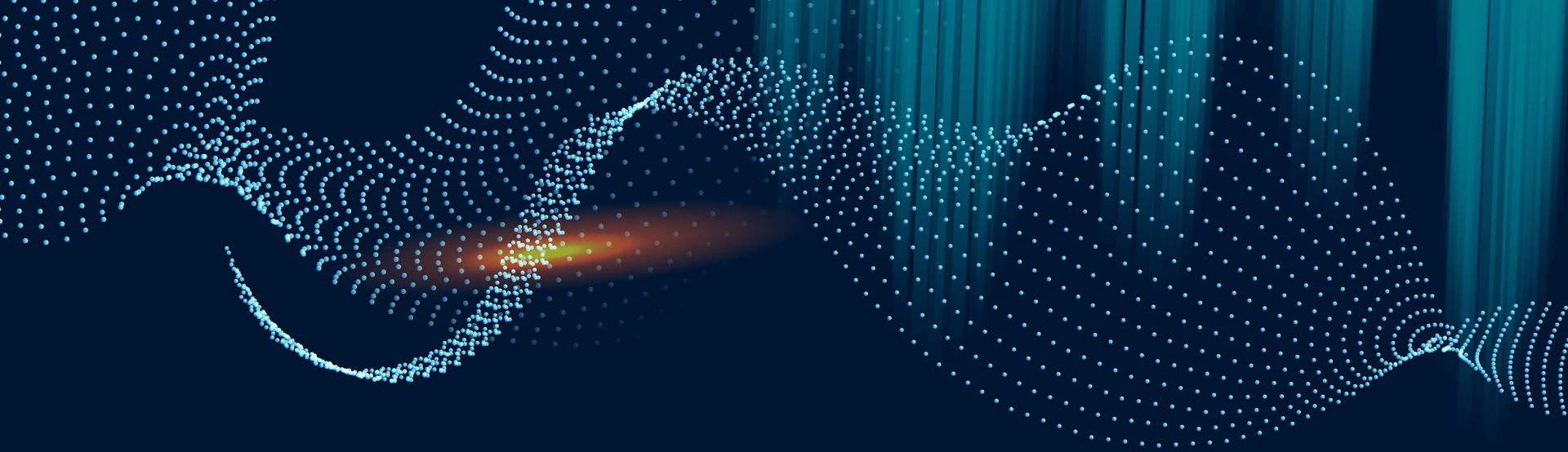
Generating Replies Based On TV Shows Scripts

Yomna Ayman
Yara Ayman

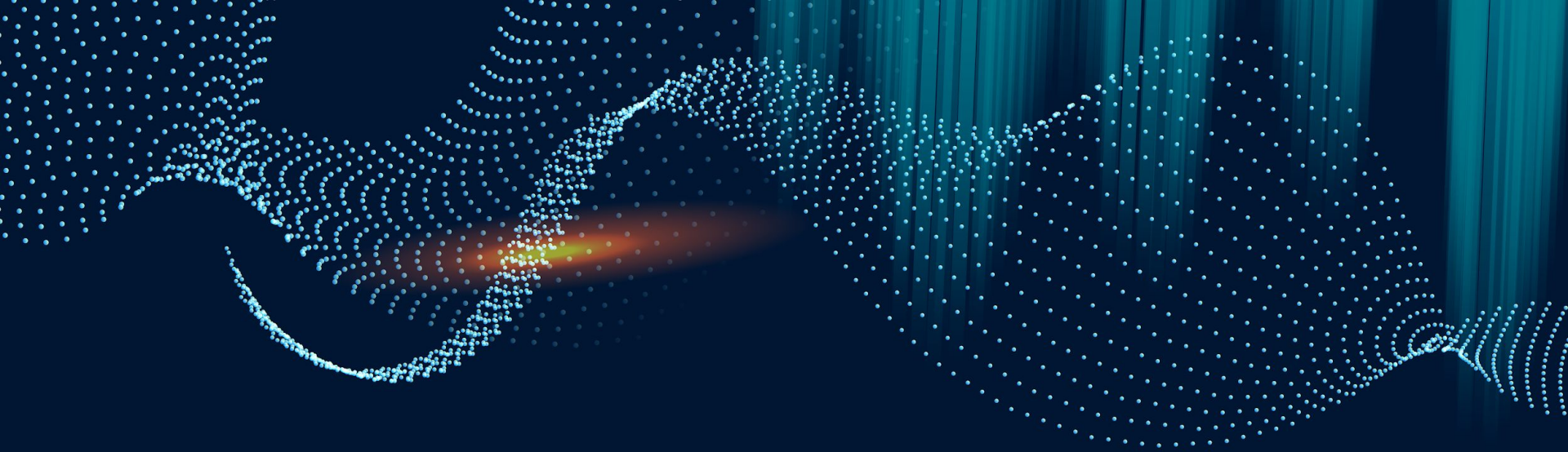


01 Introduction

Dataset Overview

Initial Dataset Columns:

- Release Date
- Season
- Episode
- Episode Title
- Name (Speaker)
- Sentence



02 | Milestone 2

Dataset Transformation for Milestone 2

To create a dataset suitable for training a conversational model:

- Each sentence is paired with the subsequent sentence to form a question and answer pair, only the top 5 speakers were considered.

Model Architecture Overview

Encoder-Decoder Architecture: Utilizes LSTM layers for sequence-to-sequence learning.

Components:

- Encoder: Processes input sequences and produces hidden states.
- Decoder: Generates output sequences based on the encoder’s states.

Encoder Architecture

- Input: Encoded sequences of fixed length (13 tokens).
- Embedding Layer: Converts token sequences into dense vectors.
- LSTM Layer: Processes the embedded vectors, returns sequences and states (hidden and cell states).

Decoder Architecture

- Input: Encoded sequences and initial states from the encoder.
- Embedding Layer: Converts token sequences into dense vectors.
- LSTM Layer: Processes embedded vectors, initialized with encoder’s states.
- Dense Layer: Produces output probabilities for each token.

Complete Model

- Combined Model: Connects encoder and decoder for end-to-end training.
- Loss Function: Categorical cross entropy.
- Optimizer: Adam optimizer.

Model Training

	Epoch 1	Epoch 25	Epoch 50	Epoch 75	Epoch 100	Epoch 145
Loss	4.0701	2.1900	1.3473	0.8451	0.3992	
Accuracy	35.53%	52.89%	69.74%	82.06%	91.52%	98.96%

Steady improvement observed over 150 epochs.

Significant reduction in loss and consistent increase in accuracy.

Examples

– Query: you are a dwarf

– answer: i am a monster as well as a dwarf you should

– Query: for all the dead may never die

– answer: they are right take him alive kill his men

– Query: what are we going to do

– answer: someone does work that poor years to see the city OUT

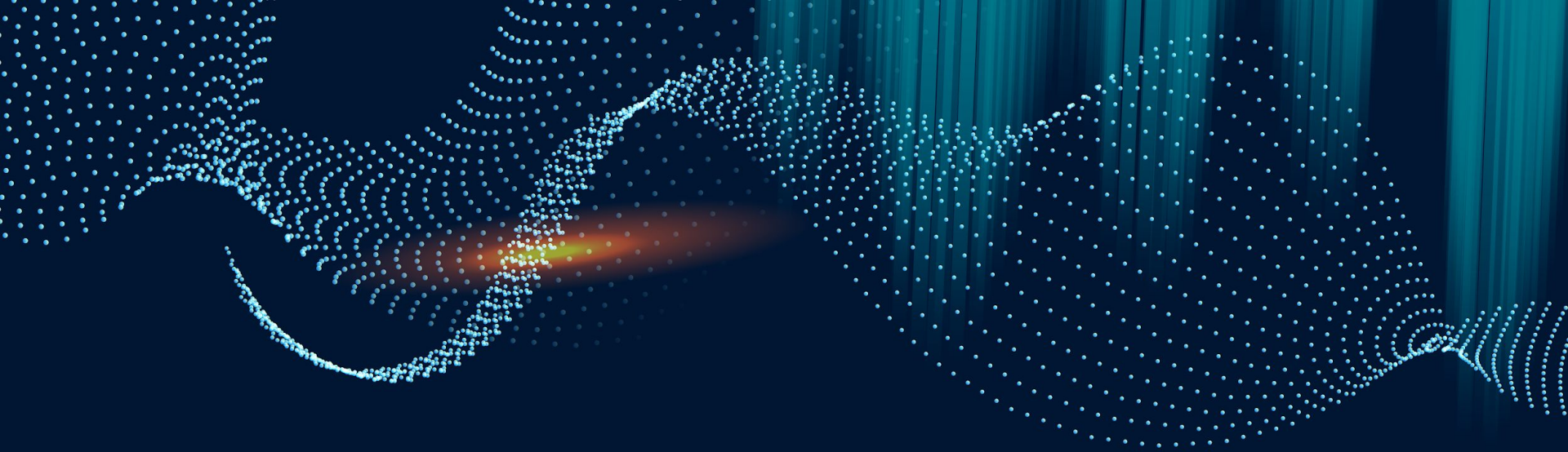
– Query: valar morghulis.

– answer: valar <OUT>

Limitations

Limited Vocabulary:

- The model's vocabulary might miss important words, showing them as an <OUT> token.
- Fixed sequence length could cut off essential information or add unnecessary padding.
- High Computational Cost:
- Training with 400 LSTM units over 150 epochs makes the model resource-intensive.
- Without proper regularization, the model might overfit, performing poorly on new data.
- Basic Text Processing:
- Simple text cleaning might not handle all language variations effectively.
- Starting embeddings from scratch instead of using pre-trained ones can hinder performance.



03 | Milestone 3

Model Choice & Data Preparation

We Used flan-T5 -base model.

- Data Loading: The dataset is loaded into a pandas DataFrame with columns: question, answer, and speaker.
- Conversion to Dataset Object: The DataFrame is converted to a Hugging Face Dataset object.
- Dataset Splitting: The dataset is split into training and testing sets with an 80/20 ratio.

Preprocessing

- Prefix Addition: Each question is prefixed with "Please answer this question as {speaker}:" to provide context.
- Tokenization: Questions and answers are tokenized using a specified tokenizer, with a maximum input length of 128 tokens for questions and 512 tokens for answers.
- Function Mapping: The preprocessing function is mapped across the entire dataset to apply tokenization.

Training Configuration

- Global Parameters:
- Learning Rate: 3e-4
- Batch Sizes: 4 for training, 2 for evaluation
- Weight Decay: 0.01
- Save Limit: 3 checkpoints
- Number of Epochs: 4
- Training Arguments: Defined using Seq2SeqTrainingArguments with configurations for output directory, evaluation strategy, learning rate, batch sizes, weight decay, save limit, number of epochs, and gradient accumulation steps.

Metrics & Evaluation

ROUGE Metric: The ROUGE metric is used to evaluate the performance of the model by comparing the generated responses to the reference answers.

[4780/4780 2:04:36, Epoch 3/4]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
0	3.070800	2.832194	0.085000	0.013975	0.081284	0.083144
1	2.773500	2.773306	0.087862	0.016973	0.084225	0.085995
2	2.561200	2.767051	0.092374	0.018433	0.087919	0.089761
3	2.427900	2.784258	0.092910	0.019242	0.088748	0.090310

Examples

– Query: valar moghulis
– Speaker: arya stark

– answer: Valar dohaeris.

– Query: valar moghulis
– Speaker: daenerys targaryen

– answer: Is that a good man?

– Query: the white walkers, what about them??
– Speaker: jon snow

– answer: They're not going to die.

– Query: what happened to you?
– Speaker: sansa stark

– answer: I was stabbed in the back.

Conclusion

Alignment with Conversational Tasks:

- FLAN-T5 pre-trained on conversational data, ideal for diverse speaker question-answering.
- Effective incorporation of conversational context yields contextually relevant responses.
- Advantages Over Traditional Models:
- FLAN-T5 captures subtle speaker behaviors, enhancing context appropriateness.
- Traditional models lack explicit dialogue training, struggling with conversational nuances.

Challenges and Limitations:

- Fine-tuning FLAN-T5 demands significant computational resources and data annotation.
- Model may encounter issues with ambiguity, response diversity, and potential biases.
- Utilizes a dataset of ~23,900 lines, relatively modest in size.
-

THANKS!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.