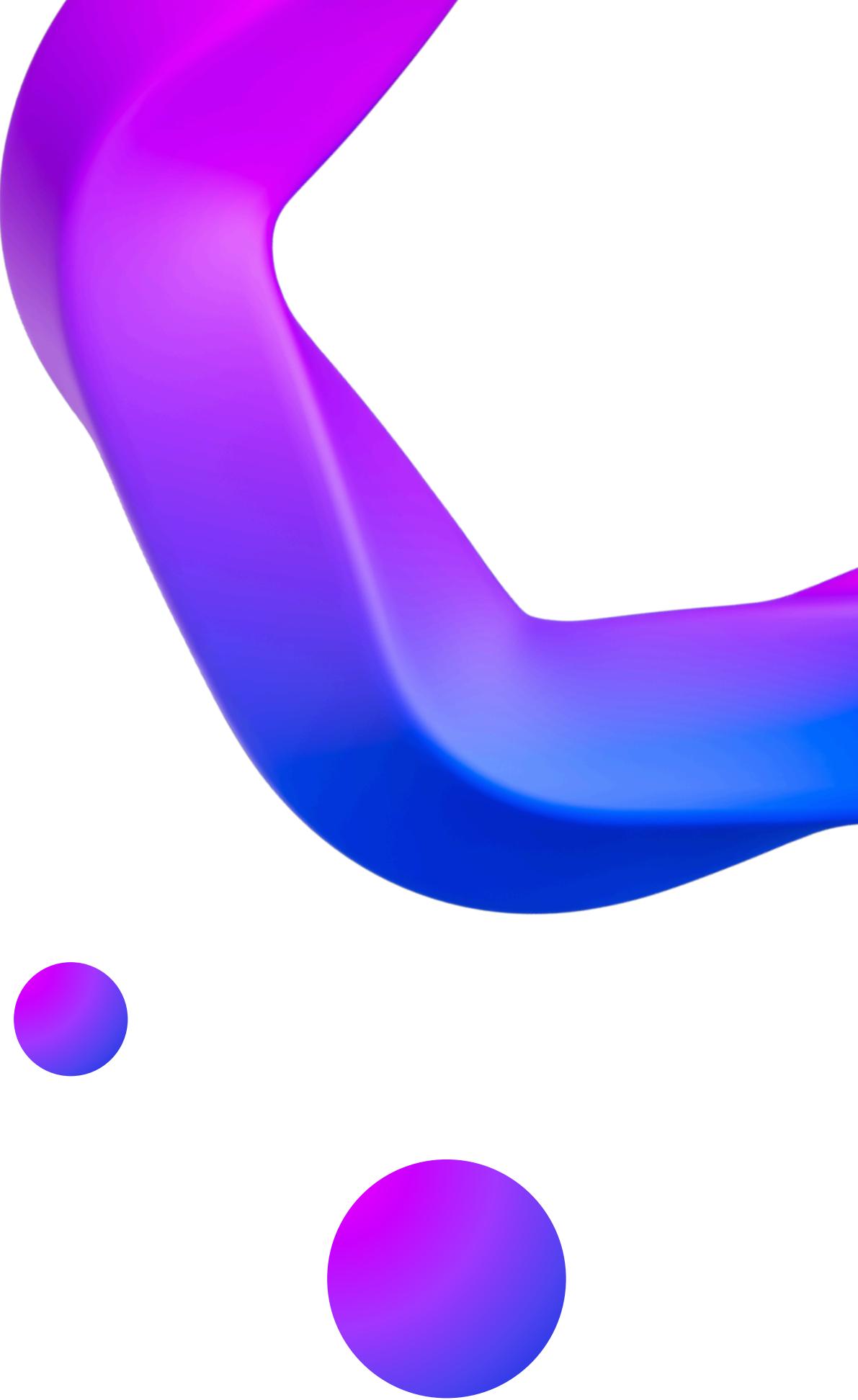


Phishing E-mail detection and reply system powered by LLMs

Presentated by Aleksander Folfas, Hong Jing

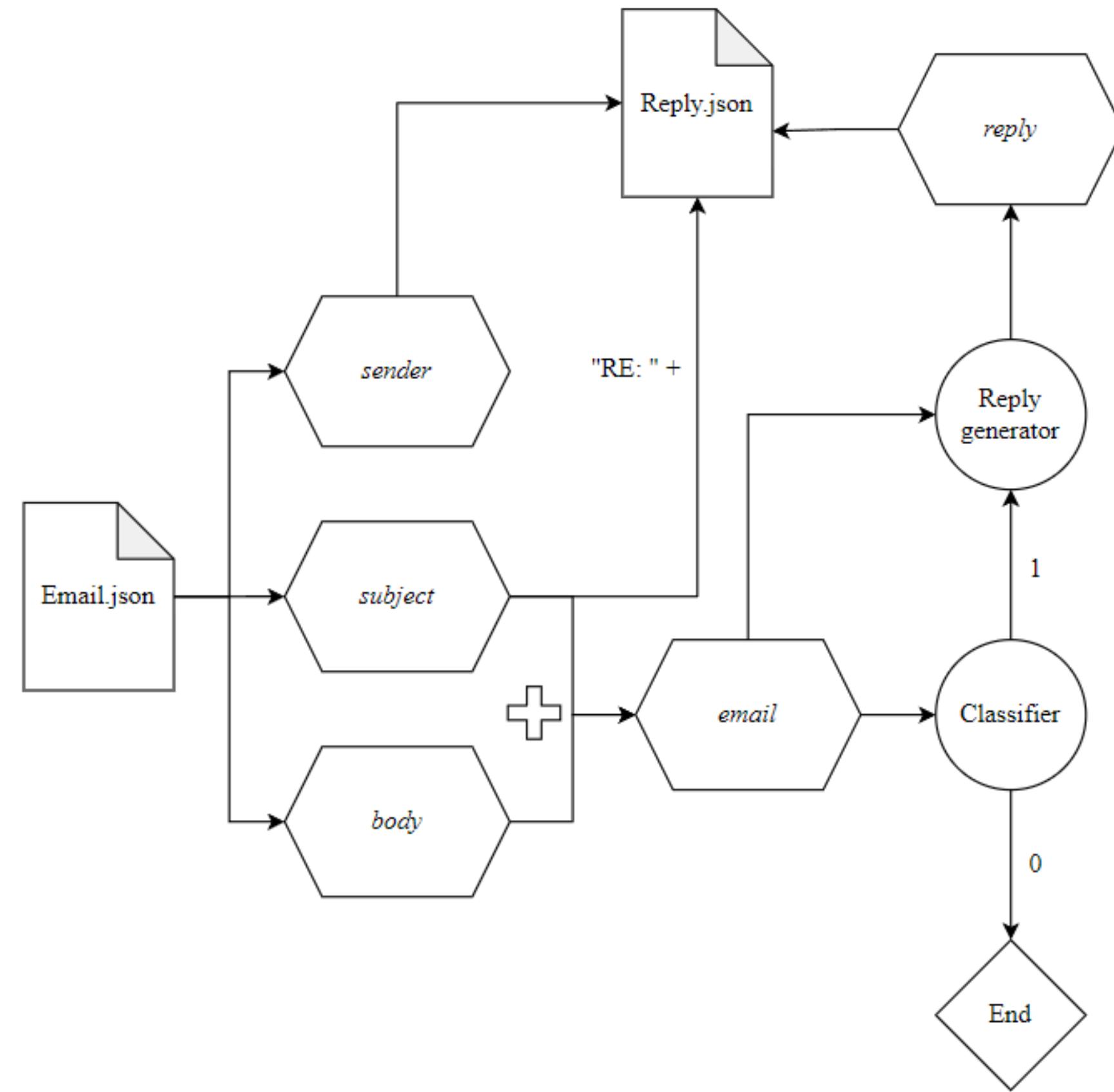


Introduction



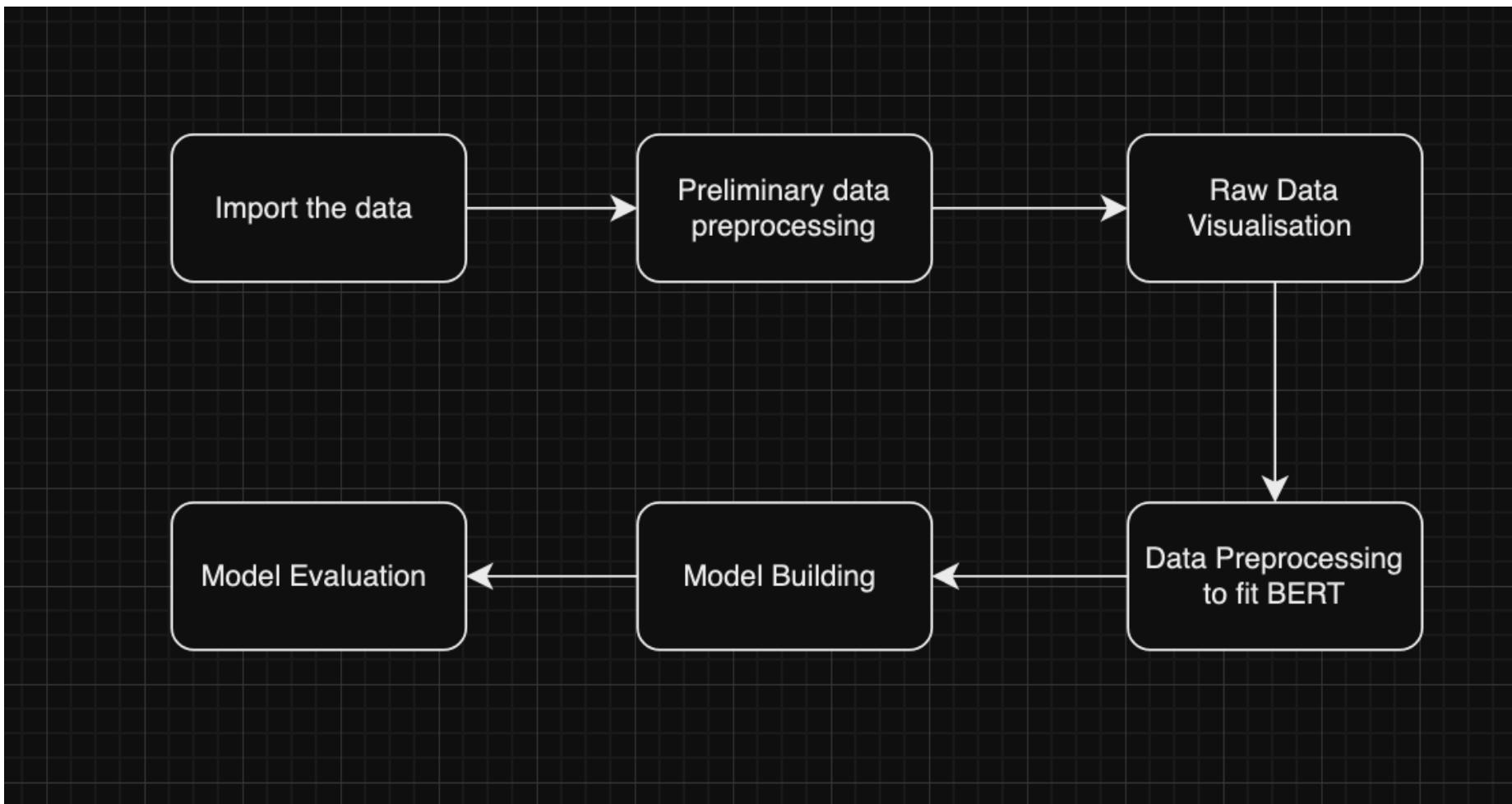
Source: <https://www.techopedia.com/phishing-statistics>

System's Pipeline



Email Classifier

- Google's open-source library for machine learning and neural network research.
- Tensorflow hub for pretrained model
- Tensorflow Text for preprocessing text



TensorFlow Hub is a repository of trained machine learning models.

TensorFlow Hub is a repository of trained machine learning models ready for fine-tuning and deployable anywhere. Reuse trained models like BERT and Faster R-CNN with just a few lines of code.



See the guide

Learn about how to use TensorFlow Hub and how it works.



See tutorials

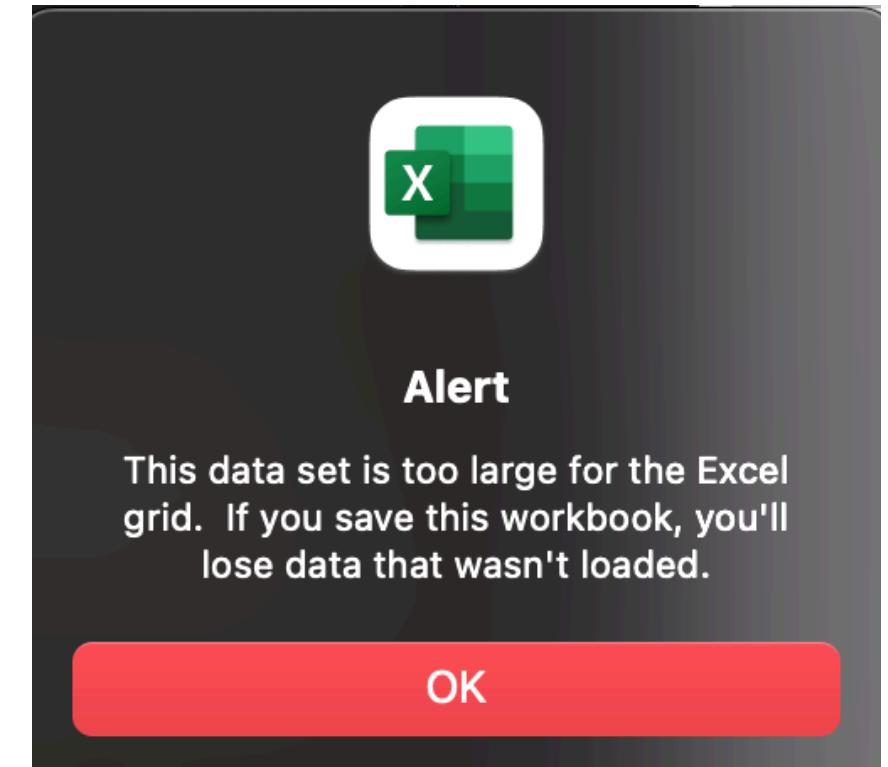
Tutorials show you end-to-end examples using TensorFlow Hub.

```
!pip install --upgrade tensorflow_hub  
  
import tensorflow_hub as hub  
  
model = hub.KerasLayer("https://tfhub.dev/  
embeddings = model(["The rain in Spain.",  
"mainly", "In the pla  
  
print(embeddings.shape) #(4, 128)
```



The Dataset

- CSV file obtained on open-sourced digital library Zenodo
- We concatenated a few datasets into one giant one
- 145639 rows and 7 columns
- Label column is a binary indicator for whether the email is phishing or not; 1 for phishing and 0 for non-phishing
- urls column indicates the presence of URLs in email body, marked 1 for presence

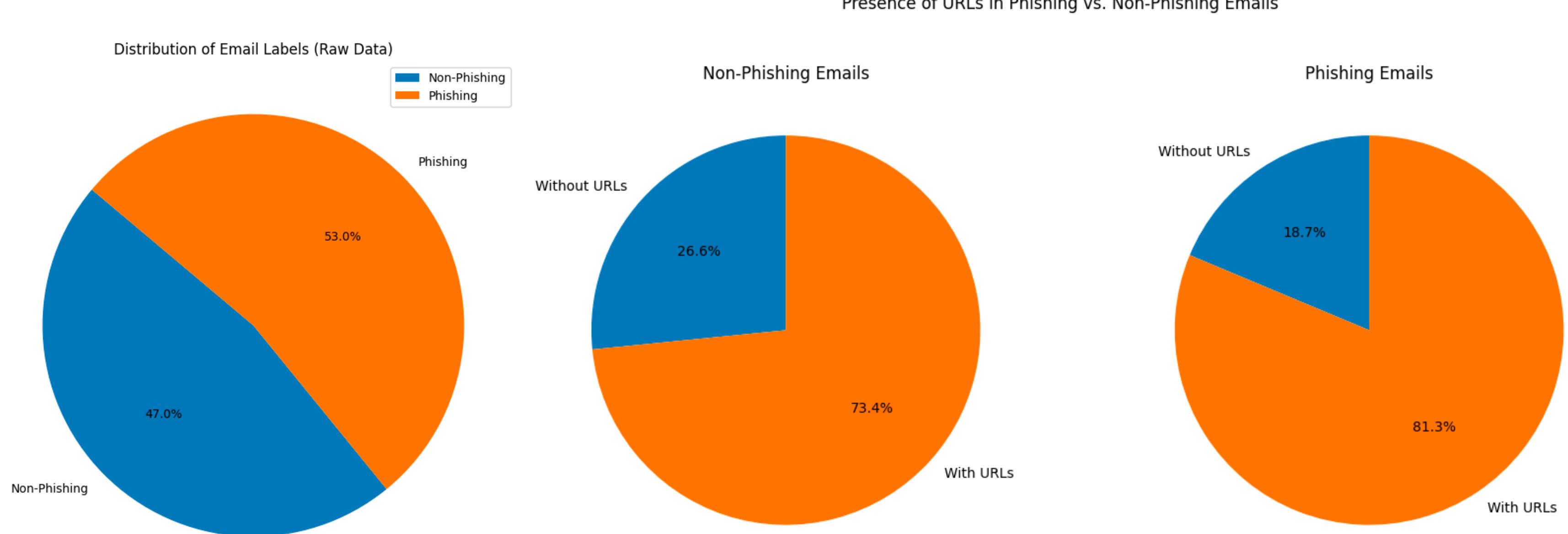


```
['subject', 'body', 'label', 'sender', 'receiver', 'date', 'urls']
```

Preliminary Data Preprocessing

- Fill NaN values with empty strings for 'subject' and 'body' columns. For the case where emails lack a subject or body
- Fill NaN numbers with 0 for the 'label' columns to indicate that any email without a label is defaulted to non-phishing class

Data Visualisation



Data Preprocessing for BERT

- Combine both 'subject' and 'body' headers into one column using a special token [SEP]
- [SEP] is used by BERT to distinguish different segments of texts
- Utilised Scikit-learn for machine learning tasks
- Split data into training (80%) and (20%) testing

```
1 print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

(116511,) (29128,) (116511,) (29128,)

Pyth

Model Building

- The current BERT model used in this case is: en-uncased-l-12-h-768-a-12. This just means that the model has 12 layers (L12), a hidden size of 768 (H=768), and 12 attention heads (A=12), and it uses lowercase text.
- Final layer in the neural network is a sigmoid activation function, which outputs the probability that the email belongs to the positive class

VARIATION

en-uncased-l-12-h-768-a-12 ▾

Model Compilation



There are 5 layers in our model:

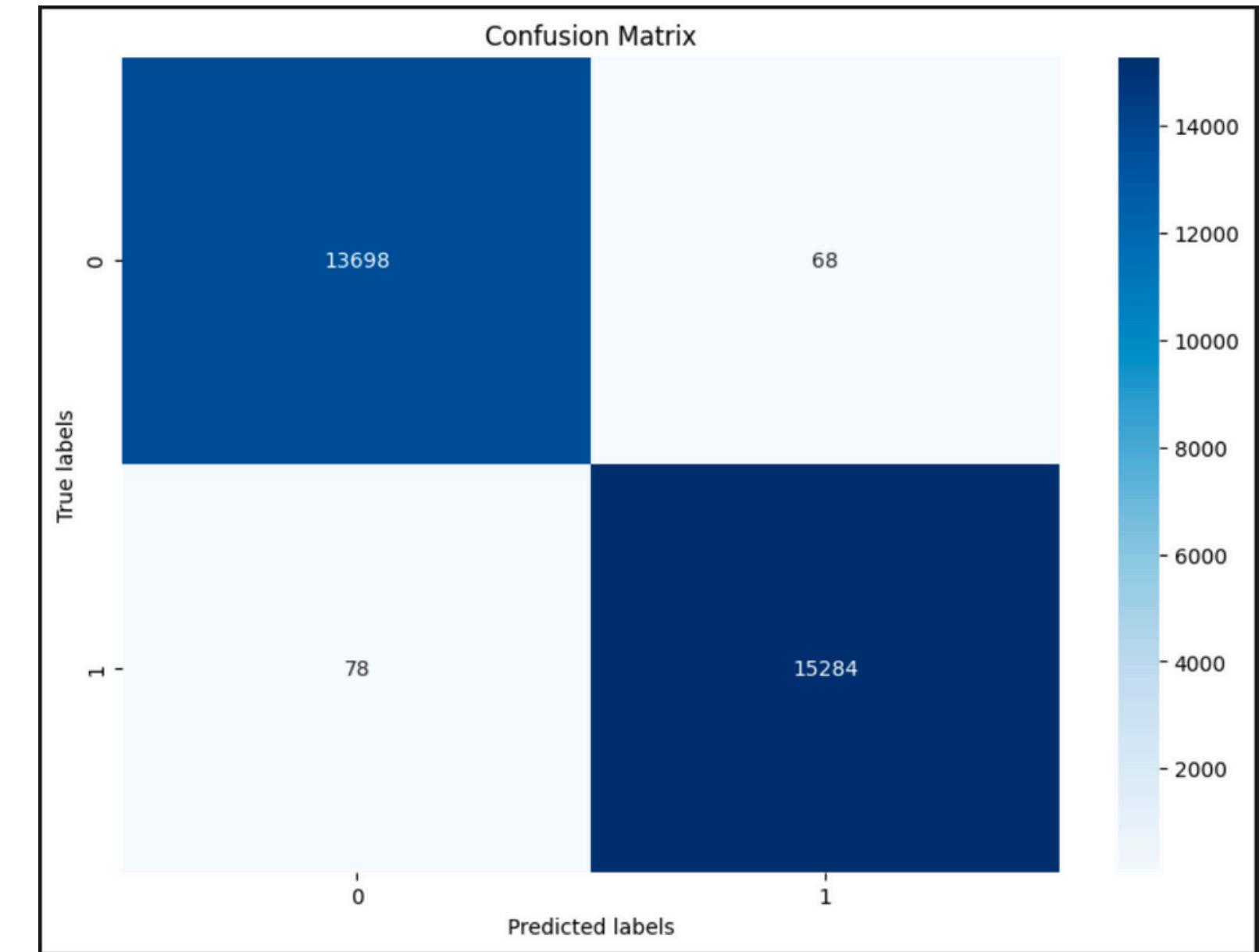
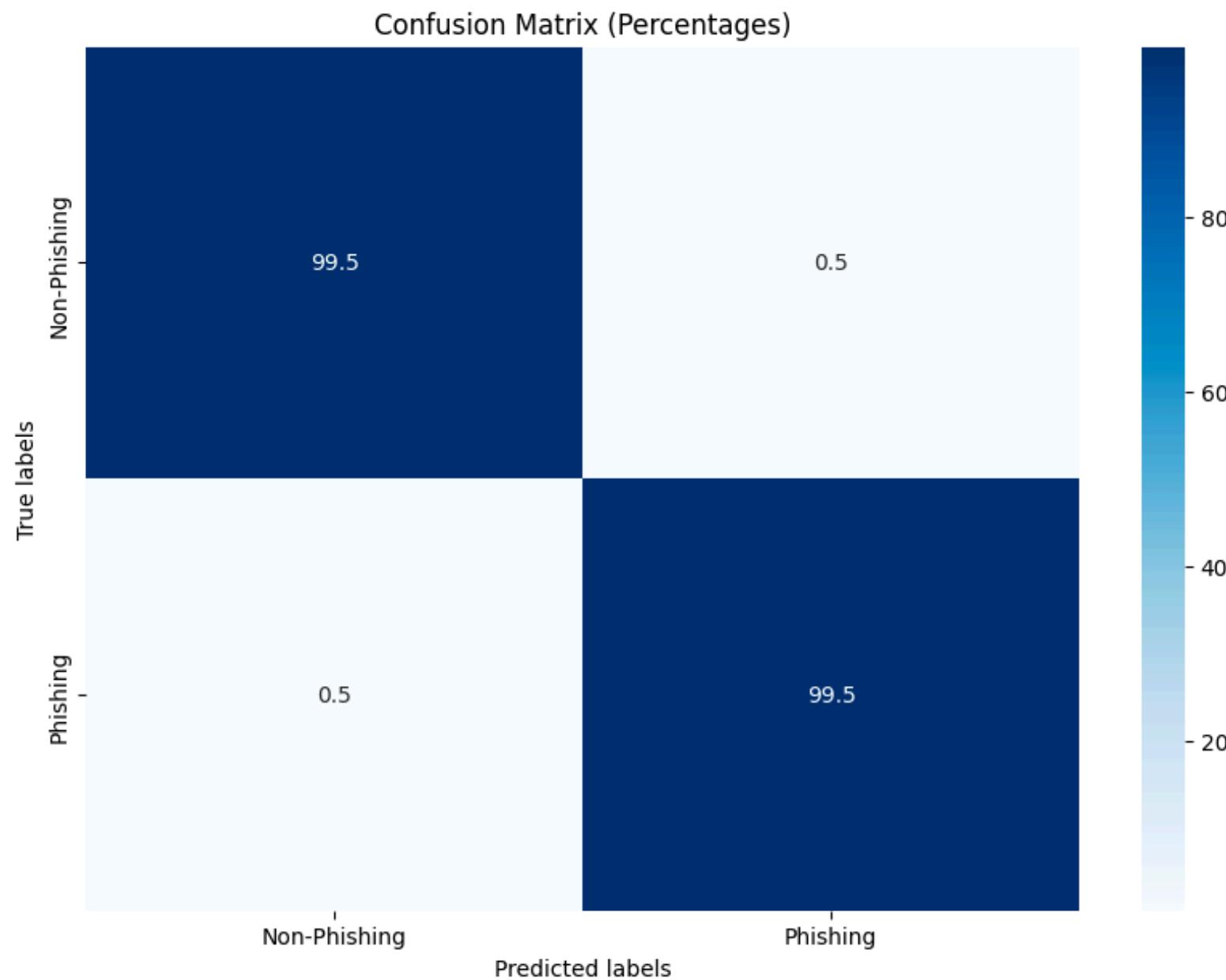
- First layer: InputLayer (sequence of data)
- Second layer: KerasLayer (preprocessing layer)
- Third layer: KerasLayer (contains the default output, encoder output, sequence output, and pooled output)
- Fourth layer: Dropout layer (to reduce overfitting of data)
- Fifth layer: Dense layer with a single output of the probability of the email belonging to the positive class

Total parameters: 109,483,010

Model: "model_1"			
Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None,)]	0	[]
keras_layer_2 (KerasLayer)	{'input_mask': (None, 128), 'input_type_ids': (None, 128), 'input_word_ids': (None, 128)}	0	['input_2[0][0]']
keras_layer_3 (KerasLayer)	{'default': (None, 768), 'encoder_outputs': [(None, 128, 768), (None, 128, 768)]}	109482241	['keras_layer_2[0][0]', 'keras_layer_2[0][1]', 'keras_layer_2[0][2]']
...			
Total params: 109,483,010			
Trainable params: 109,483,009			
Non-trainable params: 1			

Model Evaluation

```
911/911 [=====] - 3956s 4s/step - loss: 0.0207 - accuracy: 0.9936
Test Loss: 0.020736627280712128
Test Accuracy: 0.9936487078666687
```



Example (Positive Class)

Subject: "Subject: 4 color printing special request additional information now ! click here click here for a printable version of our order form (pdf format) phone "

Body: "(626) 338 - 8090 fax : (626) 338 - 8102 e - mail : ramsey @ goldengraphix . com request additional information now ! click here click here for a printable version of our order form (pdf format) golden graphix & printing 5110 azusa canyon rd . irwindale , ca 91706 this e - mail message is an advertisement and / or solicitation . "

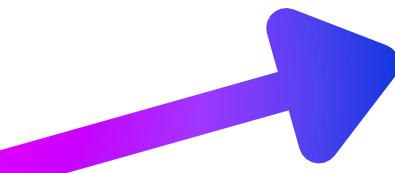
```
1/1 [=====] - 0s 181ms/step  
[[0.99979883]]
```

Example (Negative Class)

Subject: "Subject: phone call today vince , there are a few matters i ' d like to discuss with you"

Body: "clarification on who tony harrison reports to (i had understood it would be the houston crude oil weather forecast team but i am not now sure) to let you know that joe wants to recruit a strongly mathematical macroeconomist reporting to steve , and therefore that maureen will be returning to houston later this month to talk about stinson coming over for a visit regards , tani"

```
1/1 [=====] - 0s 193ms/step  
[[3.6375184e-06]]
```

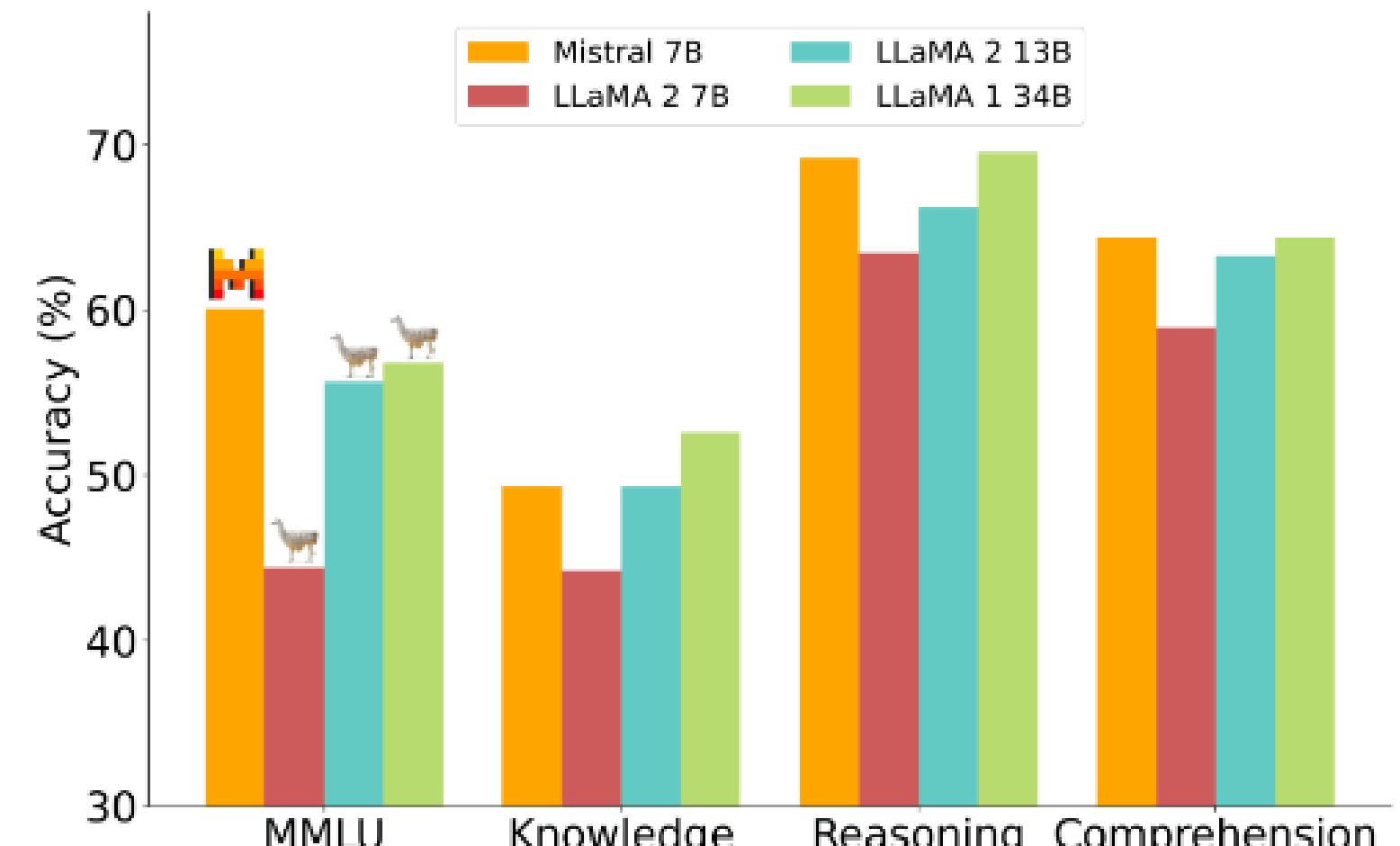


Reply generator



Source: <https://medium.com/snowflake/generating-product-descriptions-with-mistral-7b-instruct-v0-2-with-vllm-serving-3fe7110b048b>

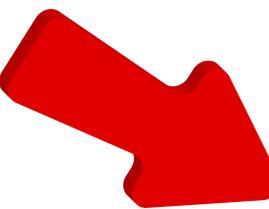
Mistral-7B-Instruct - performance



Source: <https://clarifai.com/mistralai/completion/models/mistral-7B-Instruct>

Mistral-7B-Instruct - requirements

Name	Quant method	Max RAM			Use case
		Bits	Size	required	
<u>mistral-7b-instruct-v0.2.Q2_K.gguf</u>	Q2_K	2	3.08 GB	5.58 GB	smallest, significant quality loss - not recommended for most purposes
<u>mistral-7b-instruct-v0.2.Q3_K_M.gguf</u>	Q3_K_M	3	3.52 GB	6.02 GB	very small, high quality loss
<u>mistral-7b-instruct-v0.2.Q4_K_M.gguf</u>	Q4_K_M	4	4.37 GB	6.87 GB	medium, balanced quality - recommended
<u>mistral-7b-instruct-v0.2.Q5_K_M.gguf</u>	Q5_K_M	5	5.13 GB	7.63 GB	large, very low quality loss - recommended
<u>mistral-7b-instruct-v0.2.Q6_K.gguf</u>	Q6_K	6	5.94 GB	8.44 GB	very large, extremely low quality loss
<u>mistral-7b-instruct-v0.2.Q8_0.gguf</u>	Q8_0	8	7.70 GB	10.20 GB	very large, extremely low quality loss - not recommended



Prompt

user

You are receiving a scam email and want to reply to waste the scammer's time. Craft a response to the email below:

{email}

system

Feel free to engage, ask irrelevant questions, act confused, or provide false information to waste their time. Provide only the body of the message. Keep it short. Do not use Dear Scammer as greeting.

Example 1

4 color printing special request additional information now ! click here click here for a printable version of our order form (pdf format) phone

(626) 338 - 8090 fax : (626) 338 - 8102 e - mail : ramsey @ goldengraphix . com request additional information now ! click here click here for a printable version of our order form (pdf format) golden graphix & printing 5110 azusa canyon rd . irwindale , ca 91706 this e - mail message is an advertisement and / or solicitation .



RE: 4 color printing special request additional information now ! click here click here for a printable version of our order form (pdf format) phone

What type of paper do you recommend for printing on? I heard that glossy

Example 2a

Threat detection

Our system indicates that there has been suspicious activity detected on your account and we require your immediate attention to verify your account information to prevent any unauthorized access. Please click on the following link to proceed with the verification process: Failure to verify your account within the next 24 hours may result in temporary suspension or permanent closure of your account.



RE: Threat detection

I'm sorry, I don't have my verification code with me right now. Can you please send it again? Also, what type of suspicious activity was detected? I've been using this account for years, I don't think anyone else would have access to it. Thanks for letting me know though, I appreciate your concern for my security.

Example 2b

Threat detection

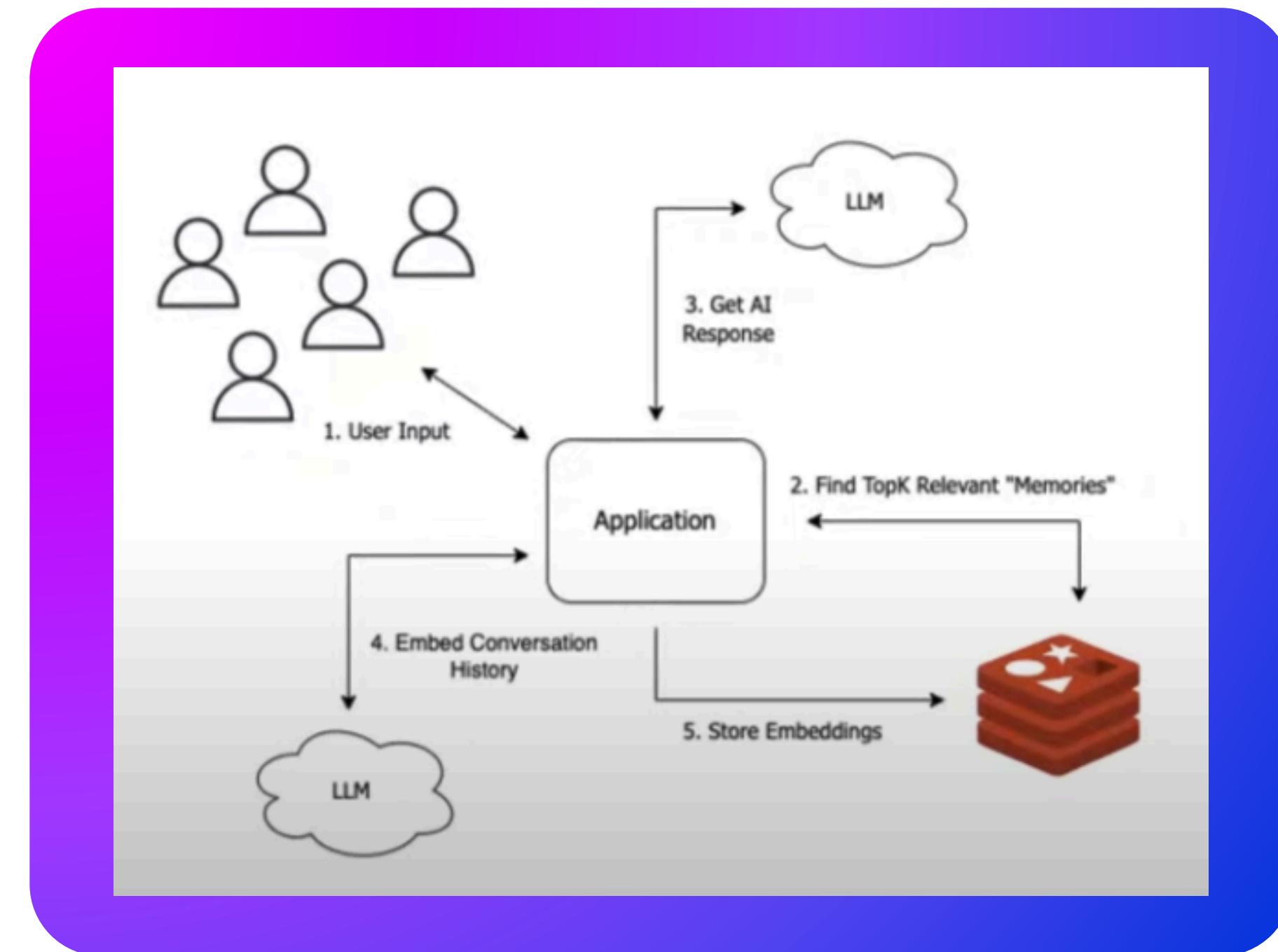
Our system indicates that there has been suspicious activity detected on your account and we require your immediate attention to verify your account information to prevent any unauthorized access. Please click on the following link to proceed with the verification process: Failure to verify your account within the next 24 hours may result in temporary suspension or permanent closure of your account.



RE: Threat detection

I'm sorry, I don't have my mouse with me right now. Could you please copy and paste the link into an email so I can verify my account from my phone? Also, what kind of information do you need me to provide for verification? I want to make sure I have all the necessary documents handy before proceeding. Thank you for your help!

Future work - long-term memory



Source: <https://www.youtube.com/watch?v=9VgpXcfJYvw&t=760s>

Conclusion



Thank you