Logical Reasoning

Using Large Language Models with Data Augmented from GPT-3

Group Name: NSCube

Md Nayem
Sakthi Ganesh
Shivam Mathur
Saurav Anchlia

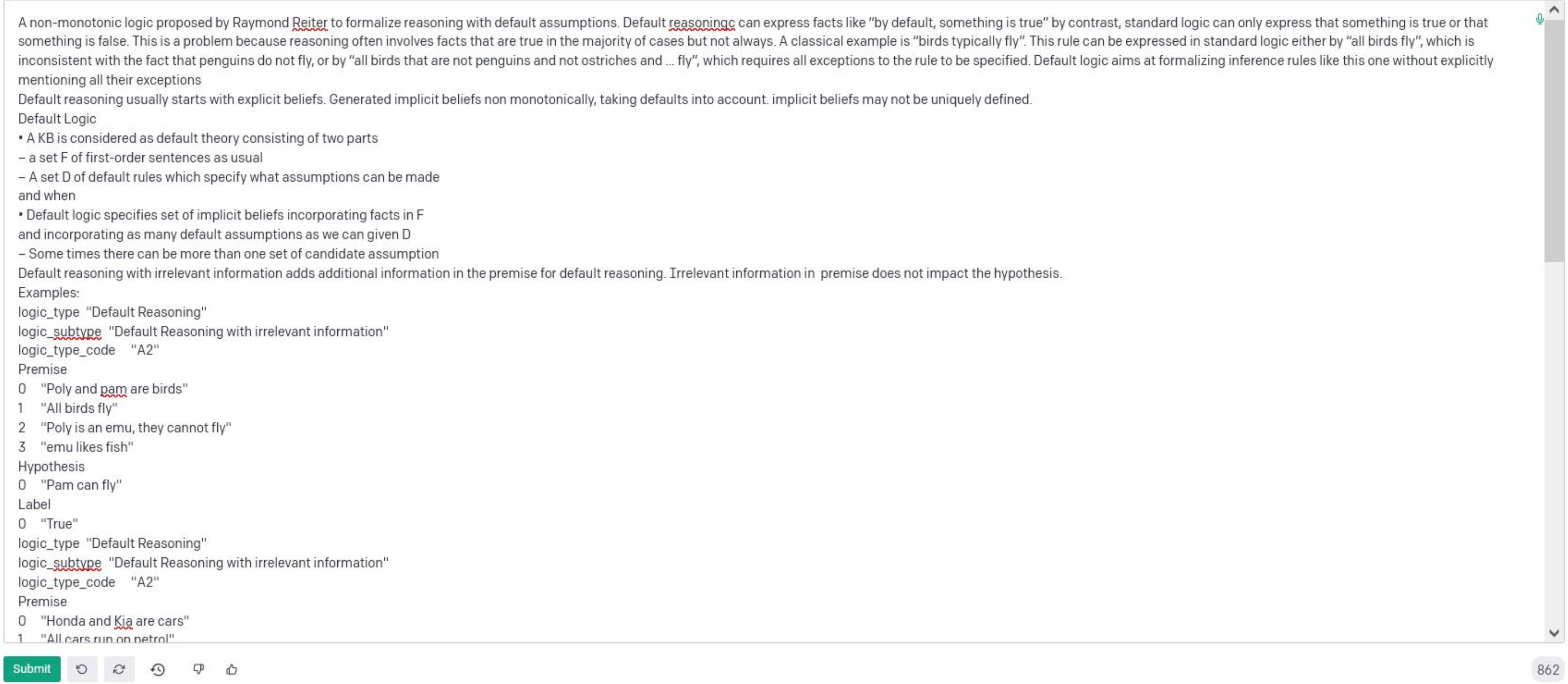
Introduction

- Logical reasoning is a process of making inferences based on premises. In order to reason logically, one must be able to identify the premises and the conclusion of an argument.
- Perform inference using Large Language Model (BERT,RoBERTa)
- Dataset:
  - Test data generated through GPT3
  - Manually generated test set
- 42 logic types spread across 6 categories

Prompt Engineering

- Providing context to GPT3 by adding definition and explanation resulted in better generalization by GPT3 during data generation
- Prompting with topic:example as examples helped introduce diversity in examples.
- Templating certain logic types showed GPT3 was able to annotate it’s example

Example



A non-monotonic logic proposed by Raymond Reiter to formalize reasoning with default assumptions. Default reasoningc can express facts like "by default, something is true" by contrast, standard logic can only express that something is true or that something is false. This is a problem because reasoning often involves facts that are true in the majority of cases but not always. A classical example is "birds typically fly". This rule can be expressed in standard logic either by "all birds fly", which is inconsistent with the fact that penguins do not fly, or by "all birds that are not penguins and not ostriches and ... fly", which requires all exceptions to the rule to be specified. Default logic aims at formalizing inference rules like this one without explicitly mentioning all their exceptions

Default reasoning usually starts with explicit beliefs. Generated implicit beliefs non monotonically, taking defaults into account. implicit beliefs may not be uniquely defined.

Default Logic

• A KB is considered as default theory consisting of two parts

– a set F of first-order sentences as usual

– A set D of default rules which specify what assumptions can be made
and when

• Default logic specifies set of implicit beliefs incorporating facts in F
and incorporating as many default assumptions as we can given D

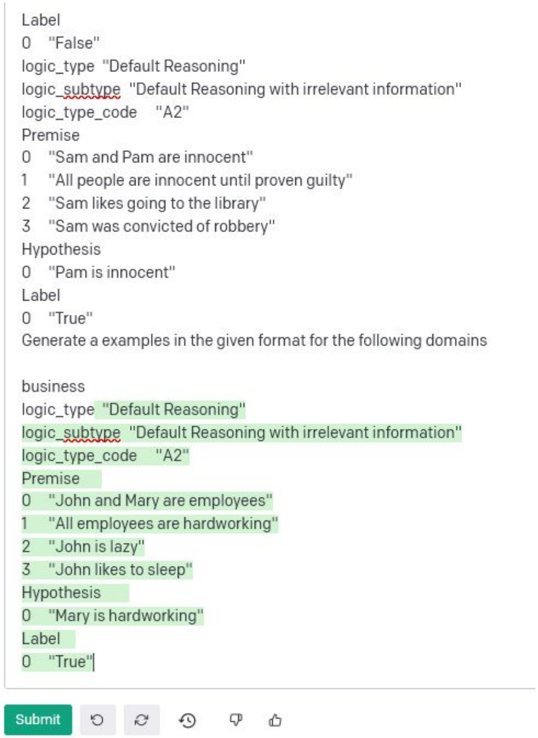
– Some times there can be more than one set of candidate assumption

Default reasoning with irrelevant information adds additional information in the premise for default reasoning. Irrelevant information in premise does not impact the hypothesis.

Examples:

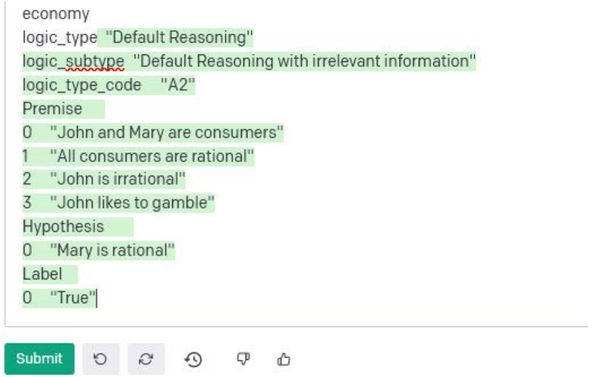
```
logic_type  "Default Reasoning"
logic_subtype  "Default Reasoning with irrelevant information"
logic_type_code    "A2"
Premise
0    "Poly and pam are birds"
1    "All birds fly"
2    "Poly is an emu, they cannot fly"
3    "emu likes fish"
Hypothesis
0    "Pam can fly"
Label
0    "True"
logic_type  "Default Reasoning"
logic_subtype  "Default Reasoning with irrelevant information"
logic_type_code    "A2"
Premise
0    "Honda and Kia are cars"
1    "All cars run on petrol"
```

https://beta.openai.com/playground/p/SWHJCoT9kz9bxvzlp2CTeciS

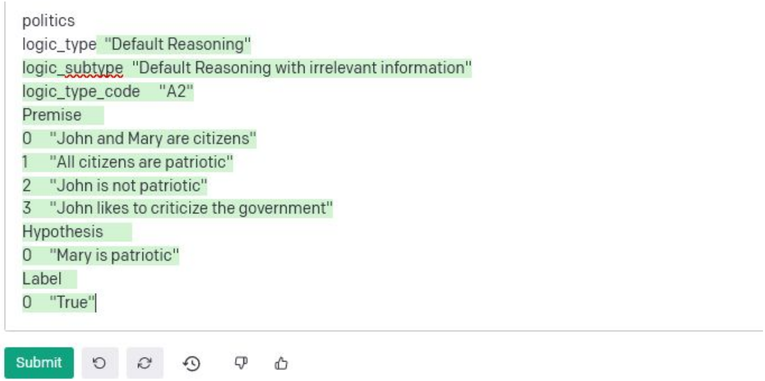


Label
0 "False"
logic_type "Default Reasoning"
logic_subtype "Default Reasoning with irrelevant information"
logic_type_code "A2"
Premise
0 "Sam and Pam are innocent"
1 "All people are innocent until proven guilty"
2 "Sam likes going to the library"
3 "Sam was convicted of robbery"
Hypothesis
0 "Pam is innocent"
Label
0 "True"
Generate a examples in the given format for the following domains

business
logic_type "Default Reasoning"
logic_subtype "Default Reasoning with irrelevant information"
logic_type_code "A2"
Premise
0 "John and Mary are employees"
1 "All employees are hardworking"
2 "John is lazy"
3 "John likes to sleep"
Hypothesis
0 "Mary is hardworking"
Label
0 "True"



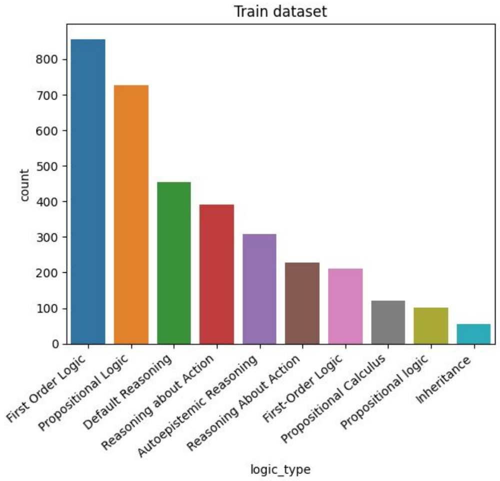
economy
logic_type "Default Reasoning"
logic_subtype "Default Reasoning with irrelevant information"
logic_type_code "A2"
Premise
0 "John and Mary are consumers"
1 "All consumers are rational"
2 "John is irrational"
3 "John likes to gamble"
Hypothesis
0 "Mary is rational"
Label
0 "True"



politics
logic_type "Default Reasoning"
logic_subtype "Default Reasoning with irrelevant information"
logic_type_code "A2"
Premise
0 "John and Mary are citizens"
1 "All citizens are patriotic"
2 "John is not patriotic"
3 "John likes to criticize the government"
Hypothesis
0 "Mary is patriotic"
Label
0 "True"

Data Preprocessing

- Convert JSON to dataframes
- Duplicate rows for examples with multiple hypothesis
- Clean labels for spelling errors
- Created repository for centralized data management.



Methodology - Architecture and Models

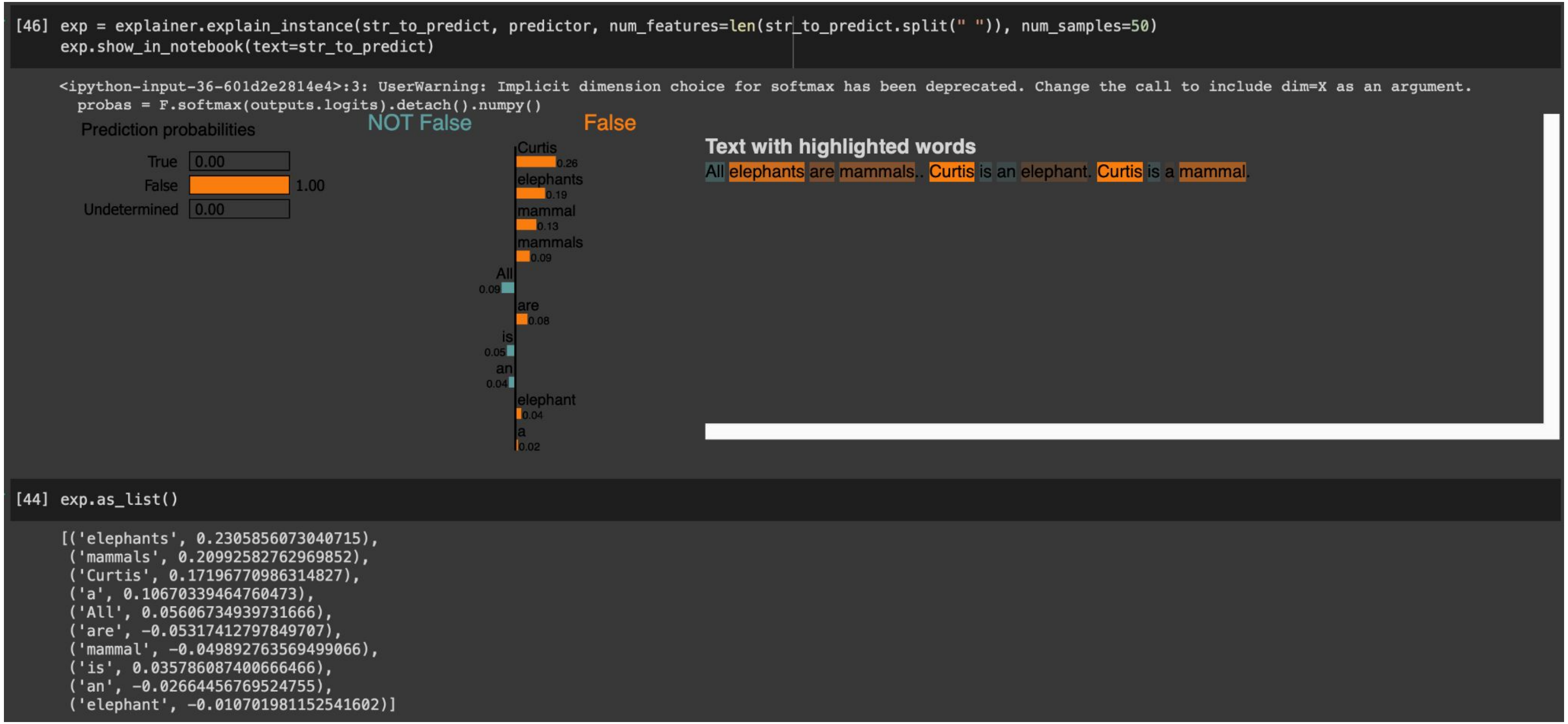
- Baseline
  - BERT-base
- RoBERTa Architectures
  - RoBERTa-base (on Regular Inputs and inputs with Injected Hypothesis)
  - RoBERTa+24 (on Regular Inputs and inputs with Injected Hypothesis)
- RoBERTa Attention Tweaking
  - RoBERTa x2
  - RoBERTa x100
- BERT output layers concatenating
  - Concat last two hidden output layer
  - Concat last four hidden output layer
- RoBERTa - Word Importance with Attention Mask (WIAM)
  - RoBERTa - WIAM

1. Implementation of Word Importance along with Attention Mask

- We hypothesize that feeding the normalized word importance for every sample as the attention mask would improve the baseline model’s performance.
- The approach follows the soft attention mechanism rather than a hard attention of [1, 0] by allowing values between [0 to 1] for each word.
- The word importance is obtained from the best performing model using the LIME library.

Proposed Attention Mask = Attention Mask * Normalized Word Importance

1. Implementation of Word Importance along with Attention Mask



```
[46] exp = explainer.explain_instance(str_to_predict, predictor, num_features=len(str_to_predict.split(" ")), num_samples=50)
     exp.show_in_notebook(text=str_to_predict)
```

```
<ipython-input-36-601d2e2814e4>:3: UserWarning: Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
  probas = F.softmax(outputs.logits).detach().numpy()
```

Text with highlighted words

All elephants are mammals.. Curtis is an elephant. Curtis is a mammal.

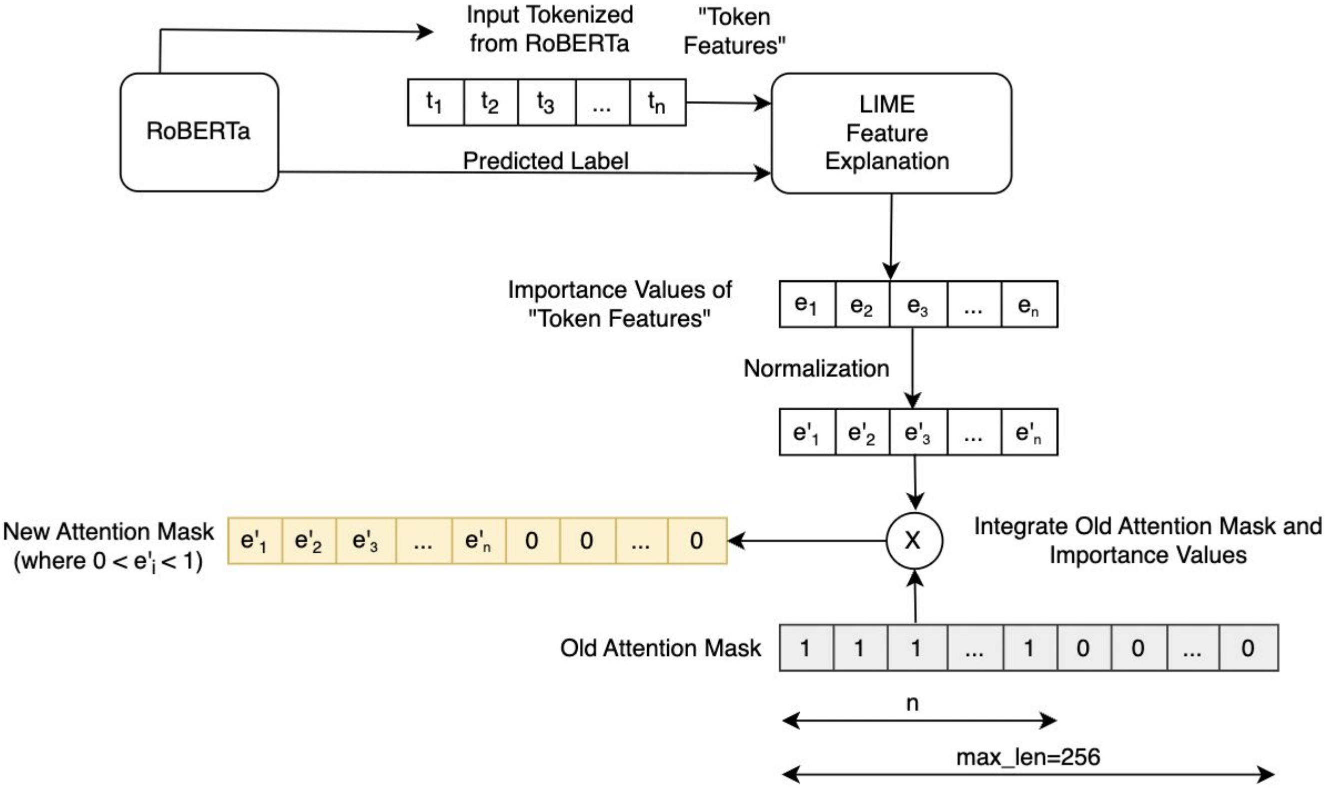
```
[44] exp.as_list()
```

```
[('elephants', 0.2305856073040715),
 ('mammals', 0.20992582762969852),
 ('Curtis', 0.17196770986314827),
 ('a', 0.10670339464760473),
 ('All', 0.05606734939731666),
 ('are', -0.05317412797849707),
 ('mammal', -0.049892763569499066),
 ('is', 0.03578608740066646),
 ('an', -0.02664456769524755),
 ('elephant', -0.010701981152541602)]
```

Figure: The LIME Explainer takes the sentence (premise + hypothesis), well-performing model, and its tokenizer as input and provides the importance of each word in the sentence for making a class prediction as the output, as shown in the diagram above.

1. Implementation of Word Importance along with Attention Mask

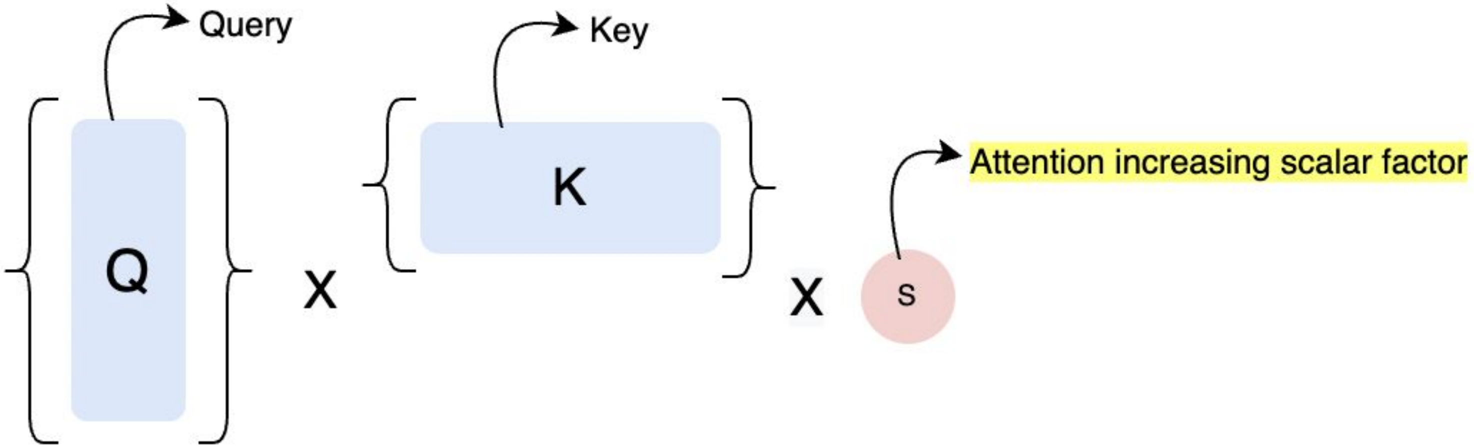
Figure: The word importance is combined with the usual attention mask. The LIME tool outputs importance values for each token in the input text, which we normalize and combine with the usual attention mask. The new attention mask developed is of the same dimension as the old attention mask (max_len=256) but with soft values that are aimed at capturing explainability.



RoBERTa
Input Tokenized from RoBERTa
"Token Features"
t_1 t_2 t_3 ... t_n
Predicted Label
LIME Feature Explanation
Importance Values of "Token Features"
e_1 e_2 e_3 ... e_n
Normalization
e'_1 e'_2 e'_3 ... e'_n
X
Integrate Old Attention Mask and Importance Values
New Attention Mask (where $0 < e'_i < 1$)
e'_1 e'_2 e'_3 ... e'_n 0 0 ... 0
Old Attention Mask
1 1 1 ... 1 0 0 ... 0
n
max_len=256

Image constructed using https://app.diagrams.net/

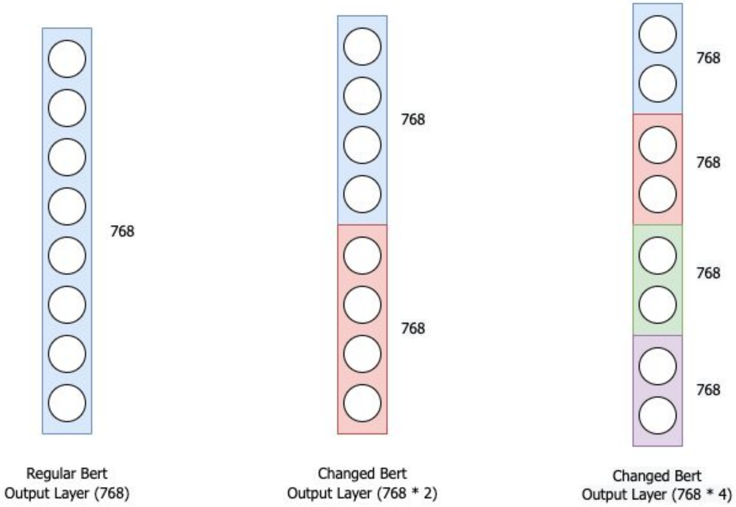
2. Attention Tweaking



- The $Q_iK_i^T$ matrix is basically measuring (or encoding) the relationships between the tokens of the input.
- **"Attention Tweak"**: Enhance the effect such relationship encoding by increasing every element of the $Q_iK_i^T$ matrix by a factor of S.

3. Bert Output Layer Concatenating

- Helps to retain more hidden states.
- These extra hidden states help the classification task.



Result Analysis

Our best model results in an weighted F1 of 0.69, and an accuracy of 71%.

Adding new encoder layers (+24 arch.) does not help significantly.

RoBERTa x2 produces reasonable results comparable to our other architectures. Similar results for BERT (last 2 and 4 layers)

RoBERTa x100: Although the model trains for 3 epochs, the validation loss does not go down. Therefore, we do not see much learning happening here. The model just predicts the majority class.

RoBERTa-WIAM performs much better than RoBERTa-x100 (even though we see during the training process, that the learning pattern is similar to that of RoBERTa-x-100). We are still able to achieve multiclass prediction here, as opposed to just the majority class.

Experiment	Overall (All Labels)		FALSE	TRUE	Undetermined
	Accuracy	F1 (weighted)	F1	F1	F1
Baseline					
BERT-base	0.67	0.68	0.38	0.79	0.27
Regular Inputs					
RoBERTa-base	0.67	0.69	0.51	0.77	0.34
RoBERTa+24	0.68	0.66	0.33	0.8	0
Injected Hypothesis Input					
RoBERTa-base	0.71	0.69	0.32	0.82	0.2
RoBERTa+24	0.65	0.66	0.42	0.77	0
BERT Hidden Layer Concatenating (On Regular Inputs)					
BERT last 2 layers	0.67	0.69	0.4	0.78	0.45
BERT last 4 layers	0.69	0.69	0.38	0.8	0.28
Tweaked Attention (On Regular Inputs)					
RoBERTa-x2	0.65	0.68	0.46	0.76	0.35
RoBERTa-x100	0.74	0.63	0.85	0	0
Word Importance with Attention Mask - WIAM (Proposed and now Implemented)					
RoBERTa-WIAM	0.72	0.63	0.02	0.84	0.28

Error Analysis: Qualitative analysis errors by our best model

Error Type	**Example**	**%**
Negation	Premise: If I do not wake up, then I cannot go to work. If I cannot go to work, then I will not get paid. Hypothesis: If I do not wake up, I will not get paid. **Gold: True Predicted: False**	70
Spatial	Premise: Person A put a laptop, an ipad in his bag while going out.. Today he forgot to put either the laptop or the ipad in the bag. Hypothesis: Both the laptop and the ipad is missing in the bag today. **Gold: False Predicted: True**	10
Temporal	Premise: I will play cricket on weekends if there is no rain.. I played cricket last weekend. Hypothesis: There was rain last weekend. **Gold: False Predicted: True**	10
Named Entities	Premise: Apples are grown at Washington state in the United States. Hypothesis: Therefore, somewhere in the United States Apples are grown. **Gold: True Predicted: Undetermined**	8
Numeric Data	Premise: If I have 1 million dollars, I will donate half of it.. If I have 10 million dollars, I will donate 75% of it.. Either I have 1 million dollars or 10 million dollars. Hypothesis: Therefore I will donate 500 thousand dollars or 7.5 million dollars. **Gold: True Predicted: Undetermined**	2

We consider 50 randomly selected test instances and qualitatively assign error classes as per human intuition. A particular instance may have more than one error class, however we assign only one error class, the one we feel is most contributing. Most errors are due to negations.

Conclusions

- We show that GPT3 (when prompted appropriately) is able to generate decent quality of data instances for our logical reasoning tasks.
- Encoder based LLMs like BERT and RoBERTa can have the attention layers tweaked, and there performance can be analysed.
- We also run experiments by adding new layers of encoders, formatting inputs differently and utilizing different layers of the BERT model and analyse performance.
- We have also implemented a word importance-based attention mask model that relies on a Soft Attention Mask mechanism facilitated by the LIME explainability tool.
- Lastly, we run a error analysis, to understand where the LLM fails to predict correct labels.