an answer had to be found to what range of perplexity scores would be interesting for the research, which changes in parameter values should be made to the model in

order to obtain different perplexity scores, which aspects of the dialogues should be evaluated and which questions should be asked to humans in order to assess these aspects.

The paper is structured as follows. Section 2 will explain the theory behind perplexity and human evaluation and, furthermore, specify and motivate the research hypotheses. The experimental setup of the research is described in section 3, which explains the steps and technical details of the generation and evaluation processes. The perplexity scores and human evaluation results are described in section 4, which is followed by section 5 about the ethical aspects and the reproducibility of the research. Limitations and additional remarks on the research will be discussed in section 6. Finally, section 7 will contain the conclusion and the future directions of the research.

2 Evaluation of Generated TV-Show Lines

The first step of the evaluation entailed a calculation of the generations' perplexity values. Next, a human evaluation was done so that the perplexity value of a generation could be related to its actual quality. This section contains a theoretical explanation of both methods and a comparison.

2.1 Perplexity

To calculate the model's loss, GPT-2 uses the cross-entropy between the training data and its predictions. The cross-entropy H measures how close the probabilities in two distributions P and Q are [8]. For two discrete distributions over words $x \in \{x_0, x_1, \dots x_N\}$, this is mathematically defined as:

$$H(P,Q) = -\sum_{x} P[X = x] log Q[X = x]$$

P describes the word distribution of the actual data while Q equals the predicted output vector for the following word's probability given the previous words $W_1, ..., W_k$ [9]. This output vector is a probability distribution over the model's vocabulary and is defined as:

$$Q[W_{k+1} = w | W_1, ..., W_k]$$

Perplexity is a metric that measures a model's certainty of its prediction. Perplexity is calculated as follows:

$$PPL = 2^{H(P,Q)}$$

As a result, as the model trains to minimize the cross-entropy, it also aims to minimize the perplexity. A lower perplexity means that the predictions follow the probability distribution of the training data better.

2.2 Human Evaluation

To get a qualitative human evaluation it's important to focus on the evaluation of a few clearly set attributes. The text attributes that should be measured depend highly on the type of text that is generated; e.g. for health record summarization tasks the correctness of the information and the capturing of key information is of high importance.

The generated *Friends* scripts were evaluated on creativity, coherence and realism. Realism in this research is defined as a measure for how plausibly a generation could be a TV-show script. Creativity is the most subjective metric, resulting in the most differing answers. It is interesting to measure the creativity because dialogues that contain less information, e.g. a conversation in which each party says hello, are more quickly deemed realistic and coherent without being an interesting generation that captures the essence of a sitcom. The coherence of the script measures if the script makes sense and if the subject remains somewhat constant throughout the dialogue.

2.3 Perplexity vs. Human Evaluation

GPT-2 is optimized for predicting the highest next word probabilities and consequently to minimize the perplexity. Natural language, according to Holtzman [7], does not produce the most probable text and rarely remains in a high probability zone but instead prefers to use more informative tokens. GPT-2 offers decoding strategies that allow for different ways of random sampling from its probability distribution in order to bring variation in text and include lower probability words. It is often thought that aiming for generations with a perplexity value that approaches the model's perplexity of human written text gives the best results [7]. We would like to argue that such strategy would indeed give a variation in the generation similar to human written text, but that this does not result in a text that actually appears more human like. Our reasoning behind this is that lower probability words can hardly ever be randomly picked while still making sense. In natural language, low probability words are deliberately picked for their specific meaning, something which can not be achieved by random sampling. Our hypothesis is thus that perplexity values approximating those of the test set belong to generations that do not make much sense, while generations with a lower perplexity make sense but are too generic to appear in *Friends*. This automatically raises the question how useful perplexity is in the case of TV-show scripts.

GPT-2 does not recognize the structure in a text, but considering that the training data is very consistent in structure, in the way that each scene indication is between square brackets and each line starts with the name of the character followed by their uttering, it is most probable that the generations will retain this structure. But since the model only does this because the probability is extremely high that a sentence would have this structure, the structure will likely dissolve when the perplexity values become higher. The model does not know how to vary a character's line without also varying in the way that character is denoted or a scene is depicted.

3 Experimental Setup

This section explains in technical detail how the scripts were obtained and how the model was finetuned. It also explains the steps for generating different scripts and calculating the perplexities of these scripts. Finally, the setup of the human evaluation survey is given.

3.1 Generation

The data preparation, finetuning of the model and generation of the scripts was done using an adaptation of an existing

GPT-2 implementation for writing movie scripts.[1] This research used the medium-sized version of GPT-2, which consists of 345M parameters and uses 12 layers. [11]

The details of the process of generating the scripts are as follows, provided step-by-step.

The original scripts as obtained from Puneeth's GitHub[2] were all in HTML form. In order to strip them from their HTML, part of van Tussenbroek's code [13] was used. The data was split into a training set of 216 episodes and a test set of 20 episodes by taking the first two episodes from the first season to put in the test set, the second and third from the second and so on. Only from the tenth season, which is shorter than the other seasons and therefore prohibited following the pattern, the 13th and 14th episodes were used for testing. The reason for separating the test and training data this way is to guarantee an even amount of data from each season and a good division between beginning and end of season episodes. This way the test set accounts for the differences that exist between the different seasons and between beginning and end-of-season episodes (end-of-season episodes are generally a lot more serious).

Next, the training scripts had to be prepared in the right format for GPT-2. This was done using GPT-2's pretrained tokenizer. This tokenizer binds pieces of frequently occurring words to an ID. The training text was tokenized by converting all the word pieces in blocks of size 512 to their corresponding IDs. After that, special tokens (for example to indicate the end of a sentence) were added to the list of ids that represents the training data.

The finetuning was done in a Google Colab notebook, which provides a free GPU. To begin the finetuning, the training data was loaded in batches of size 1. The finetuning of the text consists of training the pretrained GPT-2 medium sized model again on the training data. This was done in 3 epochs with a learning rate of 0.00002. Next, the weights and configurations of the finetuned model were saved for later use. The vocabulary from the training data was saved to the tokenizer. For every 200 batches, a sample generation was printed to allow an interim evaluation to check if the model is training correctly.

The finetuned model was then used to generate three sample texts per generation of a maximum length of 300.

Since it might not always be the best choice to pick the word with the highest probability, the probability distribution can be manipulated by sampling. When sampling is used, the next word is randomly picked from its conditional probability distribution, which makes the model's generations non-deterministic [7]. The first sampling technique that was used in this research was setting the temperature of the probability model. Lowering the temperature corresponds to skewing the distribution towards higher probability events. The second method that was applied is top-k sampling. Top-k sampling regards the top k possible tokens that have the highest relative probabilities. The third and last method used in this research is nucleus sampling. For nucleus sampling, which is also known as top-p sampling, a threshold p has to be chosen which includes the smallest set of possible tokens whose sum possibility is equal to this threshold. In order to obtain wide variations in the perplexity of the generated scripts, the code looped through different temperature, top-k and top-p values to produce 80 different generations. For each of these generations, '[Scene: ' was used as input in order to start each sample with an indication of the scene. The set of temperatures, top-k and top-p values used for the generations were `[0.7, 0.8, 0.9, 1]`, `[0, 50, 200, 400]` and `[0.8, 0.85, 0.9, 0.95, 1]` respectively.

After evaluation, extra generations were made in order to approximate the average perplexity value of the test set. The generation that came closest and was therefore used in this research, was generated using a temperature of 1, top-p of 1 and top-k of 240. Finally, this research regards 81 different generations.

3.2 Perplexity Calculation

The perplexity values were calculated for each of the generated scripts, as well as for the test set of actual *Friends* scripts using once again a Google Colab notebook.

First, the finetuned model and tokenizer were loaded. Then, the input data was tokenized in blocks of length 512 using this tokenizer. For each input file, the loss of each block was calculated using the model, and their average was taken to obtain the perplexity value of that file. The loss corresponds to the cross-entropy of the input, so the perplexity of a block was obtained by taking two to the power of the loss value.

3.3 Survey Setup

To perform a human evaluation of the scripts, a survey was set up using Google Forms. In this survey, ten different scripts were presented to anonymous participants, two of which were parts of actual *Friends* episodes. The eight generations have five different perplexities.

The perplexity values of the generations that were chosen to be used in the survey are the highest occurring value, the lowest occurring value, and the one best approaching the average of the actual scripts' perplexities. Furthermore, two different scripts that appeared similar in quality to actual *Friends* scripts were cherry picked. These scripts have different perplexity values, which resulted in five different perplexity values overall.

A first inspection of the generations, however, showed that the generations with the lowest perplexity values were of no interest to this research, since they contained only repetitions of one up to five words. For this reason the generation with the lowest perplexity value that contained more than just repetitions was included instead. This turned out to be the generation corresponding to the perplexity value 1.647. Reading the generations also revealed that the seemingly best generations had a perplexity value around 2. The reason the two cherry picked scripts were included was to assess whether generations exist that can compete with the actual scripts, and to determine what perplexity values these generations have.

For the sake of simplifying further explanation, the followings names will be used for the different sets of generated scripts. The two generated scripts with the lowest perplexity

[1] https://github.com/cdpierse/script_buddy_v2

[2] https://github.com/puneeth019/FRIENDS

score will be referred to as the 'low PPL set', and in the same way the set with the highest perplexity score will be called the 'high PPL set'. The two scripts that have a perplexity value approaching that of the test set will be called the 'similar PPL set', while the two cherry picked scripts will simply be referred to as the 'cherry picked set'.

Finally, the perplexity values of the scripts included in the survey are 1.647 for the low PPL set, 2.259 and 1.903 for the cherry picked set, 5.220 for the similar PPL set and 7.854 for the high PPL set.

Apart from the cherry picked scripts, the script parts that would be included in the survey had to be chosen as objectively as possible. Each generation corresponding to a chosen perplexity value consisted of three scripts of varying lengths, which brought some difficulties in remaining objective. Picking the scripts completely randomly would destroy the structure. Instead of randomizing, the first and last part to start with a scene indication were picked. The scripts used in the survey can be found in appendix A. From the test set of actual scripts, the first part of the first episode and the first part of the last episode were included.

The survey started by inquiring if the participant had watched *Friends* before, for which the possible answers were "yes", "no", and "partially/I don't remember much". This question was included because people who had watched *Friends* before might give stricter scores to the generations, since they can more accurately compare them to actual scripts. For the same reason, it was expected that they would attribute higher scores to actual *Friends* scripts.

Next, the ten chosen scripts were shown on individual pages where the participant had to attribute a score on the creativity, realism and coherence of the text. The evaluation had the form of a Likert scale, where each point on the scale corresponded to somewhere between very bad on the far left, and very good on the far right.

The survey was anonymous because there was no need for any personal data from the respondent, but this did open up the possibility for a user to retake the survey. As people are more likely to take an anonymous survey, we placed priority on obtaining a larger amount of answers over eliminating the possibility to tamper with the results.

4 Results

This section describes the results obtained from the research. First, the perplexity values of the generated texts are given in section 4.1. Next, the results of the survey are given and discussed in section 4.2.

4.1 Resulting Perplexity Values

The perplexity values of the generated files are between 1.129 and 7.854, with an average of 2.894. The perplexity values of the test set are between 4.877 and 5.943, with an average of 5.409. The distribution of the generations' perplexity values is given in Figure 1.

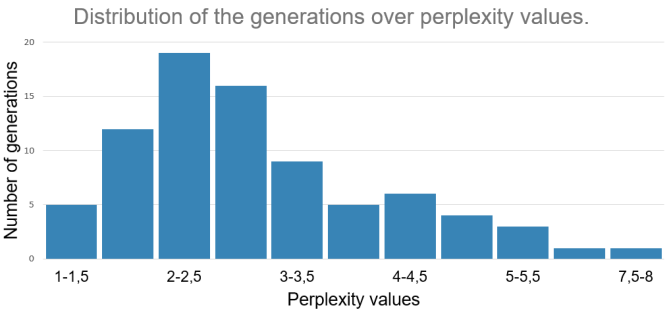


Figure 1: The resulting perplexity values of all 81 generations vary between 1.129 and 7.854.

4.2 Survey Results

The results follow from the answers of 226 survey participants. answers can be found in appendix A.

In order to compare the different scripts, the Likert scale is converted to a numeric scoring system as follows:

Very Bad	0
Bad	2.5
Okay	5
Good	7.5
Very Good	10

The scores are averaged for each script over the amount of responses.

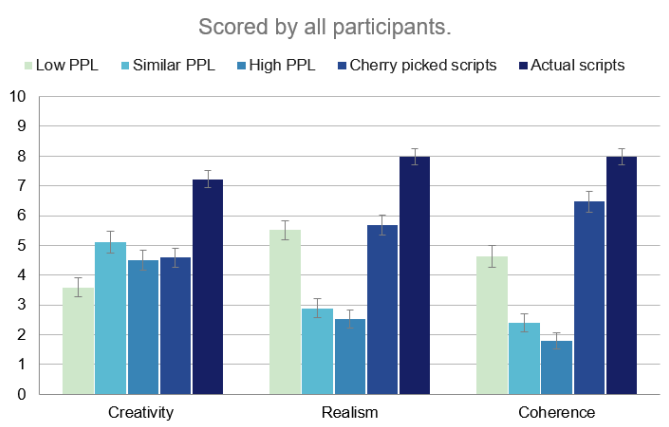


Figure 2: Results for each set of generations scored by all participants. The error bars display the 95% confidence interval.

The results proceeding from all survey answers can be seen in Figure 2. As expected, the two actual scripts from the test set scored highest on creativity, realism and coherence. From the generations, a script from the cherry picked set scored highest on creativity, as can be seen in appendix A.3 table 1. The average creativity score of the similar PPL set scored highest compared to the averages of the other sets. The cherry picked set scored highest on both realism and coherence. This shows that the best scripts are not the scripts from the similar

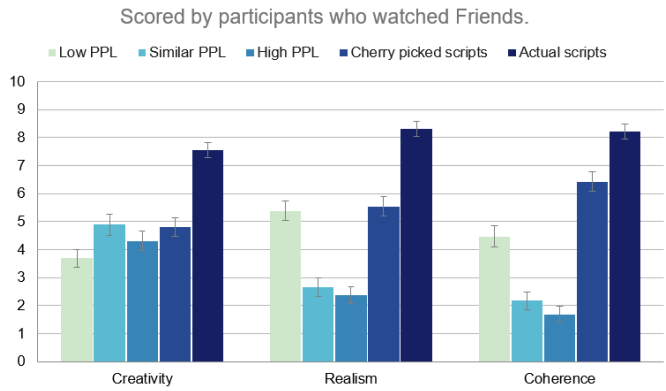


Figure 3: Results for each set of generations scored by participants who watched *Friends*.The error bars display the 95% confidence interval.

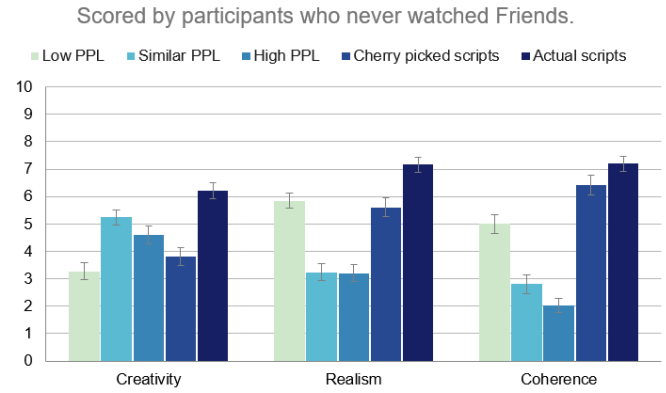


Figure 4: Results for each set of generations scored by participants who didn't watch *Friends*.The error bars display the 95% confidence interval.

PPL set, which mathematically seemed the most logical outcome, but scripts with a lower perplexity (around 2). These scripts still aren't comparable in realism and coherence to actual *Friends* scripts, since there is a rather big gap of 2.14 points in realism between the highest scoring generated set's average and the actual scripts' average, and a gap of 1.52 points in coherence. It does appear that scripts from the similar PPL set are on average seen as the most creative generations, in contrast to the expectations that the high PPL set would win on creativity. The reason it did not is probably because these generations were too random and senseless to be called creative. This set also scored the lowest on both realism and coherence. Although it was expected that generations with a high perplexity value would lose structure, this was not really the case. The only distortion in structure that occurred was the introduction of random people that do not appear in *Friends*.
On average, the low PPL set scored significantly better on both realism and coherence than the high and similar PPL set averages. The low PPL set's perplexity value is after all very close to the perplexities of the cherry picked scripts that scored on average highest on realism and coherence. But, again as expected, the low PPL set scored the worst on creativity. The generations with the lowest perplexity scores can be compared to the text generated when you keep on tapping a smartphone's word suggestions, where the most probable words are shown, which clarifies its lack of creativity.

As expected, there is a bias in the results of participants who have watched *Friends* before. The scores of the actual scripts given by participants who have watched the show are on all criteria about 1 point higher than the scores given by participants who have not, and about half a point higher than participant who are somewhat familiar with the show. Since 61.5% of the participants have seen *Friends* before, the results are highly biased. As can be seen in appendix A.3 table 3 There is only a difference of 0.47 between the highest scoring generation and the actual scripts' average realism score when only the results of participants who have not seen *Friends* are regarded, from which we can conclude that people who've never seen *Friends* before rate a good generation almost as realistic as an actual script.

5 Responsible Research

This section discusses data manipulation and the reproducibility of the methods.

Data manipulation was avoided as much as possible, but at some points in the research it was necessary. First of all, as described in subsection 3.1, none of the initial generations had a perplexity value close to the test set's average value. About ten more generations were made using slightly different parameters in order to obtain a generation with a satisfactory perplexity value. From these extra generations, only the generation with the most suitable perplexity value was used in the research. Another part of the research where a form of data manipulation had to be applied was in the survey setup. As described in section 3.3, cherry picking was avoided while selecting the scripts, except when the purpose was to cherry pick the two best scripts. Still, it was impossible to pick the other scripts completely randomly because of the structure of the texts, so the used samples were systematically taken from the generations.
The survey data could have been manipulated by participants who took the survey twice, but since this would not give them any benefits, this is very unlikely.

The largest part of the code was reused from Pierse's GitHub[3] with movie script generating code, which is well clarified in his article [10]. All extra steps taken in this research and all parameter values are given in this paper. It should be noted that it would be hard to recreate the exact same generations, because generating text twice with the methods in this research will give different texts (due to the random sampling of the language probability model). Generations with a perplexity value similar to those of the scripts used in the research would be of similar quality and would therefore give very similar results in a human evaluation.

[3]https://github.com/cdpierse/script_buddy_v2

6 Discussion

This section attempts to put the results in a broader context and discusses the limitations of the research.

The research showed that perplexity is not an adequate metric for the quality of generated TV-show lines, but that does not mean it is useless for this purpose. The survey results showed that, in order to get the best results, one should aim for lines with a low perplexity value, while keeping in mind that the lines generated would not be very interesting. More interesting results may be obtained by combining human written text with generations of a lower perplexity. This might result in a more interesting and coherent text. Instead of generating, the model could also be used to complete a given part of a TV-show dialogue.

The research was limited to only five different perplexity values, while the correspondence between the other perplexity values and human evaluation remains unknown. If one's purpose is to discover a trend in the quality of a script over a changing perplexity value, further research would be required.

It is important to note that this research involves only TV-show scripts, and that completely different results may be possible if it involved a different generation task. It is probable that perplexity does not work well with the dialogue form of the scripts, and that a generation in prose form would be better evaluated if it had a perplexity value similar to that of human written text.

This research does not clarify if GPT-2 can generate *Friends* scripts that could replace an actual script. Only a limited amount of generation techniques were applied with the goal to obtain differing perplexity values, and not to obtain the best generation possible. For the survey, two scripts that seemed to be the best candidates to appear in *Friends* were cherry picked. This was partially done to investigate if GPT-2 can generate seemingly real *Friends* scripts, but since this did not appear to be the case from the survey results, no answer can be given to this question. Additionally, these two scripts were subjectively picked by just two people, so the possibility exists that there were better candidates in the remaining generated scripts.

The survey accounted for the possible bias that may exist when a participant has seen *Friends* before. Other biases were also possible, for example from native English speakers who are better at judging the quality of English texts, or computer scientists who might be too actively looking for signs of being computer generated, but these biases were not relevant enough to include in the survey.

7 Conclusions and Future Work

This paper researched the correspondence between perplexity scores and human evaluation performed on scripts generated for the TV-show *Friends* using OpenAI's GPT-2 model. It was important to obtain varying perplexity values that were interesting for examination. In order to achieve this, the generations were made using different sampling methods using a varying top-k, top-p and temperature parameter. From all the generations, the ones with the lowest perplexity, the highest perplexity, and the perplexity closest to that of the test set, together with two cherry picked generations and two actual *Friends* scripts were assessed in a survey. The survey was taken by 226 participants who were first asked for their familiarity with the show and next to assess each script on its creativity, coherence and realism.

The relationship between perplexity values and human evaluation is as follows; generations with perplexity values lower than the test set's average perplexity value score better on coherence and realism, while they score worse on creativity than generations with a higher or equal perplexity value. The generations with a perplexity value closest to the test set's average perplexity value score best on creativity.
Perplexity captures the variation in text well, but not the quality. There is a big trade-off in creativity and coherence/realism. None of the generations scored as high in any field as the actual scripts.

In the future it would be interesting to examine how GPT-2 could be optimized for writing good TV-show scripts. It might also be worthwhile to attempt this with GPT-3. GPT-3 is the successor of GPT-2, which was released halfway through this research [5].

Perplexity as a metric on its own is not sufficient. For TV-show scripts, measuring the resemblance of character personalities and the topicality of the texts might be more relevant. Besides, the importance of perplexity highly depends on the generation end-task, as the perplexity value may say more about e.g. text summarization than it does about TV-show script generation. In the future, it would be interesting to study the relationship between perplexity and human evaluation for multiple generation tasks.

References

[1] Abualigah, L., Bashabsheh, M., Alabool, H., and Shehab, M. (2020). *Text Summarization: A Brief Review*, pages 1–15.

[2] Balodi, T. (2019). OpenAIs GPT-2 (Generative Pre-Trained Transformer-2) :"AI that is too Dangerous to Handle.". https://www.analyticssteps.com/blogs/openais-gpt-2-generative-pre-trained-transformer-2-ai-that-is-too-dangerous-to-handle. Accessed on 2020-05-08.

[3] Bernardi, R., Çakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *CoRR*, abs/1601.03896.

[4] Branwen, G. (2019). Gpt-2 neural network poetry. https://www.gwern.net/GPT-2. Accessed on 2020-04-28.

[5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[6] Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22:811–817.

[7] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

[8] Huyen, C. (2019). Evaluation metrics for language modeling. https://thegradient.pub/understanding-evaluation-metrics-for-language-models/. Accessed on 2020-05-10.

[9] Jiang, N. (2016). Perplexity vs cross-entropy. https://jiangnanhugo.github.io/2016/perplexity-vs-cross-entropy/. Accessed on 2020-05-13.

[10] Pierse, C. (2020). Film script generation with gpt-2: Fine-tuning gpt-2 to be a screenwriter using hugging face's transformers package. https://towardsdatascience.com/film-script-generation-with-gpt-2-58601b00d371. Accessed on 2020-04-28.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[12] Robitzski, D. (2020). This grad student used a neural network to write his papers: "you just can't expect a good grade.". https://futurism.com/grad-student-neural-network-write-papers. Accessed on 2020-05-08.

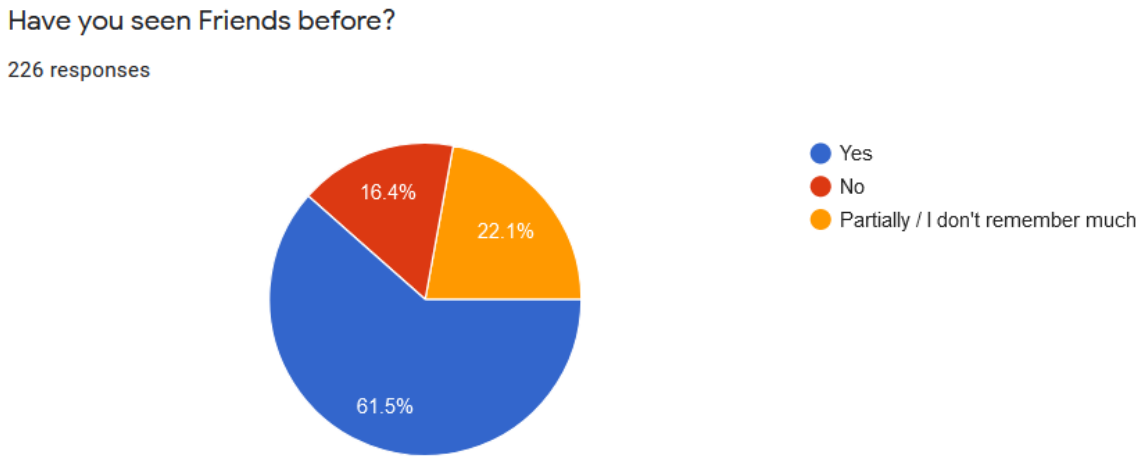
[13] van Tussenbroek, T. (2020). Who said that? comparing performance of tf-idf and fasttext to identify authorship of short sentences.

[14] Xie, Z. (2017). Neural text generation: A practical guide.

[15] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences.

A Survey Content and Results

A.1 Participants' Familiarity with *Friends*



A.2 Survey Scripts and Scores

Script 1

Perplexity Value: 1.90
Cherry Picked Set

Rachel: (entering) Ross!
Ross: Yeah, hi!
Rachel: So, what are you doing here?
Ross: Well, uh, I'm trying to raise money for charity.
Rachel: And what is that charity?
Ross: The Children's Hospital of Philadelphia.
Rachel: Really?
Ross: Yeah.
Rachel: Oh! And how much?
Ross: A lot!

Script 2
Perplexity Value: 5.22
Similar PPL Set

[Scene: Ross's apartment, he and Rachel are at Rachel's table.]
Ross: Hi.
Rachel: Yeah just wanna help Rachel out.
Ross: Uhhmmm well I have something that we could use if you guys aren't going to do anything about it. Or should I say nonstop acting. Rach.
Rachel: (Laughs) Not a problem but that is sort of a, that is a game you want to play with one of my underlings.
Ross: Nonstop acting? Rach that's a world of trouble.
Rachel: Augh that's not funny Ross!
Ross: (gets up) Yeah! (Takes her by her arm) Rach! (Shows her the actor card.) (Phoebe enters wearing underwear between her breasts and Rachel finds out that she is really really petite)
Phoebe: Ooh!
Ross: I came here to tell the truth! (Points to her chest) Rach is Phoebe. (She walks straight up to Monica with the card.) So, Rachel made a new commitment for her life and here it is I have just the one butts to throw away.

Script 3
Perplexity Value: 1.65
Low PPL Set

[Scene: Central Perk, Chandler and Joey are there.]
Chandler: (entering) Hey, guys.
Monica: (entering) Hi.
Chandler: Hey.
Monica: Hi. (They hug.)
[Scene: Central Perk, Ross is there with Monica.]
Ross: Monica!
Monica: Oh my God, Joey!
Ross: Oh my God, how are you guys doing?
Monica: I'm doing fine.

Script 4
Perplexity Value: 7.85
High PPL Set

[Scene: Rachel's apartment, Rachel is entering to shower.]
Rachel: Ahh! (She opens the door and jumps up and knocks loudly on the door.)
Monica: I'm having a shower!
Rachel: Pheebs both your hair plumper bitch!
Monica: Harbs 'em out!
Rachel: Soontoss? Listen, would you like to go to the loan committee and get the loan numbers set up so I can hold on to those?
Monica: Sure (Laughs ear to ear and walks away).
MFM: Why not?!
Rachel: He looks great! (Monica tries to get in) OG repay everything.
Monica: Come on! I think I'm in!

Script 5
Perplexity Value: 5.22
Similar PPL Set

[Scene: Monica and Rachel's, Rachel is finishing dressing and Monica is wrapping up this last few weeks of her honeymoon in preparation for the holidays.]
Joey: Great stuff. All the decorations are ready. You see where this is going. I still. I still. I still have to wait. I still have to see who is the one who? I still need to know.
[Scene: Rachels apartment. The previous night, she was making up her mind. Ross and Monica want to know if she's comfortable with sexual relationships.]
Ross: I don't know. I even asked you. [turns away pretending to not hear him] Damn. I just wish I hadn't used that the first time. All I ever know is what you believe.
Monica: All I know is a fling with that guy on top of Chandler is making me feel like a total whore.
Ross: Well it needs to be obvious that you like it.
Monica: I... I do. I really, really do. If this sounds like some kind of love story, I know what you're saying.
Joey: I'm just sayingin'. [Joey and Monica are about to walk up to Chandler in the hallway and drop the bomb. They turn to a guy close by and he, in turn, turns and backs away, keeping Ross in the picture.] After a century of this?

Script 6
Perplexity Value: 5.846
Actual Friends Script

[Central Perk, Chandler, Joey, Phoebe, and Monica are there.]
Monica: There's nothing to tell! He's just some guy I work with!
Joey: C'mon, you're going out with the guy! There's gotta be something wrong with him!
Chandler: All right Joey, be nice. So does he have a hump? A hump and a hairpiece?
Phoebe: Wait, does he eat chalk? (They all stare, bemused.)
Phoebe: Just, 'cause, I don't want her to go through what I went through with Carl oh!
Monica: Okay, everybody relax. This is not even a date. It's just two people going out to dinner and not having sex.
Chandler: Sounds like a date to me.

Script 7
Perplexity Value: 1.65
Low PPL Set

[Scene: Monica and Rachel's, Ross is still sitting there talking to her about his situation.)
Ross: Why did you not tell him?
Monica: I just can't believe that. I mean, he'd been such a great guy, and the thing was, how could you say no to him?
Ross: You're like, not sure what the big deal is?
Monica: I'm sorry, but I didn't want him to see the truth. It's just, I didn't want him to see how I'm feeling.
Ross: Oh, I'm gonna tell him. I promise.
Monica: Okay.
Ross: Y'know, I think maybe it's time.
Monica: No, I promise!
Ross: Oh, okay.

Script 8
Perplexity Value: 5.13
Actual Friends Script

[Scene: Chandler and Monica's apartment. They are having a diner party with Phoebe and Mike.]
Mike: (raising his glass) Thank you guys for having us over.
Phoebe: Oh! Yeah, this is fun, couples night.
Chandler: Yeah, I don't know why we hang out with married couples more often.
Monica: Well, because every time we do, you make jokes about swinging and scare them away.
Chandler: You mean that Portuguese couple? Yeah, like you wouldn't have done it. (she shrugs)
Ross: (entering) Hey, you guys... I have great news.
Monica: Ross, we're kind of in the middle of diner here.

Script 9
Perplexity Value: 7.85
High PPL Set

[Scene: Monica's apartment after Pete has all of the dogs shot deep.]
Petey: Are you guys still guys? Thanks to you guys for not shooting the dog! (They nod.) Uhm I guess they need my shoe! (They all go back to their seats and then sit down.)
Joey: So, what are we gonna do? Tell in all seriousness, I think the investigation has been compromised. Surprise everyone. I know where we met, and I'm positive it was just on purpose.
Chandler: (to Joey) Tribbs, ya know, the most definitely belonged to Trib. It was his coat! (Their dog starts barking from somewhere behind the couch.)
Joey: Do you know where worst is? People make the decision all the time but consider yourself lucky if that person is themselves. Look, you guys decided to take an old man's coat.
[Scene: The receptionist cubicle, Phoebe is waiting for the phone.]
Chandler: (on phone) Hi, not on the 30th, sorry. (Susan sits up, then falls onto the computer looking at her.)
Susan: Oh, oh! Hi. I've been thinking for over a minute about out. Any chance you and I could have some time to talk?
Phoebe: Yeah, well, that would solve everything.

Script 10
Perplexity Value: 2.26
Cherry Picked Set

Phoebe:	Okay, so, uh, so how's your body?
Joey:	Oh, it's great! I'm feeling great!
Phoebe:	Okay, so how about your teeth?
Joey:	Oh, you know, they're like a little harder.
Phoebe:	Yeah, well, you can see my teeth now. (laughs)
Joey:	So, how's your eye?
Phoebe:	Oh, it's great.
Joey:	So, how's your nose?
Phoebe:	Oh, it's great. (laughs)
Joey:	So, how's your lip?
Phoebe:	Oh, it's so soft.

A.3 Results per Script

Table 1: Results per individual script as rated by all participants

Script nr.	Script set	Creativity	Realism	Coherence
1	Cherry picked set	3.98	5.93	6.31
2	Similar PPL set	5.06	1.92	1.57
3	Low PPL set	2.27	5.11	4.12
4	High PPL set	4.10	2.21	1.35
5	Similar PPL set	5.17	3.86	3.23
6	Actual script	7.35	8.04	8.10
7	Low PPL set	4.91	5.92	5.14
8	Actual script	7.11	7.94	7.88
9	High PPL set	4.90	2.83	2.22
10	Cherry picked set	5.19	5.42	6.64

Table 2: Results per individual script as rated by participants who have watched Friends before

Script nr.	Script set	Creativity	Realism	Coherence
1	Cherry picked set	4.17	5.61	6.17
2	Similar PPL set	4.84	1.85	1.60
3	Low PPL set	2.30	4.77	3.81
4	High PPL set	3.83	1.89	1.17
5	Similar PPL set	4.95	3.45	2.73
6	Actual script	7.66	8.35	8.35
7	Low PPL set	5.07	6.01	5.13
8	Actual script	7.45	8.27	8.11
9	High PPL set	4.78	2.88	2.21
10	Cherry picked set	5.43	5.47	6.67

Table 3: Results per individual script as rated by participants who have not watched Friends before

Script nr.	Script set	Creativity	Realism	Coherence
1	Cherry picked set	3.51	6.69	6.49
2	Similar PPL set	4.93	2.03	1.35
3	Low PPL set	2.09	6.08	4.53
4	High PPL set	4.53	3.72	2.23
5	Similar PPL set	5.54	4.46	4.26
6	Actual script	6.35	7.23	7.43
7	Low PPL set	4.46	5.61	5.47
8	Actual script	6.08	7.09	6.96
9	High PPL set	4.66	2.70	1.82
10	Cherry picked set	4.12	4.53	6.35

Table 4: Results per individual script as rated by participants who are somewhat familiar with Friends

Script nr.	Script set	Creativity	Realism	Coherence
1	Cherry picked set	3.8	6.25	6.55
2	Similar PPL set	5.75	2.05	1.65
3	Low PPL set	2.3	5.35	4.65
4	High PPL set	4.55	2	1.2
5	Similar PPL set	5.5	4.55	3.85
6	Actual script	7.2	7.8	7.9
7	Low PPL set	4.8	5.9	4.95
8	Actual script	6.95	7.65	7.9
9	High PPL set	5.4	2.8	2.55
10	Cherry picked set	5.3	5.95	6.75