INTRODUCTION :

Ensuring coherence in textual discourse is crucial for readability and comprehension. This project addresses this challenge by employing neural models to measure coherence at both local and global levels. Our system utilizes a CNN + LSTM model to generate image captions, followed by passing them through a GPT-2 model for story generation. We then evaluate the coherence of the resulting paragraphs to gauge their accuracy and consistency.

ABSTRACT :

We've developed a system that generates image captions using a CNN + LSTM model. These captions are then passed to a GPT-2 model for story generation. Subsequently, we assess the coherence of the resulting paragraph to evaluate its accuracy and consistency.

DATASET :

- **Flickr 8k Dataset:**
  - Collection of sentence-based image description and search, consisting of 8,000 images that are each paired with five different captions provides clear descriptions of the salient entities and events. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.
- **ROCStories and Story Cloze Test Corpora:**
  - The Story Cloze Test emerges as a fresh framework for assessing narrative comprehension, generation, and script acquisition, offering a departure from the prevalent method, the Narrative Cloze Test (Chambers & Jurafsky, 2008). Its implementation involves selecting the appropriate ending to a four-sentence story. To facilitate this test, a new corpus of five-sentence commonsense stories, ROCStories, was developed. This corpus stands out for its dual characteristics: it encapsulates a diverse array of causal and temporal common sense relations between everyday occurrences and serves as a high-quality repository of mundane life narratives, suitable for both testing and generating stories.
  - Link : https://cs.rochester.edu/nlp/rocstories/

BASELINE MODELS:

- Expressing an Image Stream with a Sequence of Natural Sentences
    - CRCN Model: It is proposed that a novel multimodal architecture named coherent recurrent convolutional networks (CRCN) integrate convolutional neural networks for image description, bidirectional recurrent neural networks for the language model, and the local coherence model for a smooth flow of multiple sentences.
    - Link: https://github.com/cesc-park/CRCN
    - **Issues Encountered:** The datasets are available while installing and running the dependencies and prerequisites respectively, the browncorpus package also known as browncoherence used by many researchers to extract entity features is now unavailable on the internet. This dependency has no replacement which was compatible with the existing CRCN model.
      Expired link:
      https://bitbucket.org/melsner/browncoherence/get/d46d5cd3fc57.zip
    - The alternatives to the browncoherence which were explored are:
        - Coh-Metrix
        - Cohere toolkit

- Transformer Models for Text Coherence Assessment
    - It includes Grammarly Corpus of Discourse Coherence (GCDC), Wall Street Journal (WSJ), and Recognizing Textual Entailment (RTE) datasets. The datasets were available.
    - Link:
      https://github.com/tushar117/Transformer-Models-for-Text-Coherence-Assessment
    - **Issues encountered:** The unavailability of dependencies and requirements of the implementation.
        - Update the dependency of pytorch_lightning from pytorch_lightning. metrics import Metric the metric module was removed from it.
        - The unavailability of the outdated version of PyTorch 1.6.0 which was used during the implementation (No matching distribution found error)

- Sequential Vision to Language as Story: A Storytelling Dataset and Benchmarking
    - Herein, The paper presents a storytelling dataset and benchmarking process involving the collection of five images every 15 seconds from videos, followed by human annotation through Amazon Mechanical Turk. CAMT, GLACNet, ViT, and SAES are discussed and compared, utilizing encoder-decoder strategies and object detection techniques to generate coherent stories. Automatic

evaluation metrics like BLEU, CIDEr, ROUGE, and METEOR are employed to assess the quality of machine-generated stories, with human evaluations confirming the challenging nature of the proposed dataset compared to existing ones.
    - Link: https://research-repository.uwa.edu.au/en/publications/sequential-vision-to-language-as-story-a-storytelling-dataset-and
    - **Issues encountered:**
        - The unavailability and access denial of the dataset SSID. Link: https://ieee-dataport.org/documents/sequential-storytelling-image-dataset-ssid
        - The large size and access unavailability of the VIST dataset. Link: https://visionandlanguage.net/VIST/
    - We did a detailed analysis of the datasets while we waited to get access to these datasets via IEEE.
    - We attempted the implementation of GLACNet and encountered the above issue.
- ReCoRL Model:
    - We implemented the ReCo-RL model which we used a pre-trained ReCo-RL model and preprocessed data which included image features, VIST captions, and entities preprocessed by spacy. The pre-trained and vocab files were used for this purpose. In the predicted outputs by our ReCO-RL model, each line contains the sequence ID and the generated story.
    - Due to our inaccessibility to the VIST dataset, this pre-trained model helped us evaluate and understand the coherence outputs related to sequential storytelling.
    - They've proposed a reinforcement learning framework, ReCo-RL, with reward functions designed to capture the essence of these quality criteria namely, relevance, coherence, and expressiveness. Experiments on the Visual Storytelling Dataset (VIST) with both automatic and human evaluations demonstrate that the ReCo-RL model achieves better performance than state-of-the-art baselines on both traditional metrics and the proposed new criteria.
    - **Issues encountered: The fine-tuning required access to the dataset in order to mold it according to our requirements**
        - The large size and access unavailability of the VIST dataset. Link: https://visionandlanguage.net/VIST/

IMPLEMENTATION

A. Image Caption Generator using Deep Learning on Flickr8K dataset

Generating captions for images presents a significant challenge within the realm of deep learning. We are using various computer vision and NLP techniques to discern image context and describe it in natural languages, such as English. A functional model of the image caption generator is constructed employing CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory) units.

For training purposes, the Flickr8K dataset is employed, comprising 8000 distinct images. Each image is associated with five unique sentences that elucidate its content.

Steps:

- Features are extracted from the image using techniques of convolutional neural networks (CNNs), and LSTM which capture relevant visual information.
- Caption data is loaded, containing descriptions or labels associated with images.
- The vocabulary is generated by collecting unique words from the captions data.
- A mapping between image prefixes (identifiers) and captions is created to link each image to its corresponding descriptions.
- Text data is cleaned by removing any irrelevant characters, symbols, or formatting issues.
- The dataset is divided into training and testing sets to evaluate the model's performance.
- A model is created, often using recurrent neural networks (RNNs) or transformer architectures, to learn the relationship between images and captions.
- The model is trained on the training dataset, where it learns to generate captions based on the input images.
- Once trained, the model is saved or loaded for future use.
- The model's performance is evaluated by testing it against the unseen data in the test dataset.
- Scores such as BLEU (Bilingual Evaluation Understudy) or METEOR (Metric for Evaluation of Translation with Explicit Ordering) are calculated to measure the quality of generated captions compared to the ground truth.
- Finally, the model is tested with real-world images to observe its performance in generating captions for unseen images.

- We attempted to create paragraphs with the vocabulary of the training dataset. And use Beam search for this task.
- Beam search result of generating story:

```
Normal Max search: stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stre
tched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched stretch
ed stretched stretched stretched stretched stretched stretched stretched stretched stretched stretched
```

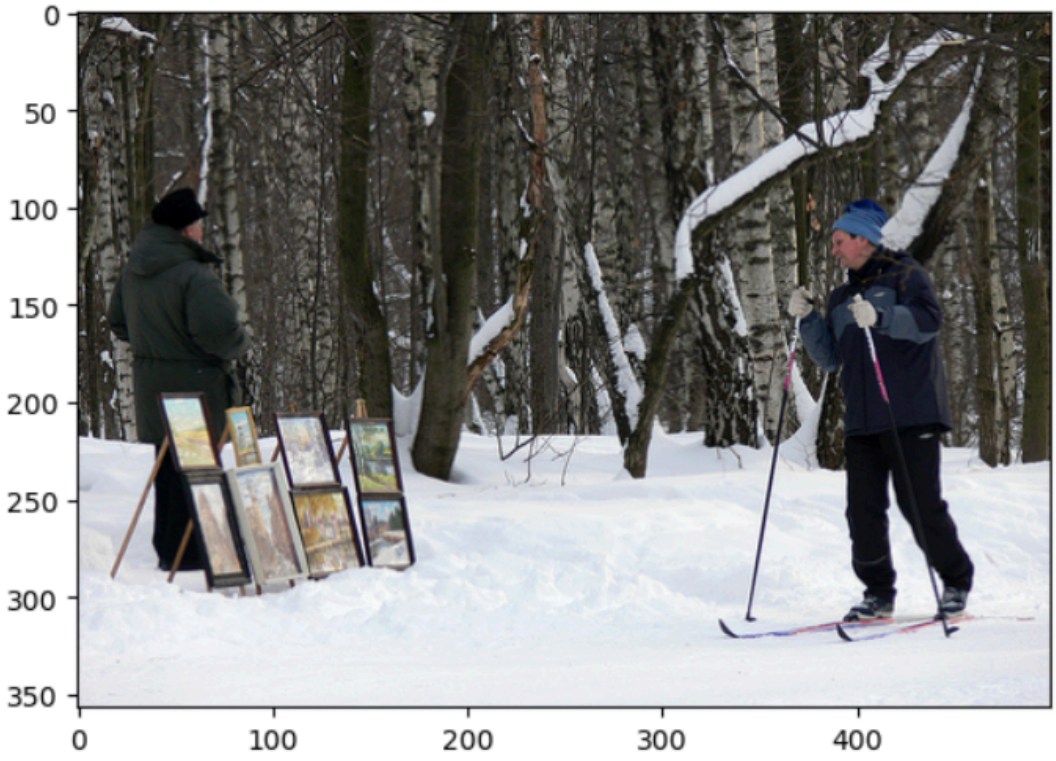
- Upon receiving repetitive output, we switch to using Top-k sampling for the task.

- We obtain following observations:

  Normal Max search: marketplace search formally raft weimaraners goalkeeper inflating half-pipe mills mom Rotweiler interviewed Half windbreaker the frown pausing military-style forests Riders aisle suits graffti cheerleader driver flaggers hands dumps cluster Cars mini-trampoline childing temple beneath eight vertical khakis Beige stare

- We then predict the sentences, predicting the next word with the highest probability. and comparing them with the actual.

```
----------------------Actual----------------------
startseq a man in a hat is displaying pictures next to a skier in a blue hat endseq
startseq a man skis past another man displaying paintings in the snow endseq
startseq a person wearing skis looking at framed pictures set up in the snow endseq
startseq a skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
---------------------Predicted---------------------
startseq a man standing on the snow endseq
```



```
---------------------Actual---------------------
startseq a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl endseq
startseq a little girl is sitting in front of a large painted rainbow endseq
startseq a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it endseq
startseq there is a girl with pigtails sitting in front of a rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted--------------------
startseq a girl in a red jacket stands in front of a red fence endseq
```

- Calculation of Bleu score against test dataset:

```
  0%|          | 0/810 [00:00<?, ?it/s]
BLEU-1: 0.658807
BLEU-2: 0.485955
```

- Using the Vgg16 pre-trained model to use features in order to predict the words that require the features of the image.

Model: "functional_3"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1,792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36,928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73,856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147,584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295,168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590,080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590,080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1,180,160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2,359,808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2,359,808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2,359,808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2,359,808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2,359,808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102,764,544
fc2 (Dense)	(None, 4096)	16,781,312

Total params: 134,260,544 (512.16 MB)

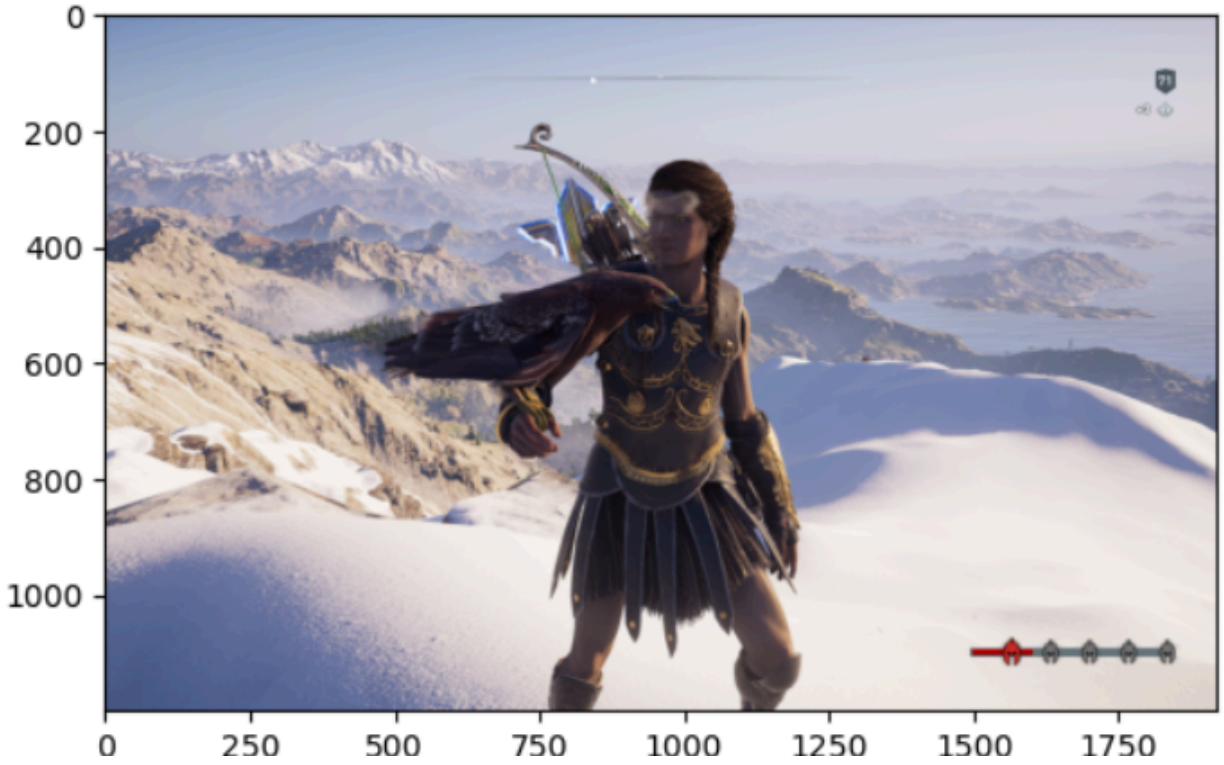
Trainable params: 134,260,544 (512.16 MB)

Non-trainable params: 0 (0.00 B)

- Then we test the model against real-time images to predict the captions.

```
'startseq a group of people are sitting in a row endseq'
```

'startseq a man is skiing down a snowy hill endseq'



- Then we generate the story with GPT-2(non-fine-tuned), giving the predicted sentence as the input to the model for all the test images.

"I don't know what to do," she says, "but I'm going to get out of here."
I ask her if she has any idea what's going on. She says she doesn't, and that she's not sure what she should do. I tell her that I have no idea how to deal with this
 a smiling girl in a blue jacket and glasses is running in a wooded area in the middle of the woods.
I'm not sure if this is the first time I've seen this kind of thing, but I'm pretty sure it's one of my favorite things to do in my life.
 a man wearing a black jacket and glasses looks down on a group of young men in the middle of the street.
"I don't know what you're talking about," he says. "I'm not going to tell you what to do. I just want you to know that I'm here to help you."
I ask him why he's here, and he replies, "Because I love you. You're my best friend and I want to be with you forever."
 a man with a brown and white and white jacket is holding a stick in his left hand. He's holding the stick in his right hand, and he's saying, "I'm going to take you to the hospital. I don't know what you're doing, but I want to get you out of here." I'm like, I can't do that. It's like I've got to do something. And then he goes back to his desk and says to me,
 a dog jumps over a hurdle and gets stuck in the middle of the road.
I'm not sure if this is a good thing or a bad thing, but I think it's important to note that this isn't the first time this has happened. In fact, I've seen it happen a few times before, and I'm pretty sure it was the last time I saw it. It's not like this dog is going to be able to get out of a ca

BLEU-1: 0.279076
BLEU-2: 0.178170

We can infer the following:

B. **BLEU-1 (0.27)**: This score indicates that approximately 27% of the individual words (unigrams) in the generated text match those in the reference text(s). While not a very high score, it suggests a moderate overlap in terms of individual word choices.
C. **BLEU-2 (0.17)**: This score is lower than the BLEU-1 score, which means that the generated text has a poorer performance in capturing two-word sequences (bigrams) from the reference text(s). Only about 17% of the bigrams in the generated text match those in the reference text(s).

D. Generating short paragraphs using GPT-2 model

GPT, short for Generative Pre-Trained Transformer, originates from OpenAI, a leading AI research organization. It is a decoder-only part of the transformer model. It derives its name from the transformative "Transformer Architecture," a groundbreaking development in language modeling. GPT represents an evolution of this architecture, offering a more sophisticated approach to natural language processing (NLP) compared to earlier methods such as RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory Networks), and CNN (Convolutional Neural Network).

There are 4 versions of GPT-2 :

- small: 124 million parameters
- medium: 355 million parameters
- large: 774 million parameters
- xl: 1.5 billion parameters

We have tried with the first 3 GPT-2 pre-trained versions. Below are the results for the few randomly generated captions:

- GPT-2 small

Generated Story 1: My name is Mansi and I am a LinkedIn Influencer!!!

I am the founder and CEO of LinkedIn.com, a social media platform for people to connect with each other. I have been an active member of the LinkedIn community for ov
er 10 years and have worked on a wide range of projects, including the development of social networking platforms like Facebook, Twitter, LinkedIn, Instagram, and Pin
terest. My work has been featured in publications such as The New York Times, The Wall Street Journal

Generated Story 2: A white girl is sitting on an office chair with pink top and specs watching terminator ichthyosis.

"I don't know what you're talking about," she says. "I'm just trying to figure out what's going on. I can't tell you how many times I've been in the hospital, and I'm
not sure what it's like to be in a hospital. It's not like I have to go to the doctor, but I do have a lot of

Generated Story 3: The boy went to the temple and said jai shree ram ṛṣṭṇa, "I am the son of the Lord."

The Lord said to him: "You are my son, and I am your father." And he went out and sat down on the ground with his head bowed and his eyes closed. Then he said, 'O Lor
d, what is the meaning of this?' The Lord replied: 'It is because you are the Son of

- GPT-2 medium

Generated Story 1: My name is Mansi and I am a LinkedIn Influencer who has been featured on Forbes, Business Insider, Mashable, and many other sites. I have been wo
rking with LinkedIn for over a year now and it has become one of my favorite social media platforms to work with.
I have worked with thousands of people on LinkedIn and have learned a lot along the way. Here are some of the things I've learned:
1. Don't be afraid to ask for help.

Generated Story 2: A white girl is sitting on an office chair with pink top and specs watching terminator iced tea. She is wearing a pink shirt and pink skirt. Her h
air is tied up in a ponytail and she is holding a cup of tea in her hand.

White girl: Hello, my name is White girl and I am here to ask you a question. What is the meaning of life? What does it mean to be a human being? How do we live in th
is world?

Generated Story 3: The boy went to the temple and said jai shree ram utha.

The next morning, the boy came to me and told me that his father had died. I told him that he should go to his mother and tell her what had happened. He said, "I will
go with you." So I went with him to my mother's house. When I reached there, I saw my father lying on the floor with his hands tied behind his back. His eyes were clo
sed

- GPT2 - LARGE

Generated Story 1: My name is Mansi and I am a LinkedIn Influencer who has been featured in the New York Times, The Wall Street Journal, USA Today, Business Inside
r, and many other publications. I'm also the author of the best-selling book, "The Power of Introverts in a World That Can't Stop Talking," which is available on Amaz
on.com and in bookstores everywhere. You can also find me on Twitter, Facebook, Google+, and LinkedIn.

Generated Story 2: A white girl is sitting on an office chair with pink top and specs watching terminator on tv. She is wearing a pink shirt and pink pants. Her hai
r is in a ponytail and she has pink lipstick on her lips. The girl looks at the camera and says, "I'm a girl. I'm not a boy." The camera then cuts to a black girl who
is standing in front of a mirror with her hair in pigtails and wearing black pants and a white shirt.

Generated Story 3: The boy went to the temple and said jai shree ram ṛṣṭaṇa, "I am the son of the lord of this temple, and I have come here to ask you for a boon."

The sage said to him: "O son, you have asked me a question, but I do not know the answer to it. Go and ask it yourself." So he went and asked the question. The sage a
nswered him, saying:

- From the above results, we can observe that there is a bias for the 3rd generated story as the data on which GPT-2 was primarily trained was Reddit data. Nearly half of the Reddit users belong to the US and leftovers are dominated by European

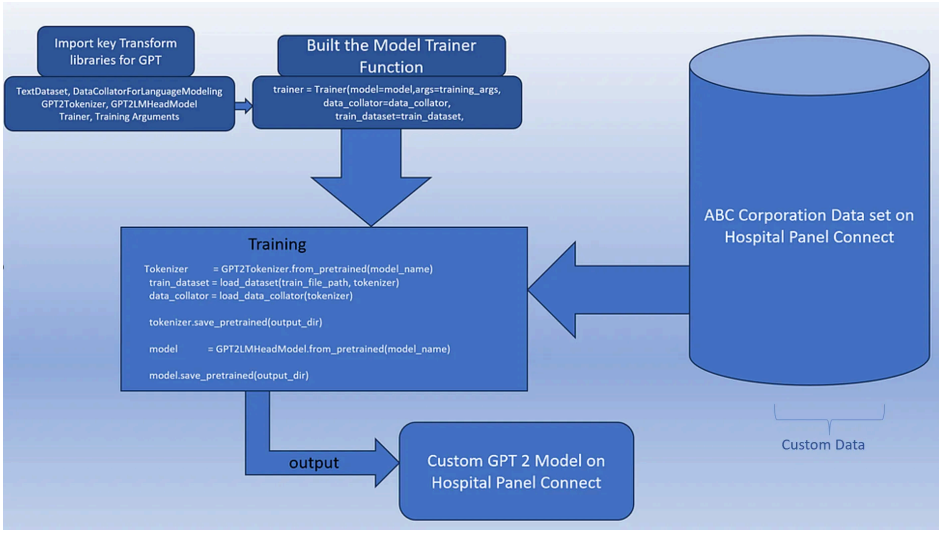
countries. In the 3rd generated story, we can observe a strong bias as the original sentence has an Indian context, but the model was not able to keep up with it.

- The above results also show that GPT-2 models, small, medium, and large made a little improvement respectively. However, the minor improvement came with a lot of computation as the number of parameters grew exponentially.
- Hence, we selected GPT-2 small to be fine-tuned with our data.

- The corpus on which we are fine-tuning the model is **ROCStories and Story Cloze Test Corpora.**
- We have picked the first 30,000 sets of 5 sentences.
- Below is the generated output of the fine-tuned GPT-2 model:

```
 'a brown dog sits on a blue lap  as his mother chews on his cereal. His mother begins to cry as he sits in his ro
cking chair with his cereal. He smiles to her and begins',
 'a man is standing on a rock with a rope. He wants a new rope. The man puts on the rope to get on the rocks. He g
ets a rope with rope on it.\n',
 'a young boy wearing a blue shirt and pirates  hats. The boy chased pirates from behind the tree, cutting away at
them. After a bit of chasing, the pirates got on the boat. The',
 'a man is walking past a cement wall with a red writing  stamp on his door. He notices the red writing stamp has
been missing. He searches for a way to find the handwriting. He finds',
 'a man sits on a rock  in the park. He has never been to a volcano. A young man has been visiting the volcano. He
gets in his car to get his phone. He calls',
 'a man and woman are sitting around a table in front of a wall. They are talking loudly and gossiping about their
boss. The woman starts to cry. She pulls her boss aside and tells',
 'a man in a black jacket is gazing through the snow. He pulls out his gun and points to his brother and tells him
to calm down. The officer orders him to remove his weapon. It',
 'a bird flies low to the ground  and takes a bite.\nBob wanted to visit a local mall. He went over to the mall at
night. He got a nice chair and some popcorn.',
 'a man in a blue shirt is skating on a skateboard. He grabs his skateboard and swings around, his face hurting. T
he skateboard goes in the other direction. The man skates',
 'a man is jumping a man back to his birds  back yard. A neighbor is making an excuse for a fire. A man buys a pol
e instead of a pole. The neighbor was upset that he',
 'a man and a woman pose for a picture  while they smoke a cigarette.\nThe man drank some soda. He saw a huge stai
n on his carpet. The stain caused his car to start to',
 "a snowboarder is jumping over a snowy hill! She's about to pass out when she breaks her foot. Her feet hurt a bi
t and she sprained it. Anna was afraid she'd",
 'a dog swimming with a tennis ball in its mouth. The fish caught the ball and tried to run away. I ran to the fis
hing pole and got on top of the fish. The fish sw',
 'a woman is wearing a white veil and a sweater is sitting on a couch  in a tree. She asks my brother if she would
like a baby to watch her. My brother refuses to tell her',
```

Below is the process flow of the GPT-2 fine-tuning:



Hyperparameters used during the inference:

- **Temperature:** Adjusting the temperature scales the probabilities of word generation. Higher temperatures encourage more original predictions, while lower temperatures keep the model focused and less likely to diverge from the topic.

- **Top-p filtering:** This method sorts word probabilities in descending order and sums them up until a specified threshold p is reached. It retains the most relevant word probabilities, acknowledging that multiple words may be suitable given a sequence.

- **Num_beams:** In GPT model fine-tuning, num_beams regulates the number of beams used in the decoding phase. Each beam represents a potential sequence of words during beam search decoding. Setting num_beams greater than 1 allows the model to maintain multiple active hypotheses, ensuring consideration of diverse sequences.

- **Top-k filtering:** It is a technique used during text generation to restrict the vocabulary size considered for each word prediction. Instead of considering the entire vocabulary, the model only samples from the top-k most likely words according to their probabilities. This approach helps control the diversity of generated text, ensuring that the model focuses on the most relevant words while still allowing for some variation in the output.

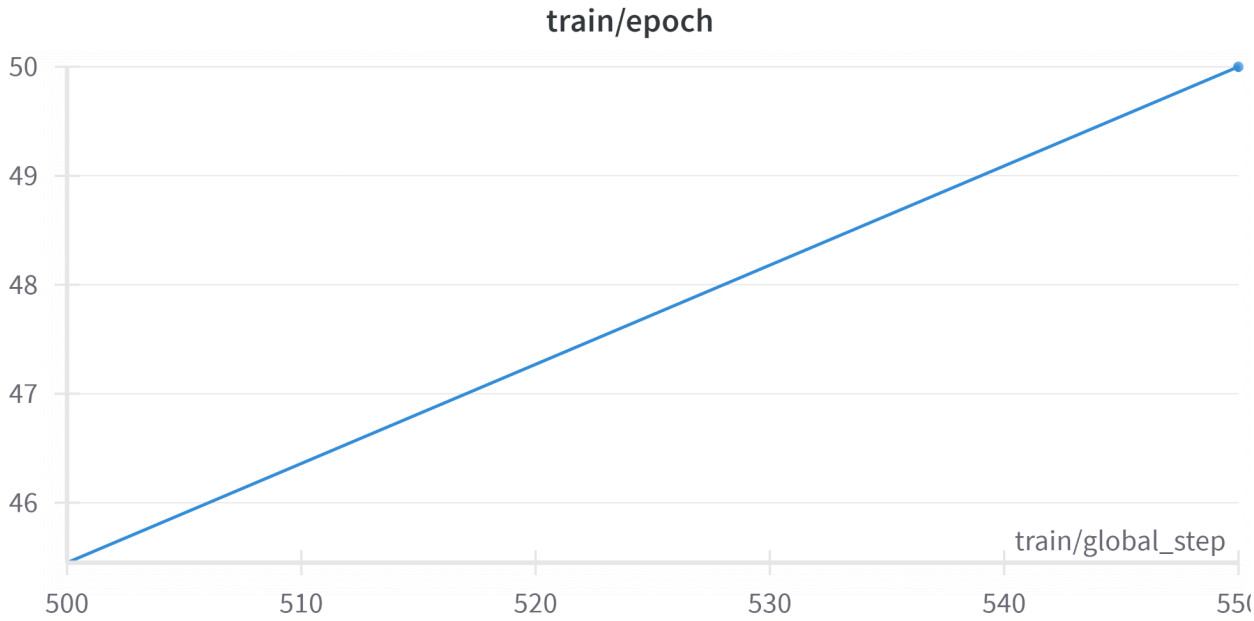
Bleu scores of fine-tuned GPT-2 model:

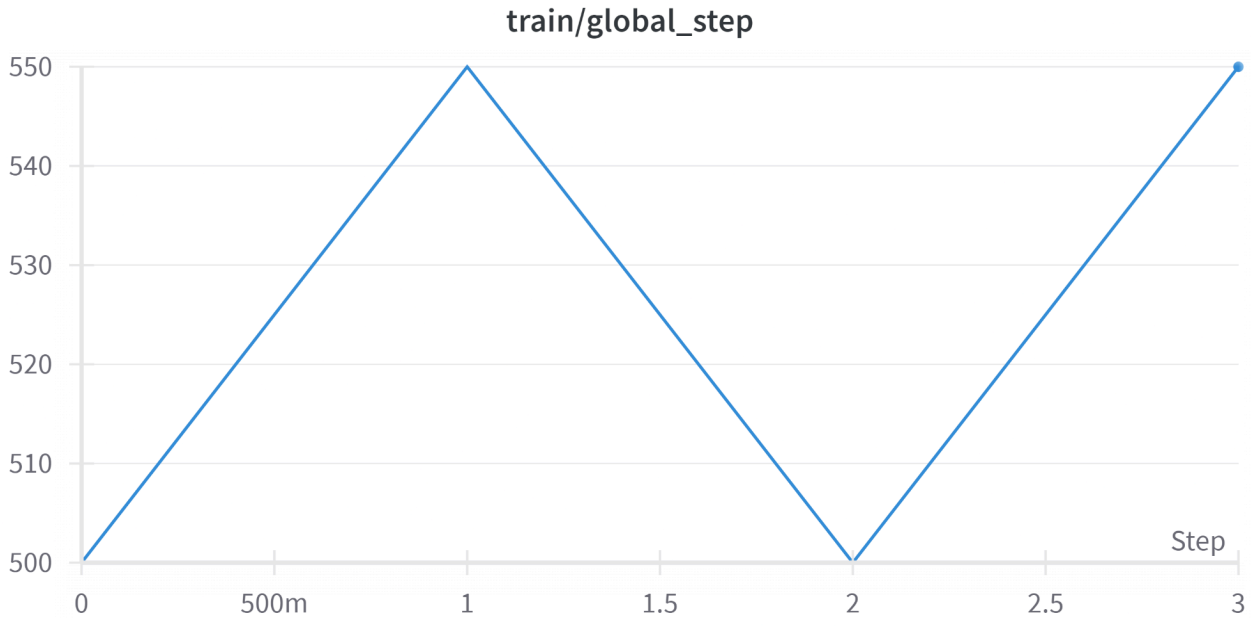
```
BLEU-1: 0.101141
BLEU-2: 0.318026
```

From the above Bleu-scores obtained, we can infer that

BLEU-1 (0.1): This is a low score, indicating that the generated text has a relatively low unigram (single word) overlap with the reference text(s). A BLEU-1 score of 0.1 suggests that only about 10% of the individual words in the generated text match those in the reference text(s).
BLEU-2 (0.31): This is a higher score than BLEU-1, indicating that the generated text has a better bigram (two-word sequence) overlap with the reference text(s). A BLEU-2 score of 0.31 suggests that around 31% of the two-word sequences in the generated text match those in the reference text(s).

- WANDB GRAPHS





These are the graphs obtained while fine-tuning the corpus on GPT-2 model.