A Comparative Study of Transformer-Based Language Models on a Project Specifications Understanding

BY:
Amen Allah Berrejeb
Supervised by:
Mr Mohamed Aziz Atitallah

Table of contents

1

Hosting Company

The company and its different departments

2

Research Concepts

A Deep Dive into the Research Concepts : Theory and Maths

3

Design and Development

4

Conclusion and Future Improvements

1

THE Hosting Company

In this section , I will describe the company and its different departments.

M&C IT CONSULTING

Main goals :

- provide IT solutions to Airlines
- Passenger Revenue Accounting
- Intelligent Process Outsourcing

mnc.aero

The departements

M&C is divided into 3 main departments:

- **ADAD Project Development & Management**
- RPA Daily services
- Cruise Line Project

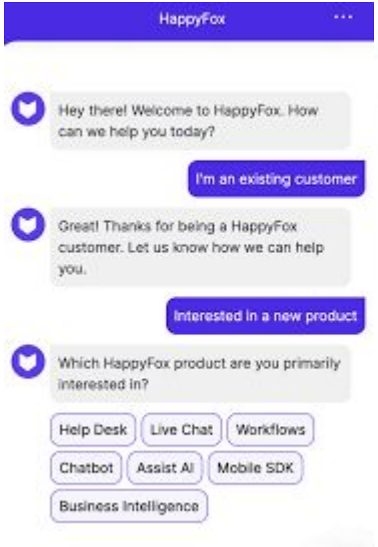
Research Concepts

2

- A brief Introduction to NLP tasks (motivation)
- Insights taken from Research papers
- Maths Behind Evaluation Metrics

Motivation (1/2)

- **NLP** is a subfield of Artificial Intelligence concerned with the **interactions** between **computers** and **human natural languages**
- **Some** applications in this field :
  - Smart Reply in Gmail Box
  - Google Traduction
  - Sentiment Analysis
  - Inappropriate Content Detection
  - **Question Answering**
  - **Language Understanding**



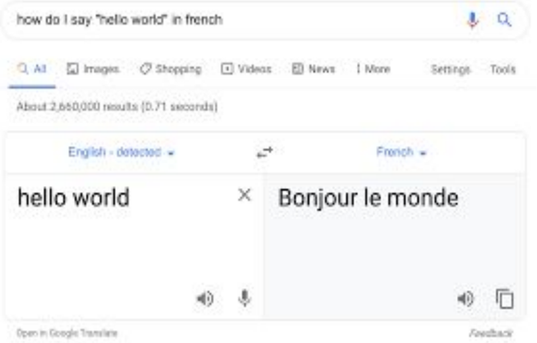
Hey there! Welcome to HappyFox. How can we help you today?

I'm an existing customer

Great! Thanks for being a HappyFox customer. Let us know how we can help you.

Interested in a new product

Which HappyFox product are you primarily interested in?



how do I say "hello world" in french

About 2,660,000 results (0.71 seconds)

hello world

Bonjour le monde

Motivation (2/2)

- **Question Answering** and **Language Understanding** are both used to resolve the problem of getting insights from the company’s text data and answer clients and developers.
- **Added** values to the company :
  - Multiple chatbots that can answers both clients and developers questions
  - First step into the AI Research & Development Department
  - Building an intelligent agent similar to chatgpt with M&C’s Private Data
  - Introducing New AI concepts and Tools to software engineers and developers

Keywords of Research

Word Representations

Attention Mechanism and Transformers

BERT / ALBERT

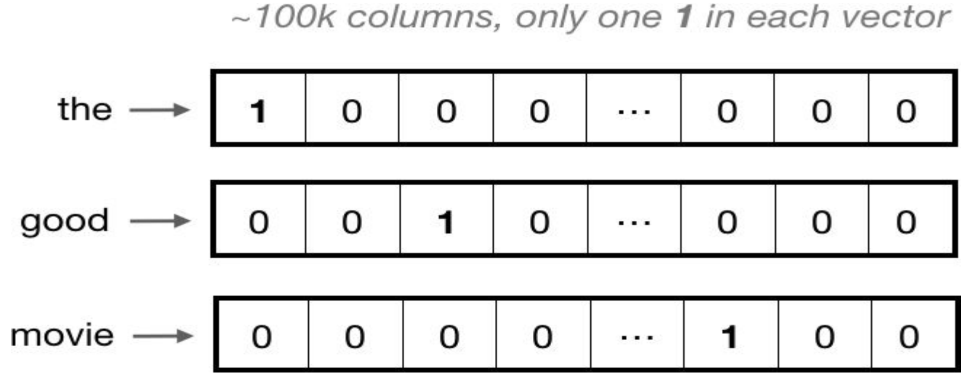
GPT :Generative Pre Trained Transformer

T5 :Text to Text Transfer Transformer

Evaluation Metrics (Mathematical Equations)

Word Representations (1/2)

- **Key idea:** Representing each word as a vector that has as many as values as there words in the vocabulary. Each column represents a word in a vocabulary



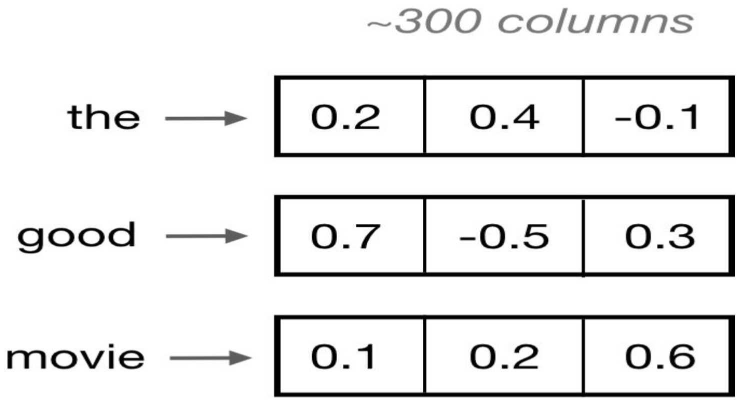
*~100k columns, only one **1** in each vector*

the →	**1**	0	0	0	⋯	0	0	0
good →	0	0	**1**	0	⋯	0	0	0
movie →	0	0	0	0	⋯	**1**	0	0

- For **large vocabularies**, these vectors can get very **long**, and they contain all 0’s except for one value. This is considered a very sparse representation.

Word Representations (2/2)

- **Key idea:** Map the word index using pre trained models and pre trained word embeddings like : Word2Vec, Glove.

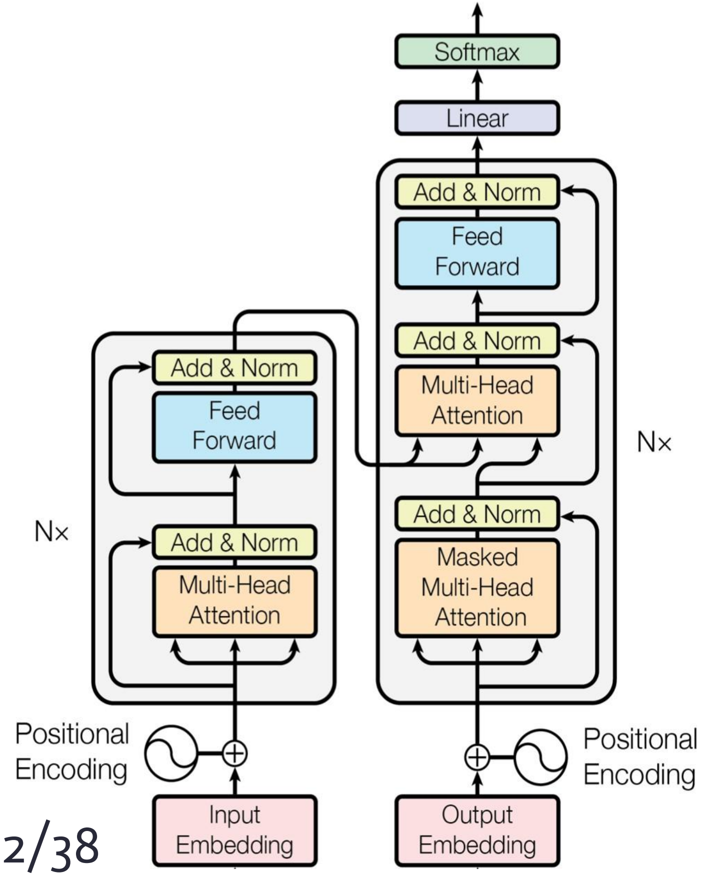


~300 columns

the →	0.2	0.4	-0.1
good →	0.7	-0.5	0.3
movie →	0.1	0.2	0.6

The word embedding can be trained **end-to-end** with the model for the downstream tasks.

Attention Mechanism : Transformer (1/2)



- Each **encoder** block has two sub-layers:
  - The First is a **mutli-head self-attention** mechanism.
  - The second is a position-**wise Fully connected feed-forward** network.
- Each **decoder** has additional third sub -layer:
  - The Third is a multi-head attention over the output of the encoder stack.
- A residual connection is added around each of the two sub-layers, followed by layer normalization:

$$\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x))$$

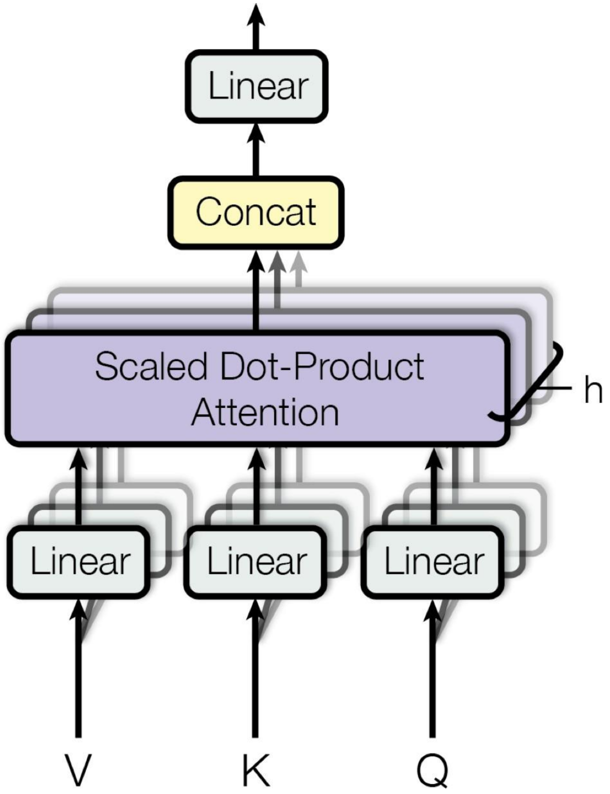
Attention Is All You Need [Vaswani *et al.*, 2017]

Attention Mechanism : Multi Head (2/2)

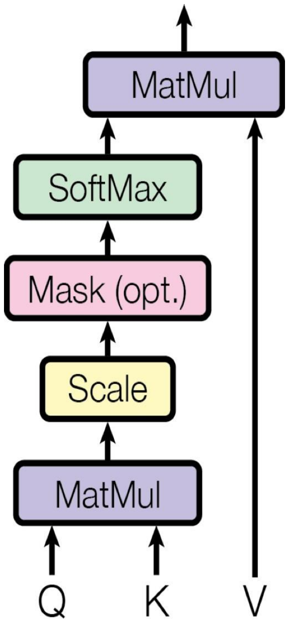
- **Project** Q, K and V with h **different**, learned linear projections.
- Perform the **scaled dot-product attention** function on each of these projected versions of Q, K and V **in parallel**.
- **Concatenate** the output values.
- **Project** the output values again, resulting in the final values.

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})\boxed{W^O}$$

$$\text{where } \mathrm{head_i} = \mathrm{Attention}(Q\boxed{W_i^Q}, K\boxed{W_i^K}, V\boxed{W_i^V})$$



Scaled Dot-Product



$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

BERT

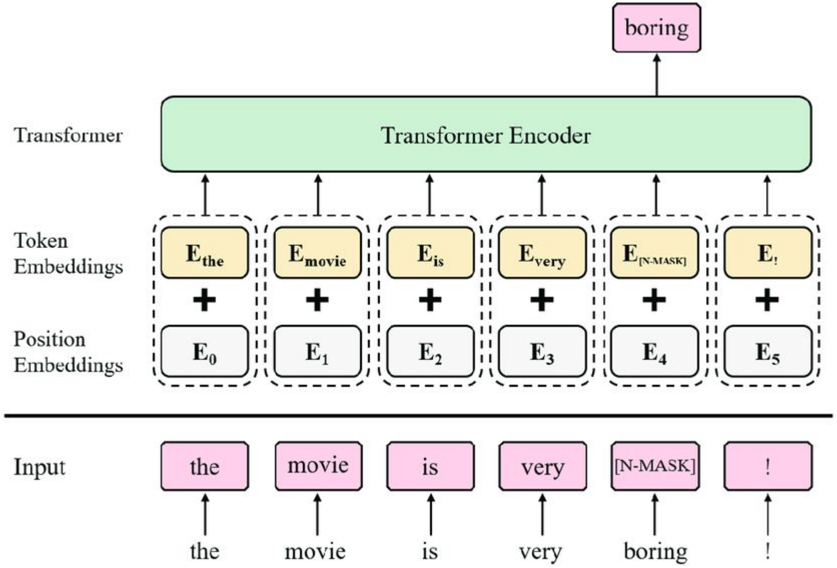


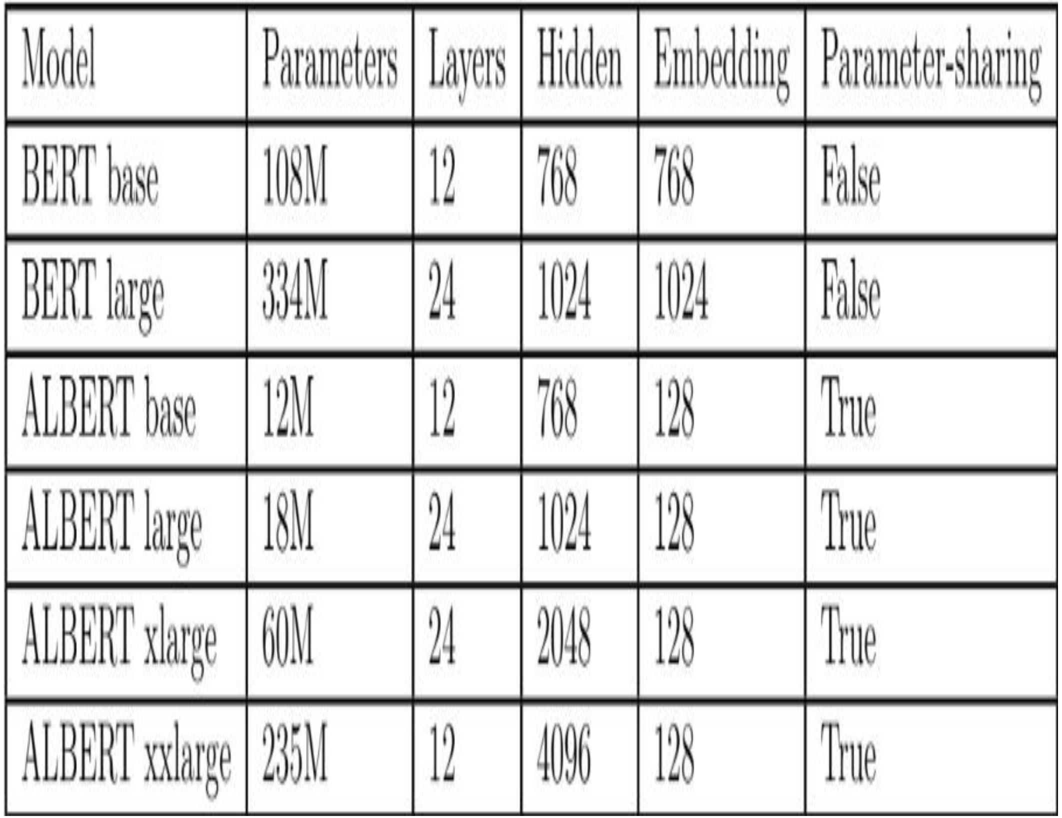
Image credit: w

- Task #1: **Masked Language Model (MLM)**
  - Mask some percentage (15%) of the input tokens at random.
  - Predict those masked tokens.

- Task #2: **Next Sentence Prediction (NSP)**
  - Binarized next sentence prediction task.
  - When choosing the sentences A and B for each pre-training example,
    - 50% of the time B is the actual next sentence that follows A (**IsNext**).
    - 50% of the time it is a random sentence from the corpus (**NotNext**).

BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [Devlin *et al.*, 2018]

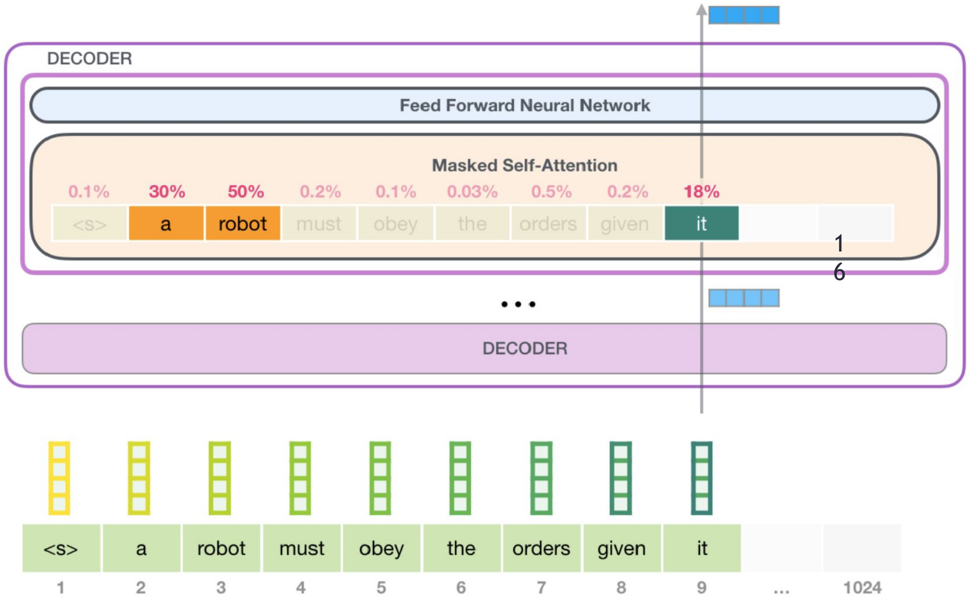
ALBERT : A Lite Bert



Model	Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT base	108M	12	768	768	False
BERT large	334M	24	1024	1024	False
ALBERT base	12M	12	768	128	True
ALBERT large	18M	24	1024	128	True
ALBERT xlarge	60M	24	2048	128	True
ALBERT xxlarge	235M	12	4096	128	True

- Replaces BERT’s Next Sentence Prediction with a **Sentence Order Prediction**
- Employs **cross-layer parameter** sharing to reduce the number of parameters, addressing BERT's **large memory footprint**
- Introduces techniques for **reducing model size** without affecting performance

Casual Language Models (CLM) :GPT



1
6

- System works in two stages:
  - First, **pre-train** a transformer model on a very large amount of data in an unsupervised manner (language modeling):

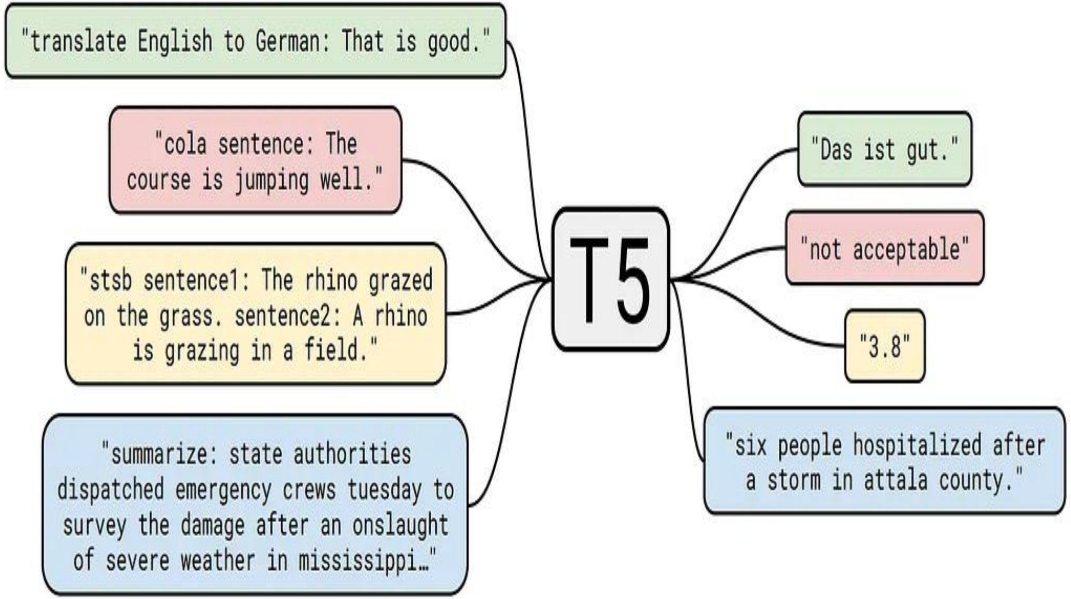
$$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \dots, u_{i-1}; \Theta)$$

  - Then, **fine-tune** this model on much smaller supervised datasets to solve specific tasks.

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m)$$

Improving Language Understanding by Generative Pre-Training [Radford *et al.*, 2018]

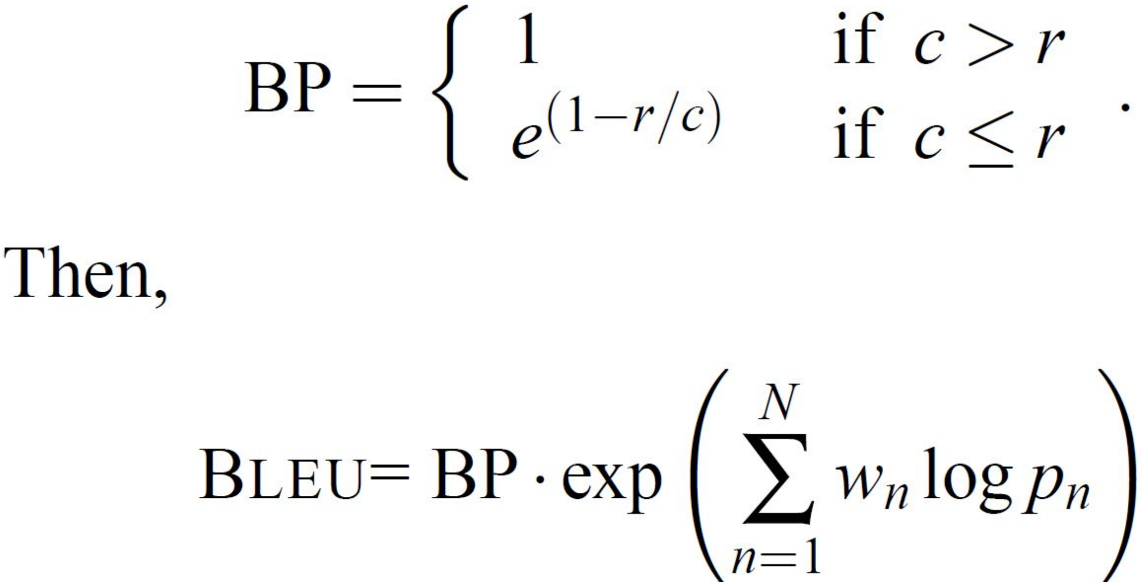
T5: Text to Text Transfer Transformer


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

- Converts all NLP problems into a **text-to-text** format, where both input and output are strings of text
- **Pre-trained** on a large text corpus using a denoising objective, similar to **BERT**
- T5 excels in multiple NLP tasks, including **translation** and **summarization**, due to its flexible **text-to-text** format

Evaluation Metrics : Bleu Score (1/2)

- **Key idea:** measure the quality of machine-generated answers by comparing it with human answers



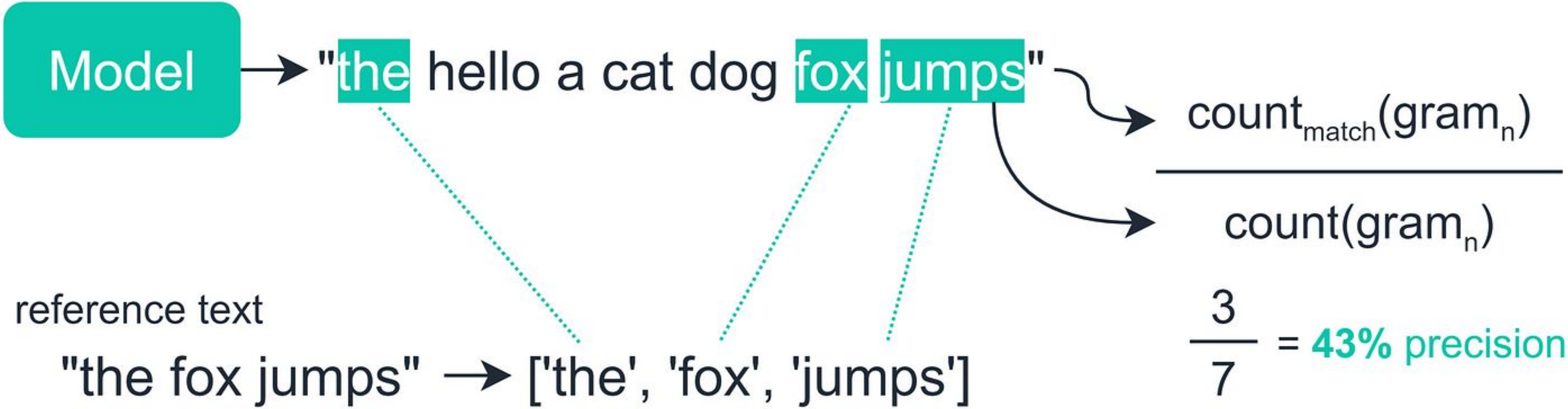
$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

Evaluation Metrics : Rouge Score (2/2)

- **Key idea:** primarily focusing on the recall of overlapping n-grams, word sequences, and word pairs.

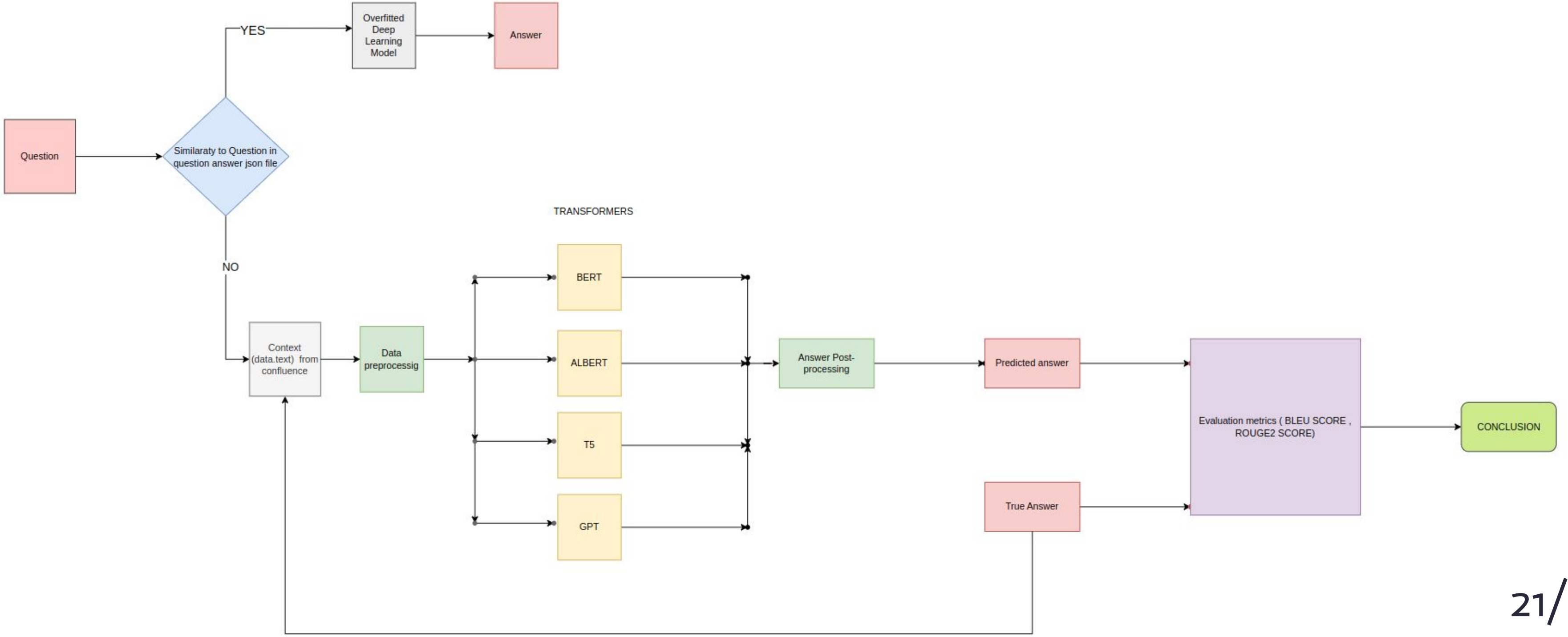

Model
"the hello a cat dog fox jumps"
$\frac{\text{count}_{match}(\text{gram}_n)}{\text{count}(\text{gram}_n)}$
reference text
"the fox jumps" → ['the', 'fox', 'jumps']
$\frac{3}{7}$ = 43% precision

3

Modelization and Application

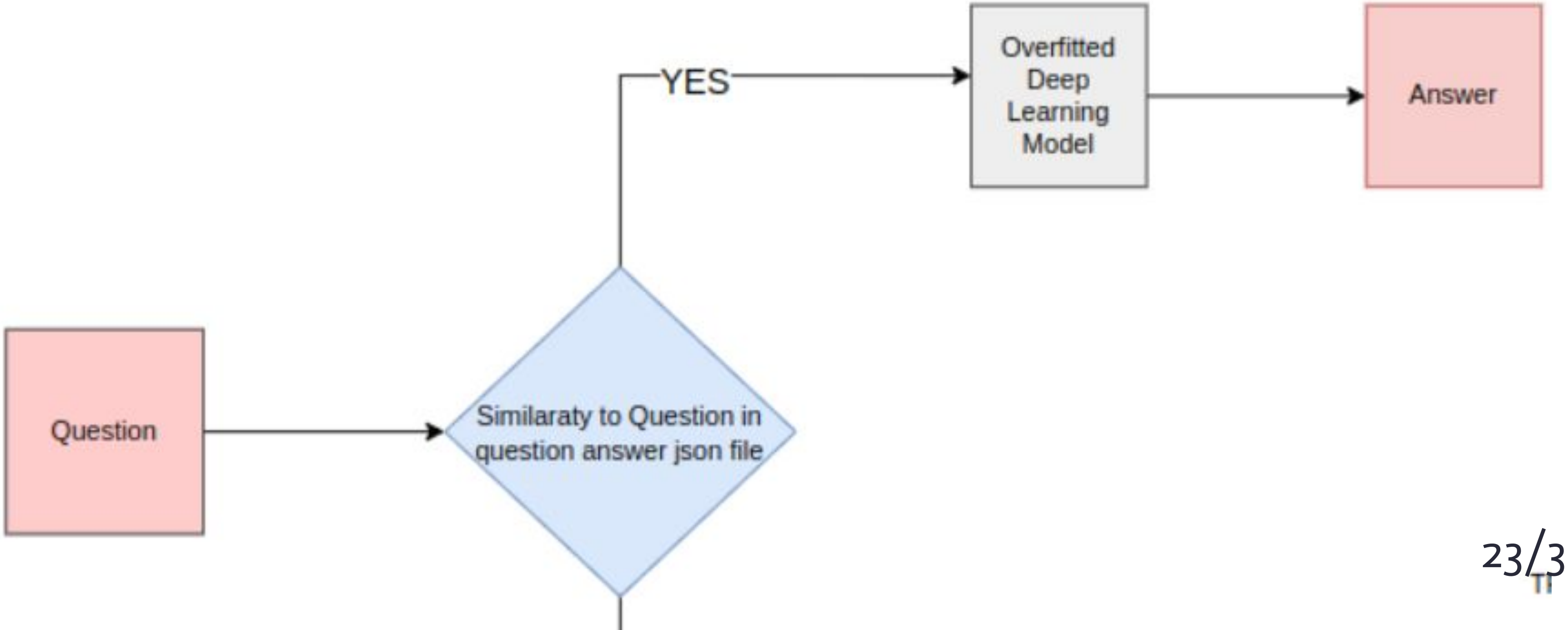
Converting Research Insights into code

Design of the Comparative Study

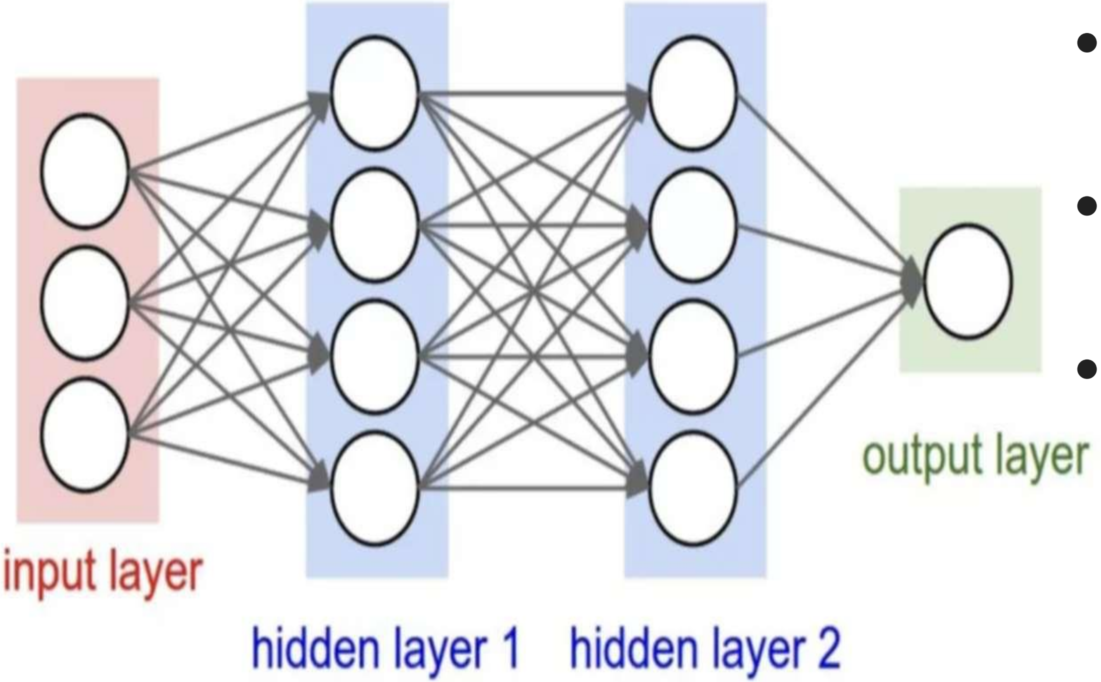

Question
Similaraty to Question in question answer json file
YES
Overfitted Deep Learning Model
Answer
NO
Context (data.text) from confluence
Data preprocessig
TRANSFORMERS
BERT
ALBERT
T5
GPT
Answer Post-processing
Predicted answer
True Answer
Evaluation metrics (BLEU SCORE , ROUGE2 SCORE)
CONCLUSION

Frameworks and Tools Used to Develop the Solution

First Part : Chatbot for Daily Conversation

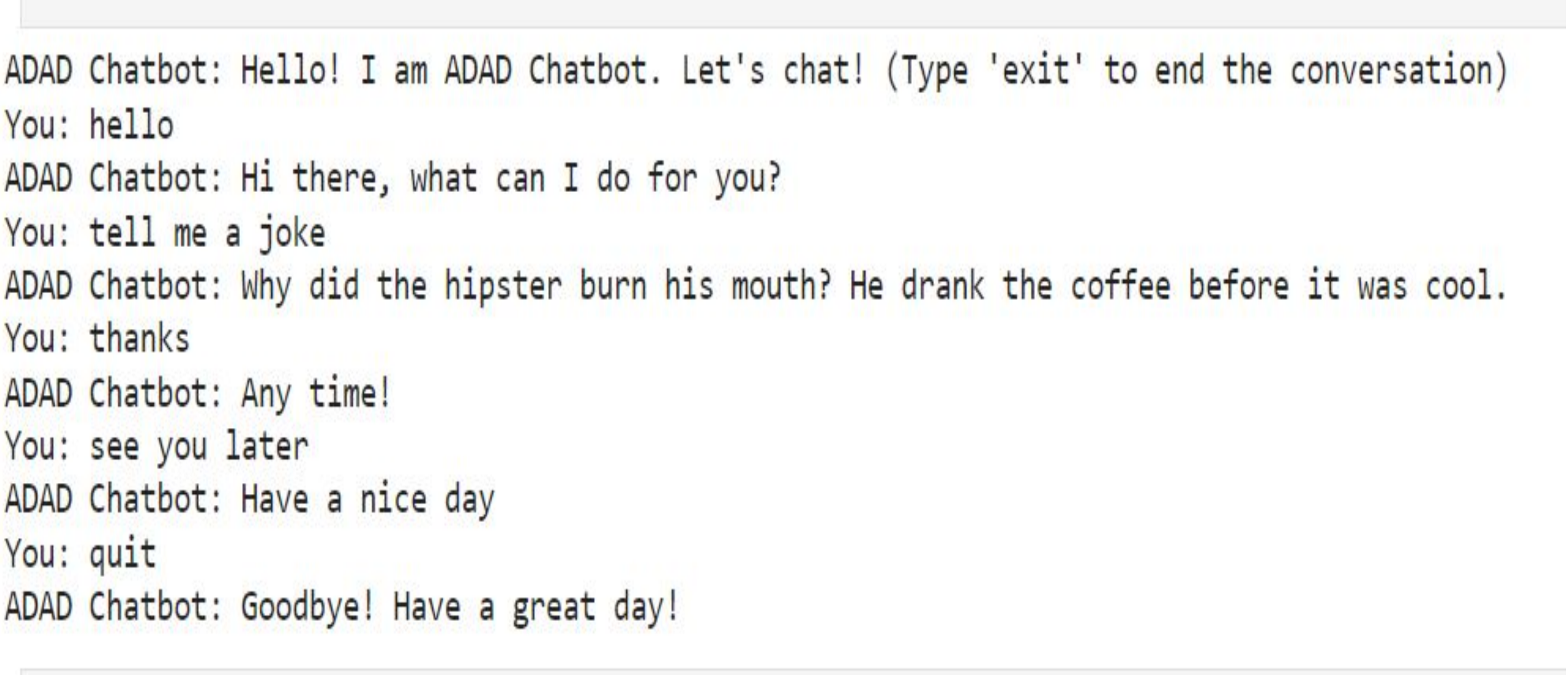

Question
Similaraty to Question in question answer json file
YES
Overfitted Deep Learning Model
Answer

Deep Learning Based chatbot (1/2)



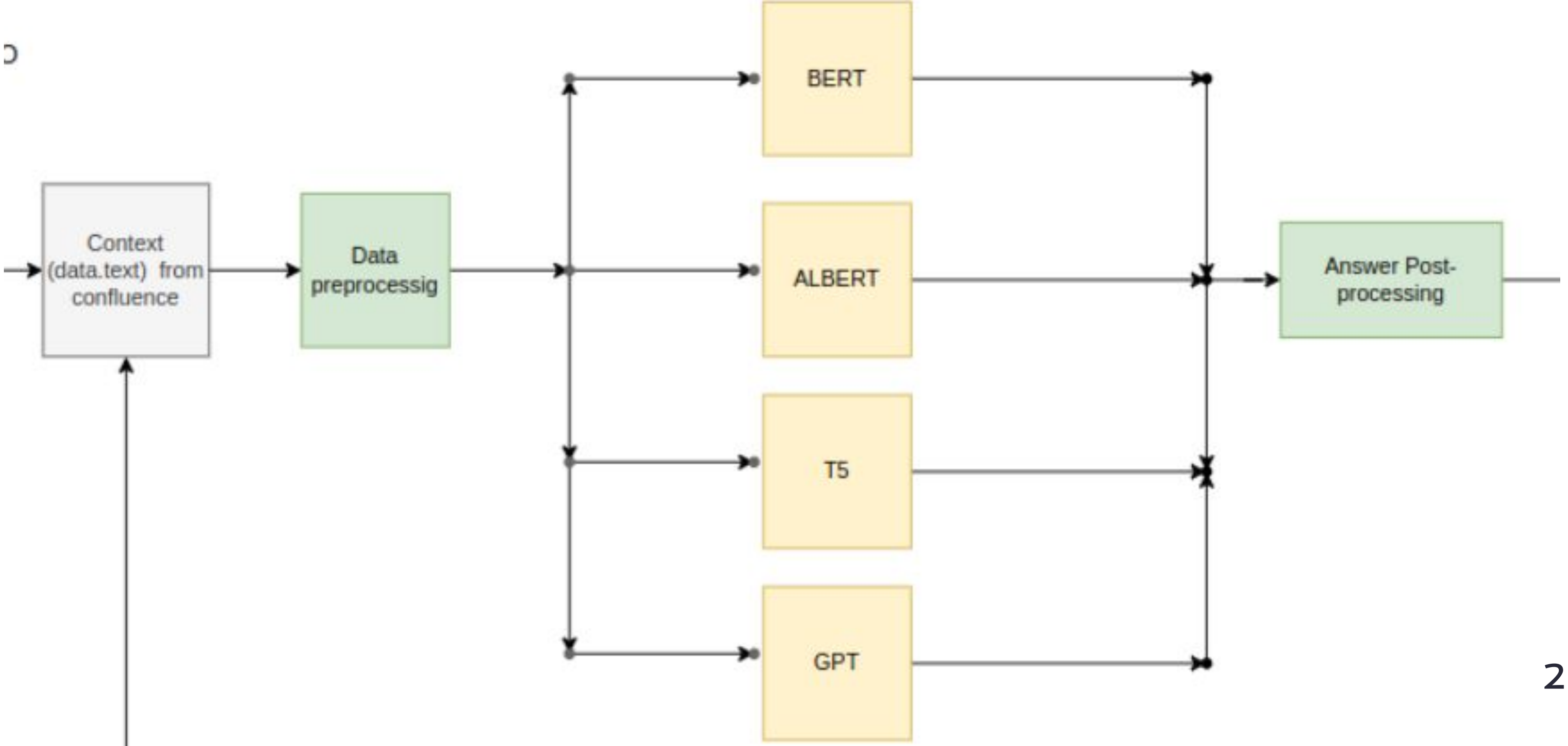
- The data is stored in JSON files, each containing pairs of questions and answers.
- The deep learning architecture is composed of three linear layers to process input data
- This structure is commonly used in building chatbots for daily conversations

Deep Learning Based chatbot (2/2)



```
ADAD Chatbot: Hello! I am ADAD Chatbot. Let's chat! (Type 'exit' to end the conversation)
You: hello
ADAD Chatbot: Hi there, what can I do for you?
You: tell me a joke
ADAD Chatbot: Why did the hipster burn his mouth? He drank the coffee before it was cool.
You: thanks
ADAD Chatbot: Any time!
You: see you later
ADAD Chatbot: Have a nice day
You: quit
ADAD Chatbot: Goodbye! Have a great day!
```

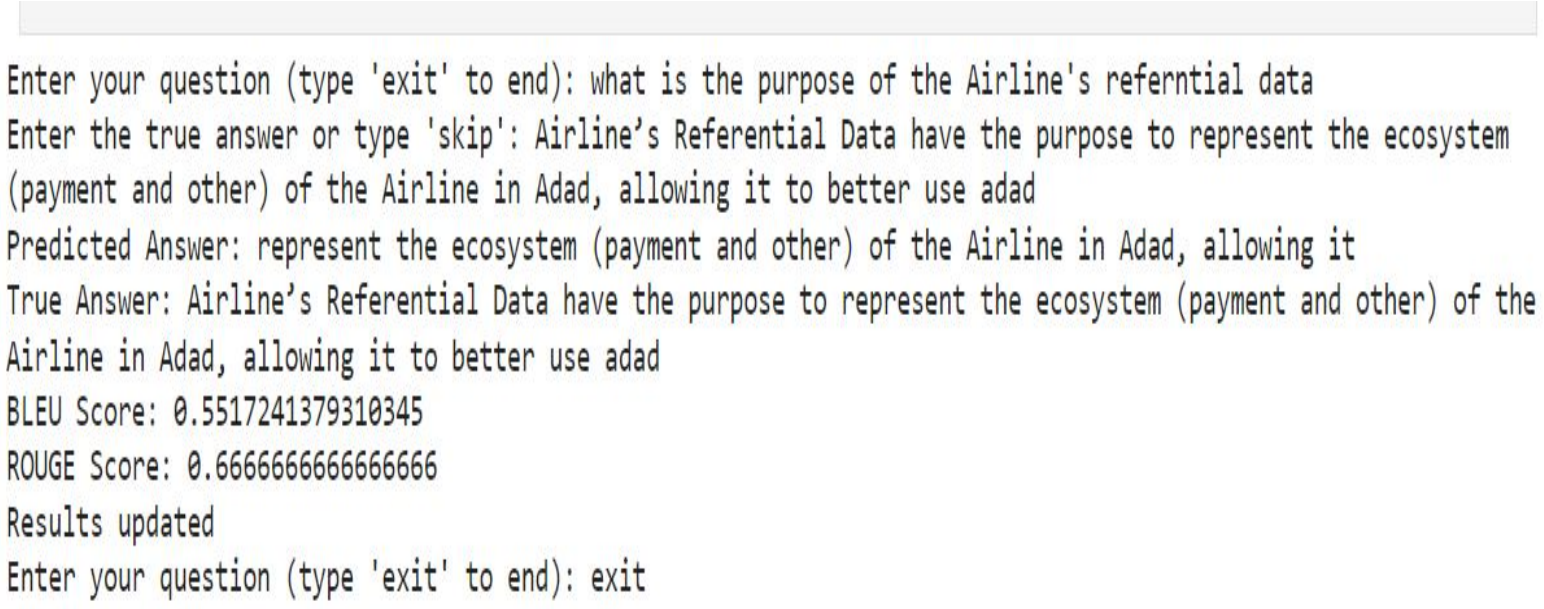
Second Part : Chatbots Based on Language Models (1/2)



Chatbots Based on Language Models (2/2)

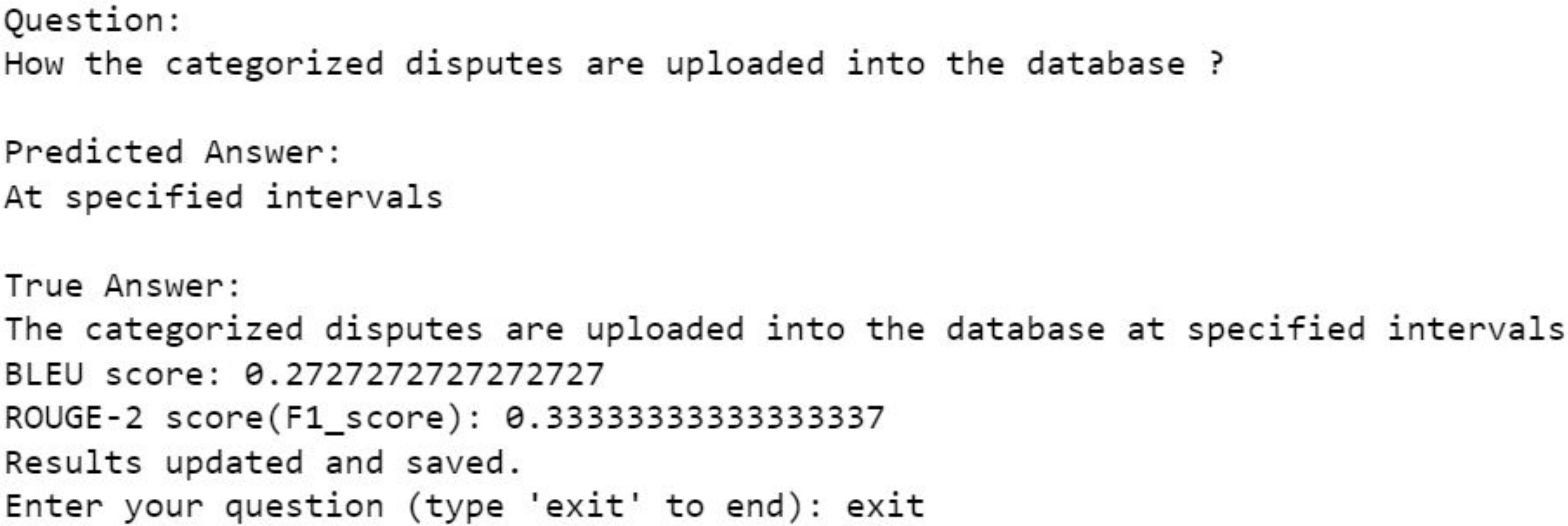
- **Key idea:** Chatbots Can understand question and answer it by context
- **Dataset**: The dataset used in this part are paragraphs from the ADAD Specifications Confluence (tool used in documentation for developers and clients)

T5 Chatbot



```
Enter your question (type 'exit' to end): what is the purpose of the Airline's referntial data
Enter the true answer or type 'skip': Airline’s Referential Data have the purpose to represent the ecosystem
(payment and other) of the Airline in Adad, allowing it to better use adad
Predicted Answer: represent the ecosystem (payment and other) of the Airline in Adad, allowing it
True Answer: Airline’s Referential Data have the purpose to represent the ecosystem (payment and other) of the
Airline in Adad, allowing it to better use adad
BLEU Score: 0.5517241379310345
ROUGE Score: 0.6666666666666666
Results updated
Enter your question (type 'exit' to end): exit
```

Bert Chatbot

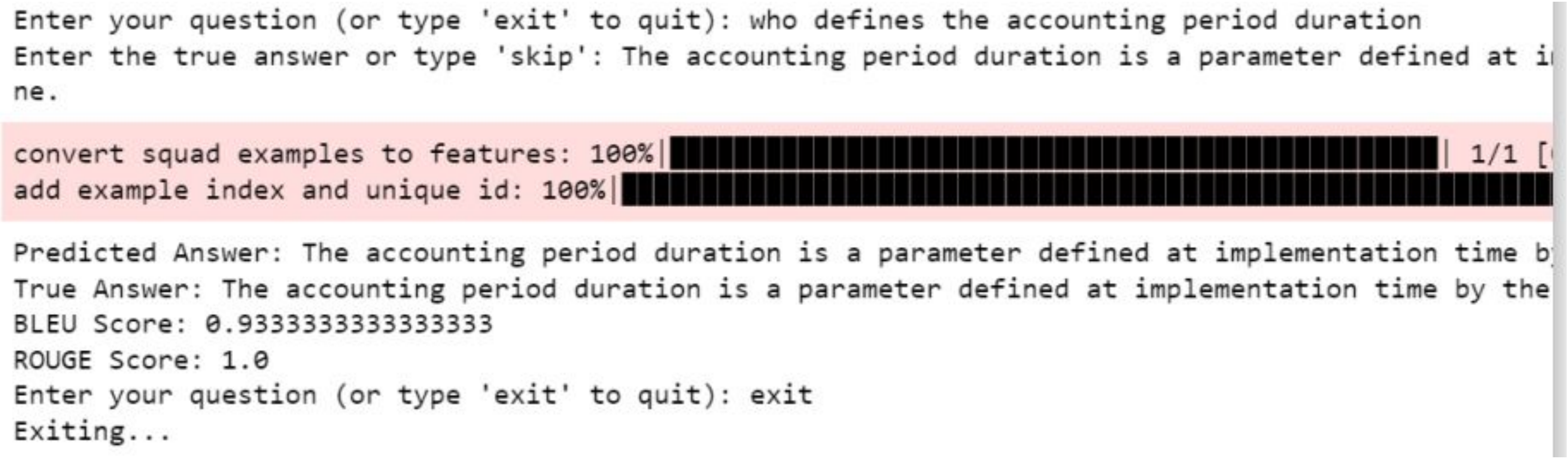


```
Question:
How the categorized disputes are uploaded into the database ?

Predicted Answer:
At specified intervals

True Answer:
The categorized disputes are uploaded into the database at specified intervals
BLEU score: 0.2727272727272727
ROUGE-2 score(F1_score): 0.3333333333333337
Results updated and saved.
Enter your question (type 'exit' to end): exit
```

ALBERT Chatbot



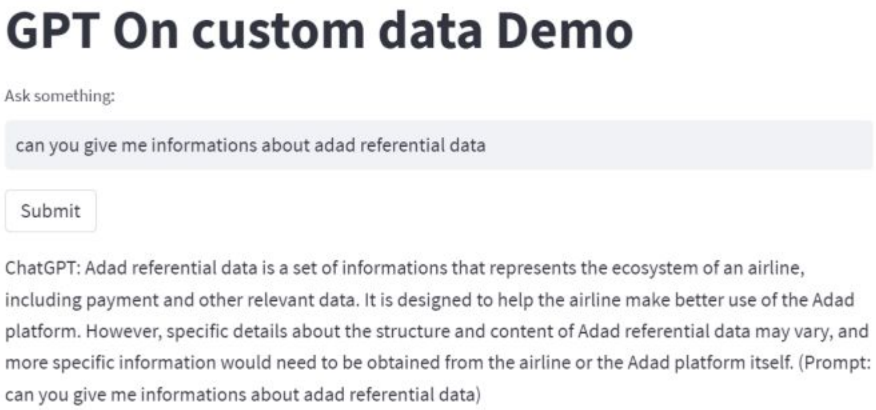
```
Enter your question (or type 'exit' to quit): who defines the accounting period duration
Enter the true answer or type 'skip': The accounting period duration is a parameter defined at i
ne.

convert squad examples to features: 100%|████████████████████████████████████████████████| 1/1 [
add example index and unique id: 100%|██████████████████████████████████████████████████████████

Predicted Answer: The accounting period duration is a parameter defined at implementation time b
True Answer: The accounting period duration is a parameter defined at implementation time by the
BLEU Score: 0.9333333333333333
ROUGE Score: 1.0
Enter your question (or type 'exit' to quit): exit
Exiting...
```

Special Use Case: OpenAI GPT

- **Problem :** GPT is not an open source model , so in order to use it we need to purchase OPENAI API
- **Solution** : Due to lack of Resources and to guarantee data privacy , I built a web app to treat some of the data using the API and deployed using Streamlit (Demo)



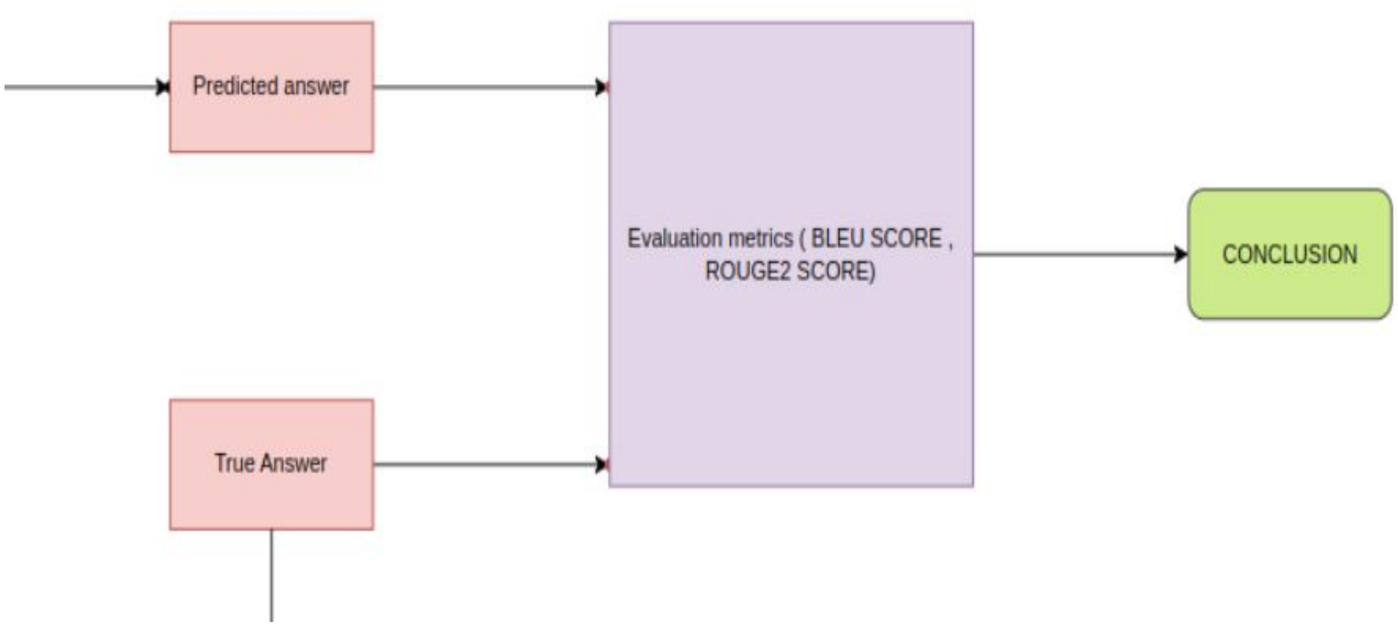
GPT On custom data Demo

Ask something:

can you give me informations about adad referential data

ChatGPT: Adad referential data is a set of informations that represents the ecosystem of an airline, including payment and other relevant data. It is designed to help the airline make better use of the Adad platform. However, specific details about the structure and content of Adad referential data may vary, and more specific information would need to be obtained from the airline or the Adad platform itself. (Prompt: can you give me informations about adad referential data)

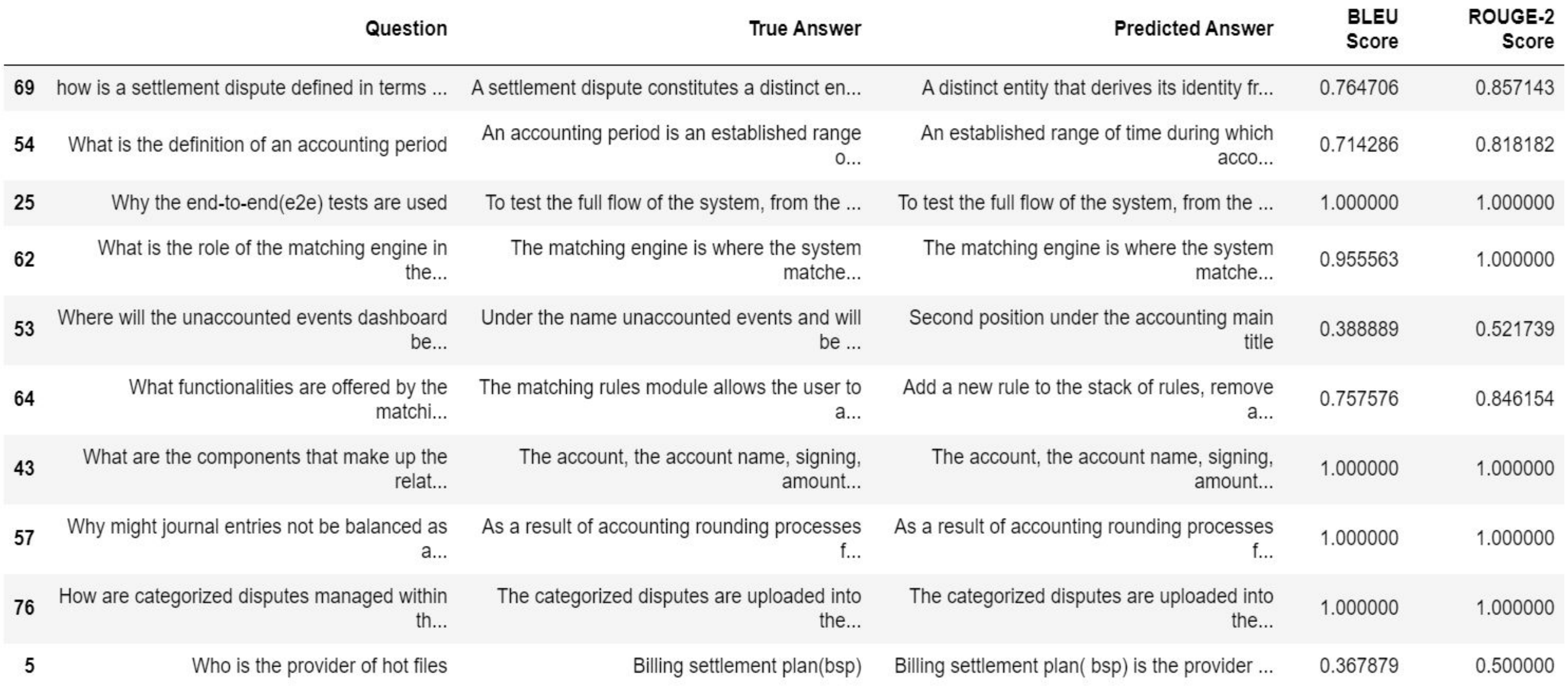
Final Part : Comparative Study of Chatbots Based on Language Models (1/2)



Comparative Study of Chatbots Based on Language Models (2/2)

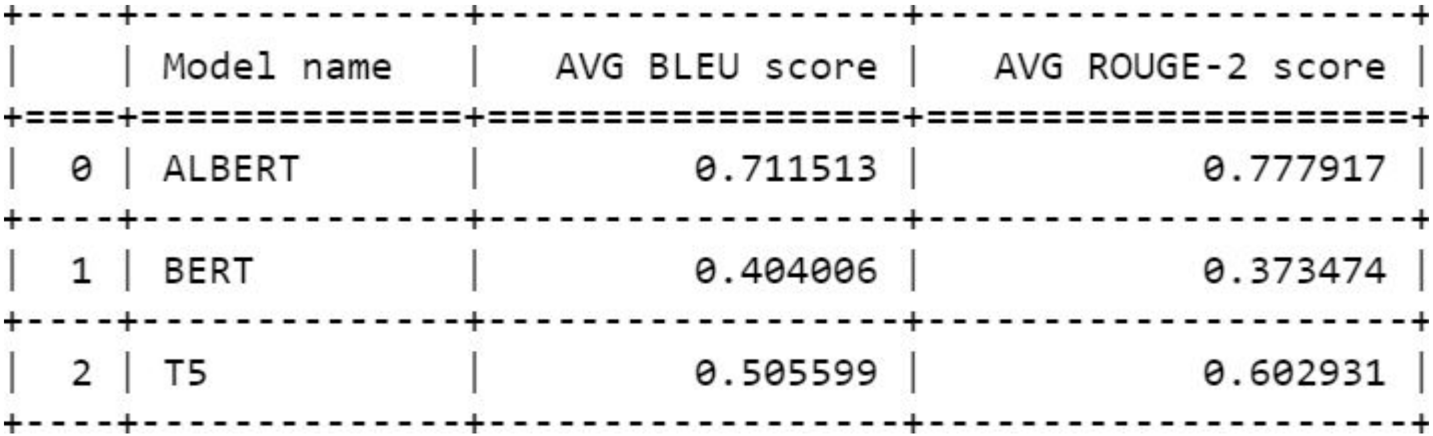
- **Key Idea:** Perform the question answering repeated task on T5, ALBERT and BERT
- **Solution**: For each model , we maintain and update a CSV file that have the question , true answer , predicted answer , and both evaluation scores and then randomly pick 10 results from the file

Example of Output File : ALBERT



	Question	True Answer	Predicted Answer	BLEU Score	ROUGE-2 Score
69	how is a settlement dispute defined in terms ...	A settlement dispute constitutes a distinct en...	A distinct entity that derives its identity fr...	0.764706	0.857143
54	What is the definition of an accounting period	An accounting period is an established range o...	An established range of time during which acco...	0.714286	0.818182
25	Why the end-to-end(e2e) tests are used	To test the full flow of the system, from the ...	To test the full flow of the system, from the ...	1.000000	1.000000
62	What is the role of the matching engine in the...	The matching engine is where the system matche...	The matching engine is where the system matche...	0.955563	1.000000
53	Where will the unaccounted events dashboard be...	Under the name unaccounted events and will be ...	Second position under the accounting main title	0.388889	0.521739
64	What functionalities are offered by the matchi...	The matching rules module allows the user to a...	Add a new rule to the stack of rules, remove a...	0.757576	0.846154
43	What are the components that make up the relat...	The account, the account name, signing, amount...	The account, the account name, signing, amount...	1.000000	1.000000
57	Why might journal entries not be balanced as a...	As a result of accounting rounding processes f...	As a result of accounting rounding processes f...	1.000000	1.000000
76	How are categorized disputes managed within th...	The categorized disputes are uploaded into the...	The categorized disputes are uploaded into the...	1.000000	1.000000
5	Who is the provider of hot files	Billing settlement plan(bsp)	Billing settlement plan(bsp) is the provider ...	0.367879	0.500000

Final Result



```
+----+--------------+------------------+---------------------+
|    | Model name   |   AVG BLEU score |   AVG ROUGE-2 score |
+====+==============+==================+=====================+
|  0 | ALBERT       |         0.711513 |            0.777917 |
+----+--------------+------------------+---------------------+
|  1 | BERT         |         0.404006 |            0.373474 |
+----+--------------+------------------+---------------------+
|  2 | T5           |         0.505599 |            0.602931 |
+----+--------------+------------------+---------------------+
```

4

Conclusion and Future Improvements

Conclusion and Future Improvements

1 Challenging Task

2 Create our Own Languag Model

3 Text to Image

4 Write Research Paper

Thank you for your “Attention”!