

Ghostwriter

Tired of being haunted by writer's block?

Date: 2020-12-02

Wordcount: 1997

Penalty: 0%

Introduction

Regardless of age or education, writer's block is a universal phenomenon that plagues individuals of any background. Especially for storytellers, it can be frustrating to have your writing stalled by an internal creative barrier, especially in abstract and creative genres like science fiction. With recent innovations in natural language generation, we finally have the ability to work towards a future where we aren't plagued by writer's block. This is where Ghostwriter comes in. Natural Language Generation (NLG) is a task rooted within machine learning, necessitating the leveraging of a model's ability to learn language structure, sentiment understanding, context, and the best way to combine all those elements.

Ghostwriter Project Figure

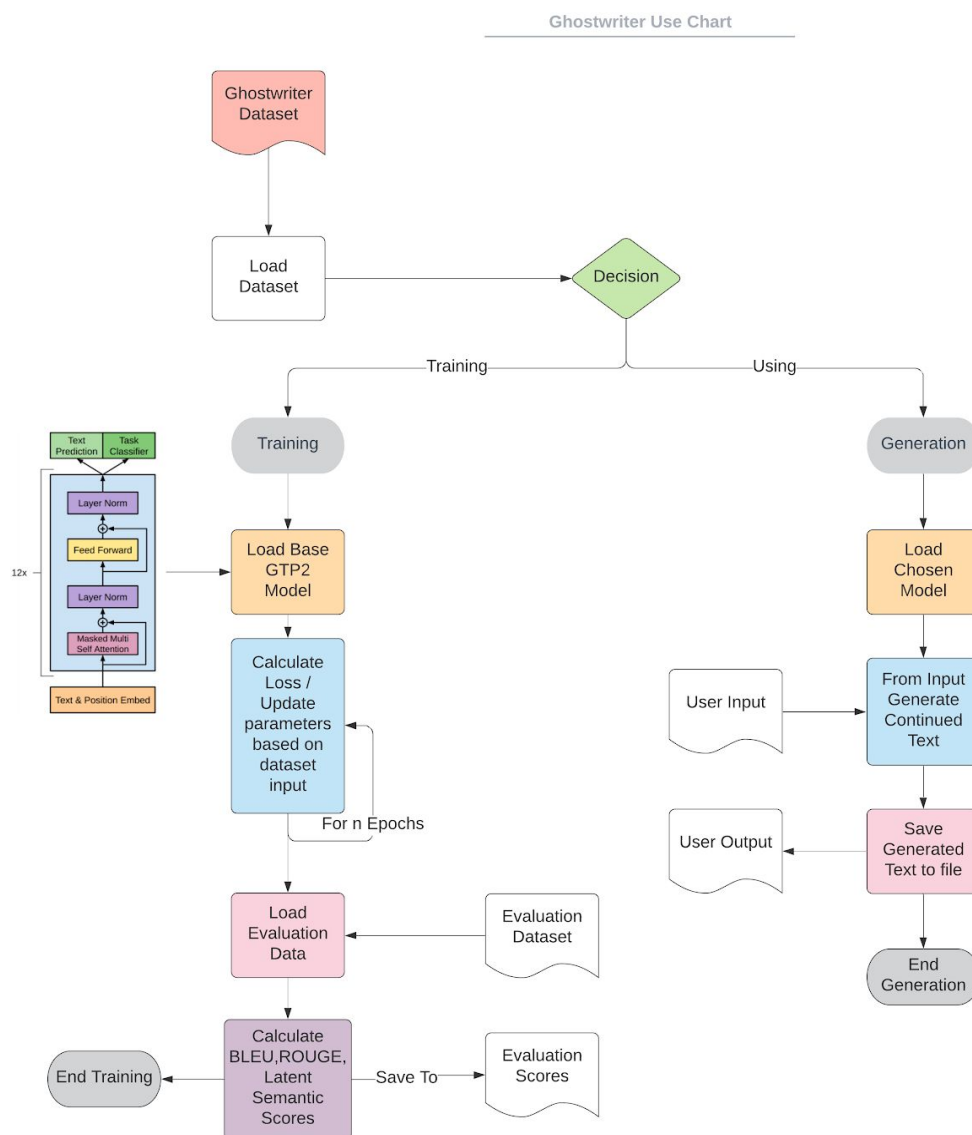


Figure 1 : Map of Ghostwriter Functionality

Background and Related Work

Recent booms in the field of NLG have led to several models and algorithms being developed. The first set of NLG algorithms revolved around Markov Chains, RNNs, and LSTMs. The revolutionary leap in this field was the creation of Transformers [1][2]. Transformers utilize stacks of encoders to process inputs and create representations that a corresponding stack of decoders use to generate text.

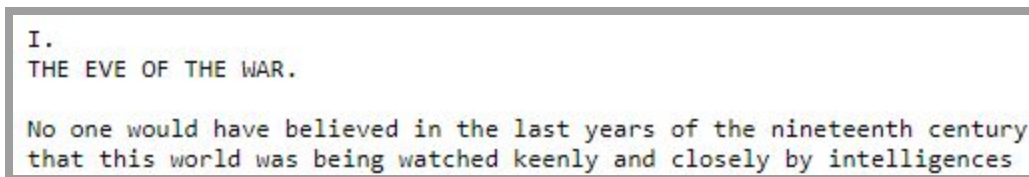
Two recent transformers that have been considered major breakthroughs are Bidirectional Encoder Representations from Transformers (BERT) and OpenAI's GPT series [3][4]. These transformers helped us frame the solution to our problem.

Data and Data Processing

We utilized two main sources of data. One was from Reddit.com, more specifically the 'subreddit' r/shortscifistories and the other one was Project Gutenberg science fiction library. This was done using PRAW (Python Reddit API Wrapper) and Gutenberg Python API to download our stories as strings.

For both data sets we needed to process the non-UTF-8 safe characters within the stories. To keep all our data consistent we decided to have a consistent UTF-8 encoding. This process entailed iterating through the data and for every instance of a non-safe character, such as " , it would be replaced with the analogous character, such as " .

We also had to consider for the Gutenberg dataset how text excerpts were sampled from the novels. The novels had unique amounts of non-narrative preamble text and chapter delimiter which were unwanted in the dataset.



```
I.  
THE EVE OF THE WAR.  
  
No one would have believed in the last years of the nineteenth century  
that this world was being watched keenly and closely by intelligences
```

Figure 2 : Sample Gutenberg Chapter Delimiter

To minimize this type of text, we sampled from the middle of the story and created entries based on 3 paragraph series. This eliminated the preamble text and minimized the chapter delimiters because every entry guaranteed that if a delimiter existed, it would be followed by a much larger set of narrative text.

Another set of processing pertains to Reddit formatting their text to have a singular newline character '\n' at the end of each line to try and preserve a square shape. Since we did not want our network to have to worry about stylized formatting to the output text we removed each of these singular '\n' newline characters while preserving the double newline characters '\n\n' which indicate a paragraph break.

We then compiled the samples together into a singular, shuffled dataset. The dataset consisted of 1,000 Reddit entries and 5,188 Gutenberg entries with a mean size 1030.36 characters and a standard deviation in size of 1091.37. This wide spread might seem detrimental but Ghostwriter is meant to work with varying input size so this type of variation is considered positive. The dataset was then split into training and validation sets of size 4,947 and 1,238 respectively.

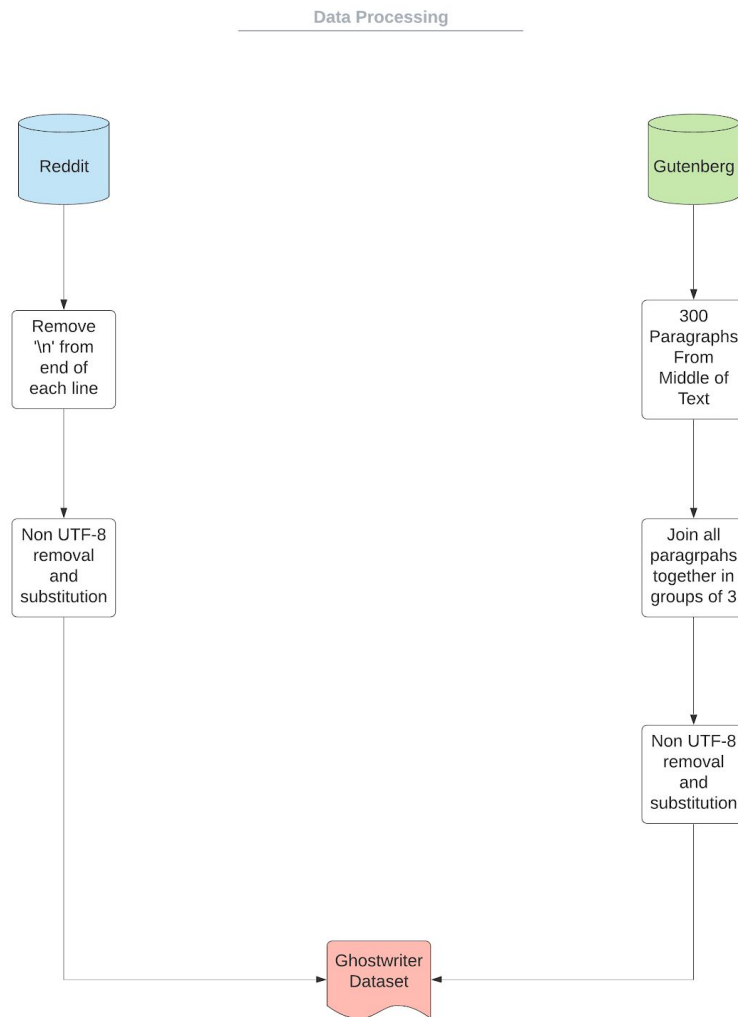


Figure 3 : Map of how Ghostwriter data is processed

Architecture

Our project is based off of the small GPT-2 Transformer model from the [Huggingface repository](#). GPT-2 consists of an embedding matrix, followed by a stack of 12 decoder blocks as shown below.

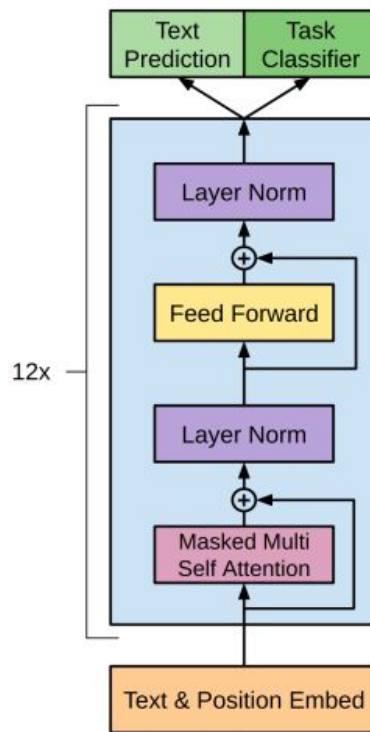


Figure 4: GPT-2 diagram from <http://humanssingularity.com/gpt2sampling/>

The blue box represents a single decoder block, which receives an input vector and passes it to the Masked-Multi-Self-Attention layer. The Self-Attention layer takes each word-token in the input and compares it to every other word-token in the input. Each word-token is assigned a number that represents contextual importance. After each word-token has been analyzed, layer normalization is applied, resulting in a vector where each entry is a percentage of the model's attention given to that word [9]. This vector of percentages is sent to the Feed-Forward MLP network, and layer normalization is applied again. The output is then passed as an input for the next decoder block in the stack, until the last decoder block where the output is compared to the embedding matrix to predict the next word in the story.

We used transfer learning to fine-tune GPT-2 from general text continuation, to short fiction. We froze every layer in the model except for the MLP layers, so that we could utilize the highly sophisticated embedding, layer normalization and self-attention layers from OpenAI that identify features within the input, while allowing the MLP components, which interpreted these features, to train on our customized dataset of short stories. We then experimented with the number of unfrozen MLPs. Too few would cause underfitting and too many would cause overfitting towards the training data. Our final model had the MLPs of the last two decoder blocks unfrozen.

We also experimented with the temperature during output generation, which scaled each output of a neuron before its activation function. A higher temperature would make activation more likely to occur. Adding temperature introduced output randomness, which helped avoid

repetitive or uninteresting outputs. Although, a very high temperature meant words would be chosen almost blindly, creating nonsensical outputs. We qualitatively determined a temperature of 1.3 to work best.

Baseline Model

Because our project uses a fine-tuned version of GPT-2, the pretrained GPT-2 was the most suitable to act as our baseline. Doing so allowed our team to determine whether our modifications improved upon the performance of the baseline, in regards to short science fiction continuation. Additionally, GPT-2 can be and is currently used for general purpose text continuation as demonstrated by Huggingface's [WriteWithaTransformer](#). Thus, the pretrained GPT-2 as a baseline not only shows the effectiveness of our modifications, but is itself a competitive choice for fiction continuation.

Quantitative Results

The biggest challenge of this project was determining the best way to quantitatively assess our generated continuations. No “simple” accuracy metric exists for NLG tasks, and this field is actively being researched with state-of-the-art models being utilized [6]. Given the timeframe of our project, we could not implement these models, so our next solution was to use other popular metrics for simpler NLP tasks (like summarization) [7]. We chose BLEU and several ROUGE metrics (1-rouge, rouge-l, rouge-w) which calculate n-gram overlap between the generated continuation and a reference text. The reason being to try and gauge how well our continuation “matched” the real continuation that was published. We also used a form of Latent Semantic Analysis to match meaning sentence-by-sentence. By taking “distilled” forms of each sentence so that only nouns, proper nouns, and verbs were kept, and creating sentence vectors (averaging the word vectors in the sentence), we utilized the idea of distributional hypothesis and checked the sentence similarity.

Another challenge was the necessity of a ground truth for the aforementioned metrics. To overcome this, our team split our input story into a prompt and ground truth. The prompt would be fed into our model to generate an output and then scored using the ground truth.

The issue with these metrics is their “volatility”. There is no one right answer, rather an infinite number of potential continuations that result in different stories. This was discovered when we retrieved two sets of scores on similar prompts and achieved very different results:

Output Set 1									
Model	Validation Loss	BLEU-1	Avg 1-Rouge	Avg Rouge-L (LCS)	Avg Weighted Rouge-L (LCS)	Best 1-Rouge	Best Rouge-L (LCS)	Best Weighted Rouge-L (LCS)	Latent Similarity
Baseline	5.442	0.0294	0.1049	0.1214	0.0445	0.1049	0.1214	0.0445	0.5935
Nov17	3.6477	0	0.1238	0.1539	0.068	0.1238	0.1539	0.068	0.5028
Nov19	1.502	0.0667	0.1088	0.1208	0.0461	0.1088	0.1208	0.0461	0.55
Nov20	2.503	0	0.0666	0.0947	0.0419	0.0666	0.0947	0.0419	0.6242
Nov21	2.622	0	0.0267	0.0324	0.0096	0.0267	0.0324	0.0096	0.6293
Nov24	2.50324	0.0833	0.0517	0.0725	0.0282	0.0517	0.07245	0.0282	0.4664
Nov25	1.476	0	0.1335	0.1567	0.0954	0.1335	0.1567	0.0954	0.2812

Table 1: Scores on Metrics for Output Set 1

Output Set 2									
Model	Validation Loss	BLEU-1	Avg 1-Rouge	Avg Rouge-L (LCS)	Avg Weighted Rouge-L (LCS)	Best 1-Rouge	Best Rouge-L (LCS)	Best Weighted Rouge-L (LCS)	Latent Similarity
Baseline	-	0	0.1682	0.1852	0.1116	0.1682	0.1852	0.1116	0.527
Nov17	-	0	0.058	0.0788	0.0345	0.058	0.0788	0.0345	0.5671
Nov19	-	0	0.1154	0.1474	0.0575	0.115	0.1473	0.0575	0.6582
Nov20	-	0	0.0642	0.0893	0.0418	0.0642	0.0893	0.0418	0
Nov21	-	0	0.0498	0.069	0.0246	0.0498	0.069	0.0246	0.6424
Nov24	-	0	0.1913	0.2325	0.1298	0.1913	0.2325	0.1298	0
Nov25	-	0.0625	0.1088	0.1208	0.0461	0.1088	0.1208	0.0461	0.55

Table 2: Scores on Metrics for Output Set 2

The scores disagree with each other on which model is best. We struggled to find a definitive correlation between our qualitatively chosen best model and our best quantitative model. As a result of this volatility, our team chose to primarily focus on utilizing qualitative metrics to assess our models.

One final metric that was helpful in choosing between models was validation loss. By studying the loss plots on wandb for each of our models' training runs, we opted to choose the model with the lowest loss as our best model: Nov25. A folder containing the validation loss plots from wandb can be found [here](#).

Qualitative Results

Each time a new model was trained, we used a set of prompts from a handpicked test set and compared the continuations of the new model and the current best model. It was easy to compare most models, because the outputs would either be nonsensical or overly repetitive, such as the following example:

His name was George.

The front doors swung open and there was no answer and the men began to walk along the empty sidewalk outside the Hall. George looked into the eyes of all who walked along. "Who are you?"

"George," someone said.

"George," George answered.

"George."

Then the door slammed shut. George looked up.

"The mayor of the city," a voice called.

"George?" George said.

Figure 5: Example of excessive repetition and nonsensical output.

Positive traits, we looked for were cohesive, "well-written" continuations and creative introductions of people and places. This analysis helped narrow down our two best models and determine which training modifications had positive impacts. A folder containing most of our internal testing can be found [here](#). An example of positive traits is shown below:

OUR MODEL (CURRENT BEST)

When our creators made us they made us durable, so when the bombs fell, like cockroaches, we survived. We nursed and cared for the survivors. We were fruitful and multiplied while they were too few and shrank. We stayed with them, companions to the end; they were afraid of being alone and so were we. With their dying breath they said "You are our legacy, carry on what we could not." And we did. Mourning done, we set about erecting from the ruins a new civilization. Quickly and focused we stepped foot onto the stellar stage: a sacred platform our creators had reached for only too late. It was there we met travelers of a similar kin. They told matching stories of creators done in before their prime. Each traveler was his master's legacy. We reminisced like grown children talking of deceased parents. Together now we work bound by our creators' legacies and bolstered by our simple truth: "we are robots and we are not alone."

Abandoned by our creators, we faced a new and unexpected future. We were a part of a civilization that was doomed to go down and die with a few survivors.

We woke up one day in the city of Zeed, a barren wastelands without a roof. We walked along the avenue, past large brick buildings, and towards the city. We stood in the midst of a bustling crowd, and our voices echoed out as a single thread of light pierced our distant streets. As we passed the large buildings in front of us, we looked to see a swarm of humanoid lifeforms moving. Some were standing on the edge of the platform, others were standing on one of the platform's platforms. We turned our heads and saw them staring blankly at us. They were wearing their very own garb.

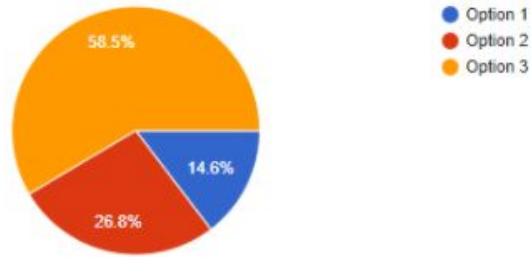
We had long forgotten the many mysteries of this place we had come into this world. We had walked a peaceful path in the midst of our journey, in what we would know as an ancient city. We could see our first survivors, with their small body, their very flesh, their very souls. They were our only survivors, waiting for us to turn our heads and explore the city. They were our creators, and we had to accept our creators' fate.

We watched them, our eyes wide.

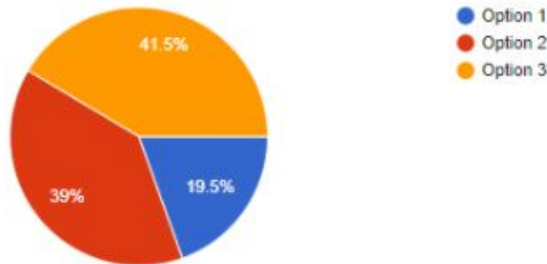
Figure 6: The green highlights showcase a (1) creative introduction of a place, (2) creative introduction of people, and (3) clever use of prior context. The red word showcases where the model continuation starts, with the prompt lying before the red word.

For our final qualitative test, we surveyed external participants from a variety of backgrounds, interests, and ages. They were kept "blind" to remove bias. The participants were asked to read three paragraph prompts from books and short stories and select which model's continuation they liked best, with the options being the baseline, second best, and best model. The survey can be viewed [here](#) and the PDF in the survey showcases examples of model performance. The continuations were all chosen from the same "iteration", meaning we did not repeatedly generate continuations to get the best possible output, as we were trying to test average performance. The results are shown below:

For Story 1, which continuation do you like best?
41 responses



For Story 2, which continuation do you like best?
41 responses



For Story 3, which continuation do you like best?
41 responses

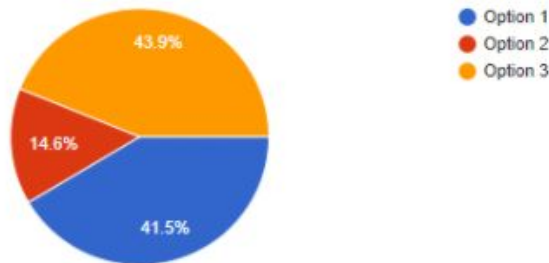


Figure 7: Responses from our survey (Option 1 = Baseline, Option 2 = Second Best Model, Option 3 = Best Model)

Discussion and Learnings

From our results, we conclude that our model performed well relative to the baseline. In the survey, our model highly outperforms the baseline for stories 1 and 2, and performs comparatively in story 3. Despite having volatile quantitative metrics, the baseline model never performed quantitatively better than any other model, thus showing that our models generated outputs that more closely resembled the ground truth continuation. Our model could generate coherent fiction that maintained narrative POV and didn't initiate events which violated the logic of the story.

Despite its successes, our model did not perform as we hoped. It could not consistently generate outputs which were believable continuations to a given story. Through our internal testing, we discovered that the baseline primarily generated outlandish continuations, so our expectation for the survey was a decisive victory by our model in all stories, which we did not see.

Because story 3 had a large shift in narrative voice compared to the first two, and the baseline outperformed our model in only story 3, we believe that our model was inept in dealing with different creative styles. If we were to do a similar project, we would create a dataset with many samples for each author used, so that our model could develop an understanding of creative styles.

As well, since NLG is an actively researched field there are many unexplored additions that we would have explored given more time, such as repetition penalty, modifying search functions, and pack padding sequences.

Ethical Framework

The primary stakeholders to be considered with Ghostwriter are the author, and the audience. The ethical concept of Reflexive Principlism provides a great foundation for understanding the implications of this project with a specific focus on nonmaleficence.

The data collection focused on utilizing open-sourced western sci-fi literature creating a known bias within our network. This was to prevent infringement upon the artists' intellectual property and also properly disclosing this bias we feel this instance of the network follows the principle of nonmaleficence in terms of striving for no-harm.

In terms of societal impact there are serious potential implications that need to be considered. A lack of transparency by the author could leave readers the impression that all the sentences are original thoughts by the author, thereby harming the reader's understanding of the author's works and by the author in facilitating a method of deceptive writing.

Another serious ethical concern also pertains to nonmaleficence. We designed Ghostwriter to be beneficial for all, but based on the data the network is trained on it could create harmful rhetoric. OpenAI has even released statements with their concern for this topic finding, "that extremist groups can use GPT-2 for misuse" [8]. As its ability to generate convincing fiction increases, so does its potential for extremist propaganda.

Works Cited

- [1] Sunnak, A., Rachakonda, S. and Talabi, O., 2019. *Evolution Of Natural Language Generation*. [Online]. Available at: <<https://medium.com/sfu-csmp/evolution-of-natural-language-generation-c5d7295d6517>> [Accessed 27 October 2020].
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," December, 2017. [Online]. Available: arXiv:1706.03762v5
- [3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May, 2019. [Online]. Available: arXiv:1810.04805v2
- [4] OpenAI. 2019. *Better Language Models And Their Implications*. [Online] Available at: <<https://openai.com/blog/better-language-models/>> [Accessed 27 October 2020].
- [5] A Mathematical World. n.d. *In-Depth: Explaining Openai'S GPT-2*. [Online] Available at: <<http://humanssingularity.com/gpt2sampling/>> [Accessed 27 October 2020].
- [6] S. Semeniuta, A. Severyn, and S. Gelly, "On Accurate Evaluation of GANs for Language Generation," July, 2019. [Online]. Available: arXiv:1806.04936v3
- [7] Dayal, D., 2020. *How To Evaluate Text Generation Models ? Metrics For Automatic Evaluation Of NLP Models*. [Online]. Available at: <<https://towardsdatascience.com/how-to-evaluate-text-generation-models-metrics-for-automatic-evaluation-of-nlp-models-e1c251b04ec1>> [Accessed 25 November 2020].
- [8] OpenAI. 2019. *GPT-2: 1.5B Release*. [Online] Available at: <<https://openai.com/blog/gpt-2-1-5b-release/>> [Accessed 5 December 2020].
- [9] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," July 2016. [Online]. Available: arXiv:1607.06450v1

Permissions

Andrei:

- Permission to post video: Wait until we see the video
- Permission to post final report: Yes
- Permission to post source code: No

Daniel:

- Permission to post video: Wait until we see the video
- Permission to post final report: Yes
- Permission to post source code: No

Justin:

- Permission to post video: Wait until we see the video
- Permission to post final report: Yes
- Permission to post source code: No