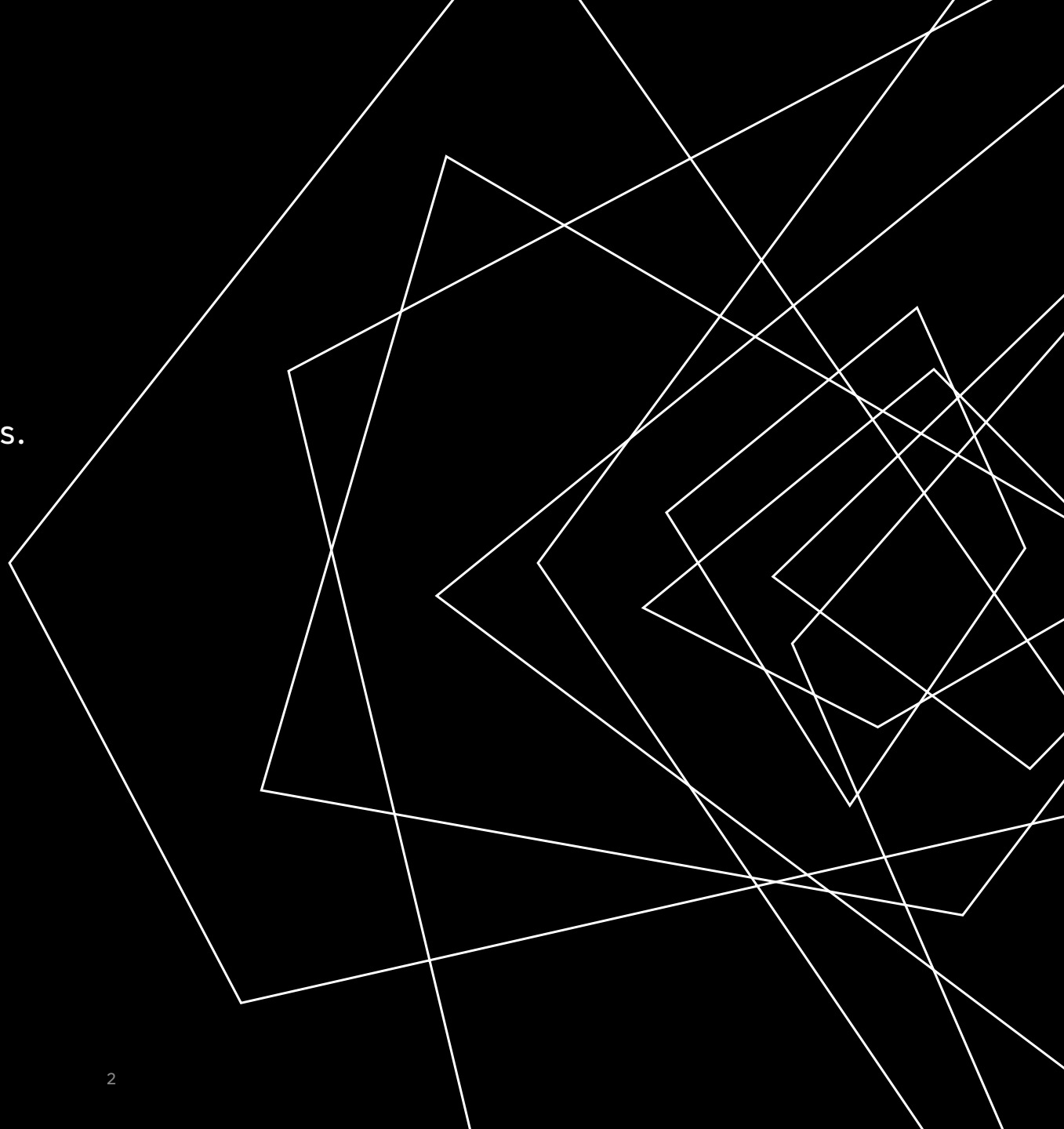
GRAMMATICAL ERROR CORRECTION

Names to be added

INTRODUCTION

- The goal of this project is to design an NLP based model for automatic Grammatical Error Corrections.
- And to design a robust model, the data with equal capability is needed, so for this C4-200M dataset was considered as it contains millions of instances with Correct and Incorrect sentences.
- And the pre-trained model used for this is T-5, which is a text-to-text based approach.



DATA PREPARTION AND PREPROCESSING

- The C4_200M dataset contains 10 files each with a set of sentences and labels, from that we considered one file named: C4_200M.tsv.0000-of-00010, which contain 30000 rows.
- Only one file out of 10 is used by taking the resources and available computational power into consideration, as more data may require more resources.
- The data in CSV file was inserted to a data frame and as an initial caution the null values are dropped, and the padding was done to balance the weights.
- And then the split of data for training as well for testing was done in a [90:10] manner.

MODEL TRAINING AND EVALUATION

- The **t5-base** model and tokenizer are loaded using **T5ForConditionalGeneration** and **T5Tokenizer**. A custom dataset class, **GrammarDataset**, is defined for handling the tokenization and preparation of data for training.
- A new column **input_token_len** is added to the test DataFrame to store the length of the tokenized inputs.
- Training arguments are set using **Seq2SeqTrainingArguments**. This includes settings like batch size, learning rate, number of epochs, evaluation strategy, etc.
- A custom **compute_metrics** function is defined to compute Rouge metrics for model evaluation.
- The **Seq2SeqTrainer** is initialized with the model, training arguments, training and evaluation datasets, tokenizer, and the data collator.
- The model is trained using the **train** method of **Seq2SeqTrainer**.

RESULTS

- After the training part was ran with all the parameters defined by setting epochs and metrics the loss table is shown here.

•As well some other results after the training are as follows: **Training Runtime**: 10,081.218 seconds (about 2.8 hours),

Training Samples per Second: 26.782.

Training Steps per Second: 0.279.

Total FLOPs (Floating Point Operations): Approximately 1.99 x 10^16.

Training Loss: 0.6712.

- And the manual testing for the model was done and the respective results are attached in the following slide.

Step	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
500	0.763400	0.628792	71.224700	60.843600	70.482700	70.513800	17.331700
1000	0.678100	0.603869	71.444600	61.254500	70.707600	70.743500	17.316600
1500	0.656000	0.591053	71.598400	61.518200	70.867500	70.903300	17.300400
2000	0.644400	0.585335	71.653500	61.625700	70.919300	70.954400	17.299000
2500	0.637800	0.582086	71.696400	61.697600	70.964500	71.000800	17.296600

MANUAL TEST RESULTS:

```
input_text = "I am enjoys, writtings Articles ons AI and I also enjoyed write articling on AI."
num_return_sequences = 1
corrected_text = correct_grammar(input_text, num_return_sequences)
print(corrected_text)
```

```
['I enjoy writing articles on AI and I also enjoy writing articles on AI.']
```

```
text =  """Today gift shows are popular in many countries, and purpose of these shows finds talented people, and hel

Firstly, result this programme has  a massive effect on the society, because many people get a chance to represent t

secondly, many audiences, and viewers watch this shows, so it is a big chance for companies by sponsoring in this pr

As a result, the aim of  producing this shows impressive, so part of the society following this shows for entertaini

"""
print(correct_grammar(text, num_return_sequences))
```

```
['Today gift shows are popular in many countries, and the purpose of these shows is to find talented people, and he
lp them to introduce themselves to each other.Actually, many people now watch these shows, and during this years fi
nd more fans that increase the Viewer, and many sponsors.']
```

THANK YOU