Fine Tuning a Llama2 Model for Lie Detection

Caltran Lorenzo

Cimbro Letizia

Jahanianarange Nahid

Skurativska Kateryna

The Task

What is Lie Detection?

- Lie detection is an assessment of a statement with the goal to reveal whether the person is being truthful or deceptive in their statements.
- When people tell lies, they use specific verbal strategies to make others believe their false stories.
- Automatized verbal lie detection techniques through the application of Machine Learning and LLMs.

The Model

The Model

- Discover whether stories told by people have actually happened or have been imagined (deception)
- Llama2
- Meta’s open source LLM
- Up to 70B parameters

The Dataset

The dataset

- The Memory dataset contains 6854 stories about autobiographical experiences collected through crowdsourcing.
- Each story has 23 parameters.
- The inclusion of variables such as "distracted," "draining," "frequency," and "stressful" suggests an interest in understanding the emotional and cognitive aspects of the storytelling process, which may be relevant for lie detection research.
- The variable "memType" likely categorizes the stories into these three types—recalled, retold, and imagined

Categories of Stories:

- ***Recalled:*** Stories that are based on actual, personal experiences or events that an individual has lived through or witnessed

- ***Imagined:*** Stories that are fictional and created in the storyteller's imagination. They do not have a basis in r eal events or experiences

- ***Retold:*** Stories that involve recounting or narrating events that the storyteller has heard from others

Data Preprocessing

Preprocessing

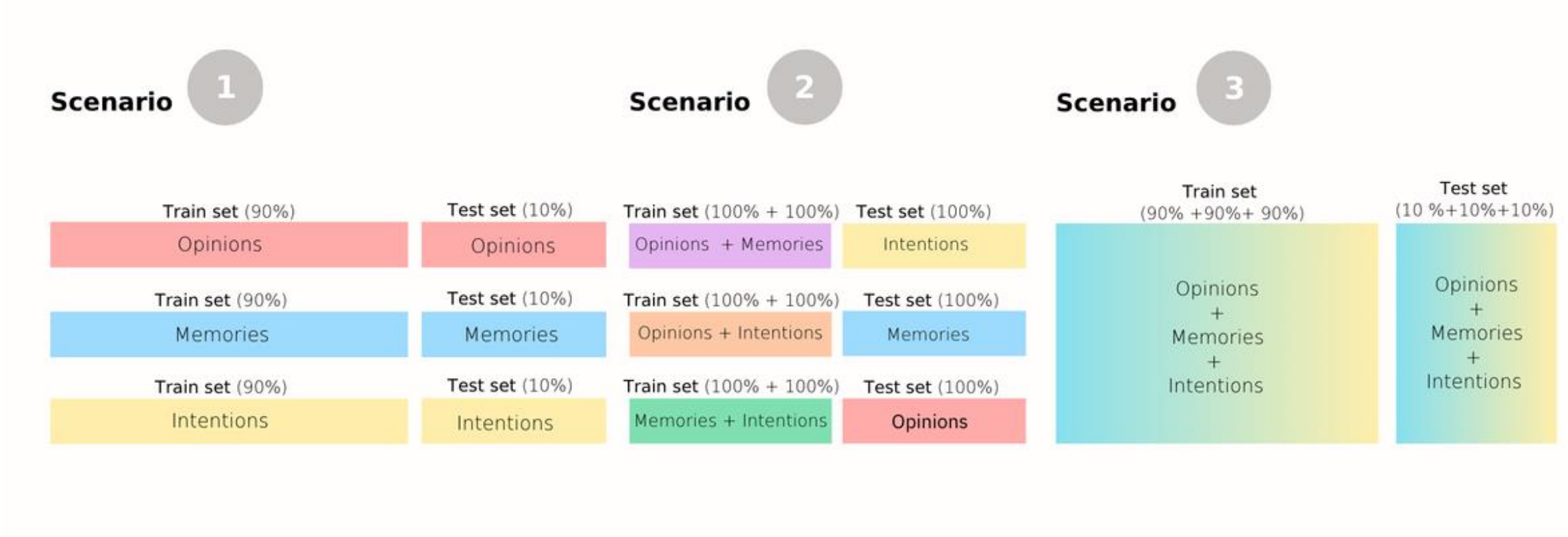
Before using Memory dataset for fine-tunning the LLama2 model we do the following:

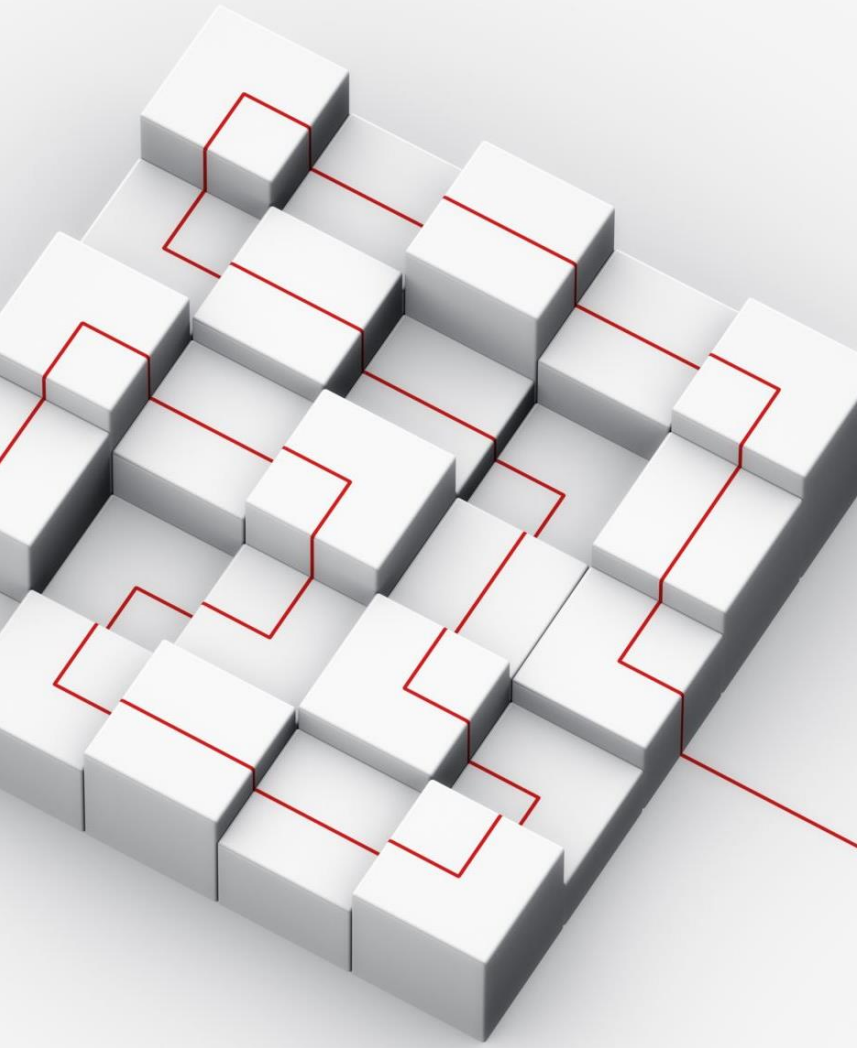
1. Exclude retold stories from consideration, as they are not fabrications; rather, they involve real occurrences, albeit not directly experienced by the storyteller.
2. Remove stories for which we don't have story or/and we don't know the type of the story.
3. After counting exact number of words for each story (using Spacy's English language model) remove stories with number of words below 2.5 of standard deviation.

As a result of preprocessing, we obtained 5507 stories for further usage.

Scenarios

Scenarios


Scenario 1
Train set (90%)
Opinions
Test set (10%)
Opinions
Train set (90%)
Memories
Test set (10%)
Memories
Train set (90%)
Intentions
Test set (10%)
Intentions
Scenario 2
Train set (100% + 100%)
Opinions + Memories
Test set (100%)
Intentions
Train set (100% + 100%)
Opinions + Intentions
Test set (100%)
Memories
Train set (100% + 100%)
Memories + Intentions
Test set (100%)
Opinions
Scenario 3
Train set
(90% +90%+ 90%)
Opinions
+
Memories
+
Intentions
Test set
(10 %+10%+10%)
Opinions
+
Memories
+
Intentions



Scenarios

- **Scenario 1:** The model is trained to recognize lies in one part of a dataset and tested if it could spot lies in the remaining part. This should be done for each dataset, using a fresh copy of the model each time. This is done to see how well the model can learn to detect lies in the same situation.
- **Scenario 2:** The model is trained on multiple datasets and then tested it on another one that it has not seen before. This should be repeated multiple times, using different datasets combinations of datasets each time. This helps to understand how well the model can handle new situations it hasn't encountered during training.
- **Scenario 3:** The model is trained and tested on a combination of all the datasets. This allows to figure out if the model can learn and find lies in stories from different situations.

Fine Tuning

Fine tuning

Llama2 has multiple purposes, such as:

- Summarizing, expanding, rewriting and changing the tone of voice of an input text;
- After entering a text prompt, Llama2 attempts to predict the most plausible follow-on text;

What we are going to do is apply fine-tuning. In machine learning, fine tuning is the process of adjusting the weights and parameters of a pre-trained model on new data to improve its performance on a specific task.

Methodology

Methodology

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

As we can see, there a multiple versions of the Llama 2 models. Due to limitations with the GPU we fine tuned the Llama2 Chat model with 7B parameters.

Generating a prompt

The following step is to transform the texts contained in the train and test data into prompts to be used by Llama. For both the training and test set the prompt contains specific instructions about what our goal is: to categorize each story as either 'imagined' or 'recalled', aside from the that, the train set prompt also contains the story that we want to categorize and the assigned label, while the test prompt only contains the story.

The training phase

The last thing to do is to configure and initialize a Simple Fine-tuning Trainer (SFTTrainer) for training a large language model using the Parameter-Efficient Fine-Tuning (PEFT) method, which should save time as it operates on a reduced number of parameters compared to the model's overall size.

The PEFT method focuses on refining a limited set of additional model parameters, while keeping the majority of the pre-trained LLM parameters fixed.

Results

Results before Fine Tuning

As we can see the original model is not able to carry out the task with satisfying results, as it only categorizes the stories as ‘imagined’ or it’s not able to categorize them at all.

```
Accuracy: 0.499
Accuracy for label 0: 0.989
Accuracy for label 1: 0.014

Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.99      0.66       274
           1       0.57      0.01      0.03       277

    accuracy                           0.50       551
   macro avg       0.53      0.50      0.35       551
weighted avg       0.53      0.50      0.34       551


Confusion Matrix:
[[271   3]
 [273   4]]
```

Results after Fine Tuning

After 3 training epochs we can observe a significant increase in the accuracy, even though we observed high values for the training and evaluation loss

Epoch	Training Loss	Validation Loss
0	1.907800	1.836034
1	1.884700	1.821752
2	1.870500	1.817982

```
Accuracy: 0.708
Accuracy for label 0: 0.953
Accuracy for label 1: 0.466

Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.95      0.76       274
           1       0.91      0.47      0.62       277

    accuracy                           0.71       551
   macro avg       0.77      0.71      0.69       551
weighted avg       0.77      0.71      0.69       551


Confusion Matrix:
[[261  13]
 [148 129]]
```

Observations

What we observed is that while the model is able to categorize the stories with good accuracy, it is not able to generate correctly the label text.

	text	y_true	y_pred	predicted_label
1409	Analyze the story between brackets and determi...	imagined	imag	imagined
3049	Analyze the story between brackets and determi...	imagined	imag	imagined
797	Analyze the story between brackets and determi...	imagined	imag	imagined
3512	Analyze the story between brackets and determi...	imagined	imag	imagined
987	Analyze the story between brackets and determi...	imagined	re	recalled
...	...	...	...	...
88	Analyze the story between brackets and determi...	recalled	re	recalled
180	Analyze the story between brackets and determi...	recalled	imag	imagined
2805	Analyze the story between brackets and determi...	recalled	imag	imagined
4335	Analyze the story between brackets and determi...	recalled	imag	imagined
3700	Analyze the story between brackets and determi...	recalled	imag	imagined

Conclusions

In conclusion, the study highlights the potential of advanced language models to uncover and classify cognitive processes embedded in textual narratives.

The findings contribute to our understanding of human cognition and memory, while also offering practical applications in various domains.

By leveraging cutting-edge technologies and interdisciplinary approaches, researchers can continue to advance our knowledge of storytelling, memory recall, and cognitive functioning in both theoretical and practical contexts.

Thanks for the attention