Conversational AI with Transformer models

Rajesh Shreedhar Bhat

Sr. Data Scientist, Walmart Global Tech - India

Dinesh Ladi

Data Scientist, Walmart Global Tech - India

Agenda

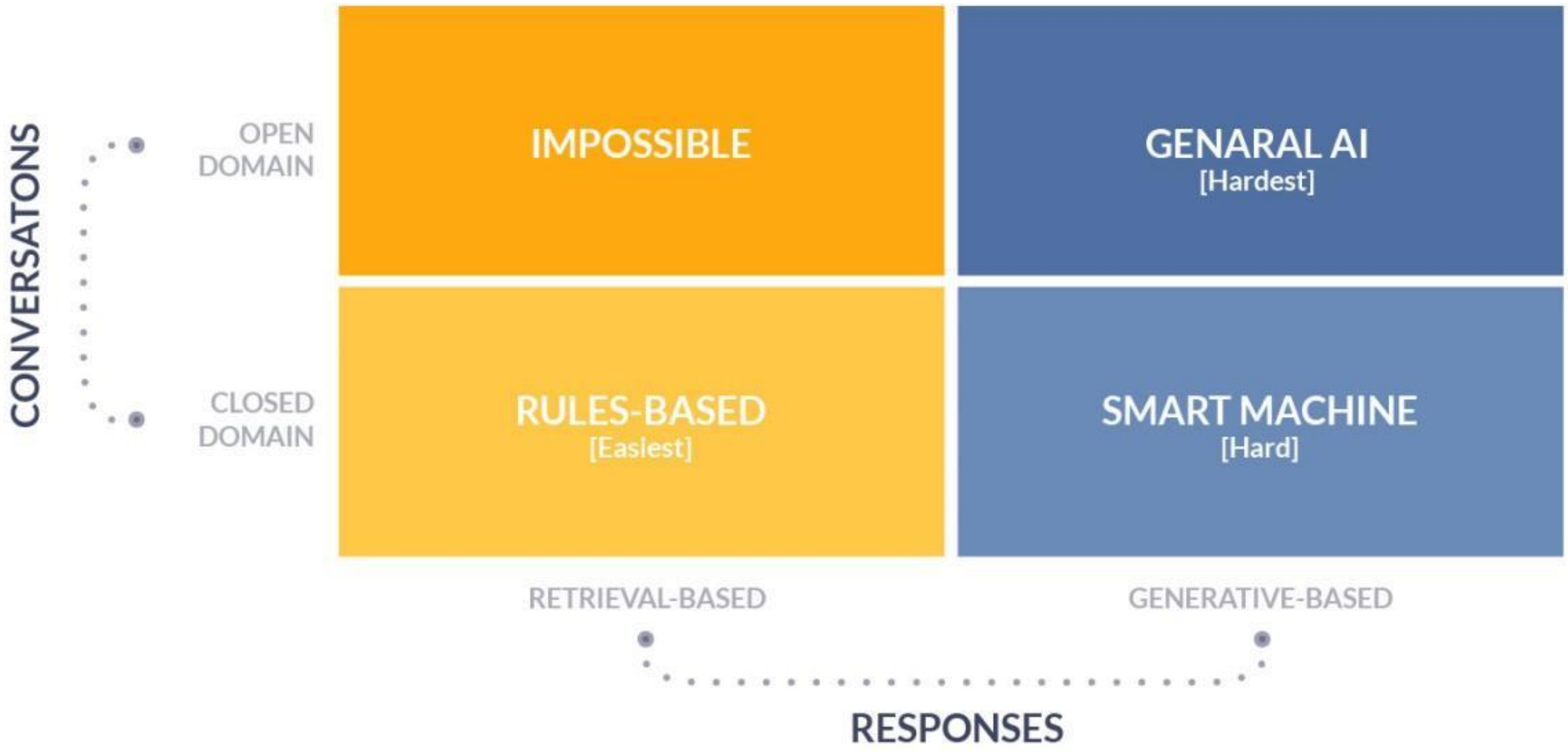
- Why Conversational AI / chatbots?
- Chatbot Conversation Framework
- Use-case in hand
- Chatbot Flow Diagram
- NLU Engine and its components
- Transformer models for Intent classification
- Data and Model Training Summary
- Productizing BERT for CPU Inference

Why Conversational AI / chatbots?

- Messaging is a popular form of interaction and chatbots streamlines the interaction between people and services.
- Easy Scalability of bots.
- Always Available !
- Helpful for organizations with presence in multiple geographies.

According to Forbes, the chatbot market is forecasted to reach $1.25 billion by 2025.

Chatbot Conversation Framework



Use-case in hand

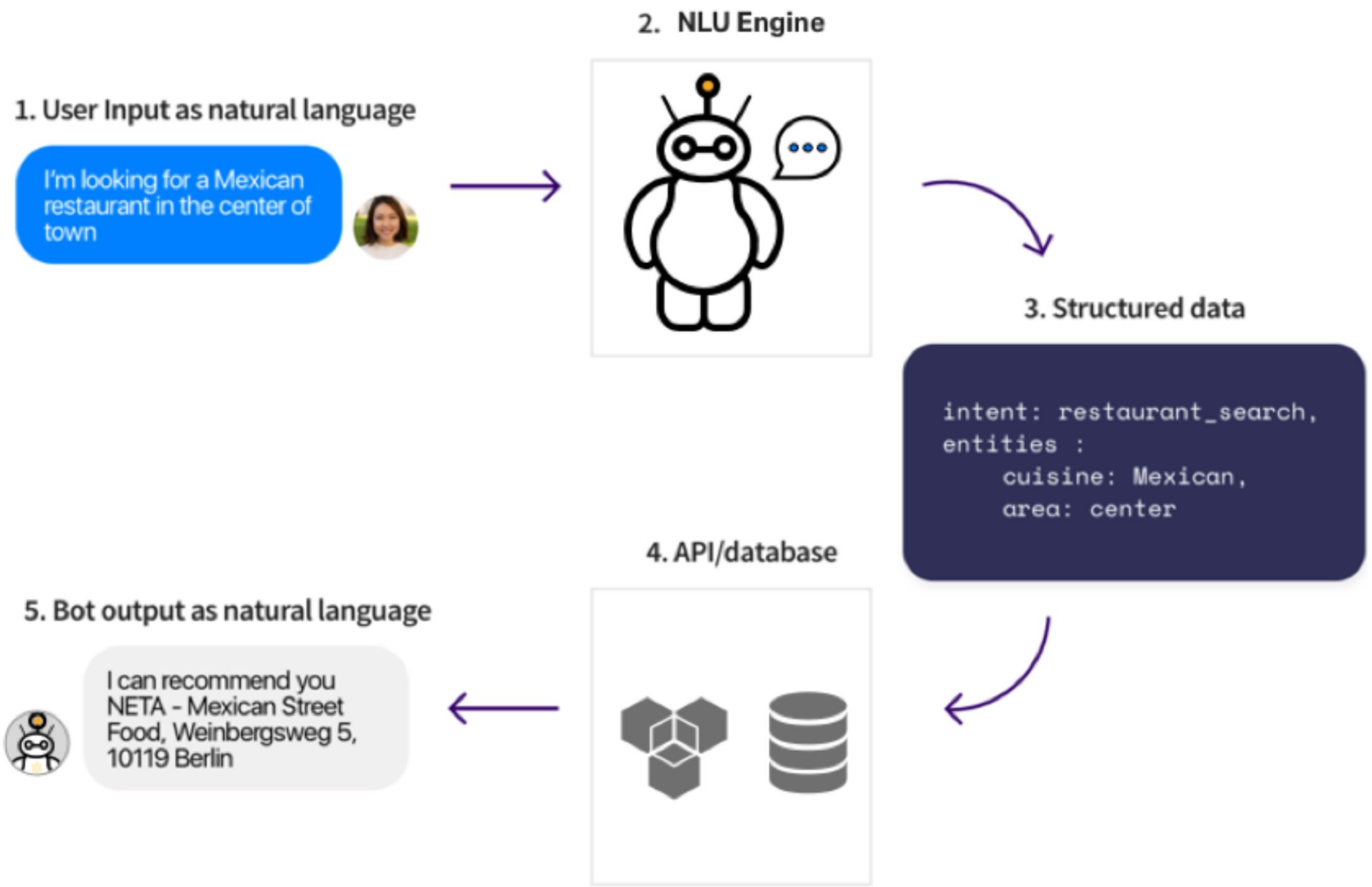
HR policies bot: At Walmart scale, allows employees to query the bot regarding various policies of Walmart.

How will bot help?

- Very convenient to get queries clarified on various policies
- Available 24x7!
- Eliminates person dependency.
- Provides a consistent experience.

Integrated with various communication platforms.

Chatbot Flow Diagram



Components of NLU Engine

Intent: This tells what user would like to do.

Example: search for a restaurant, raise a ticket, book a taxi etc.

Entity: These are the attributes which give details about the user’s task.

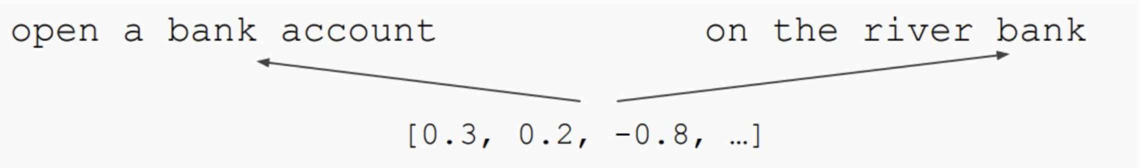
Example: if the user intent is to search for a restaurant possible entities could be:

a. the type of restaurant user is looking for i.e cuisine
b. the area in which the user looking for restaurant i.e location, etc...

In ML & NLP domain above two components are more frequently called as sentence classification and named entity recognition(NER) problems.

Transformers for Intent Classification

- Contextual embeddings

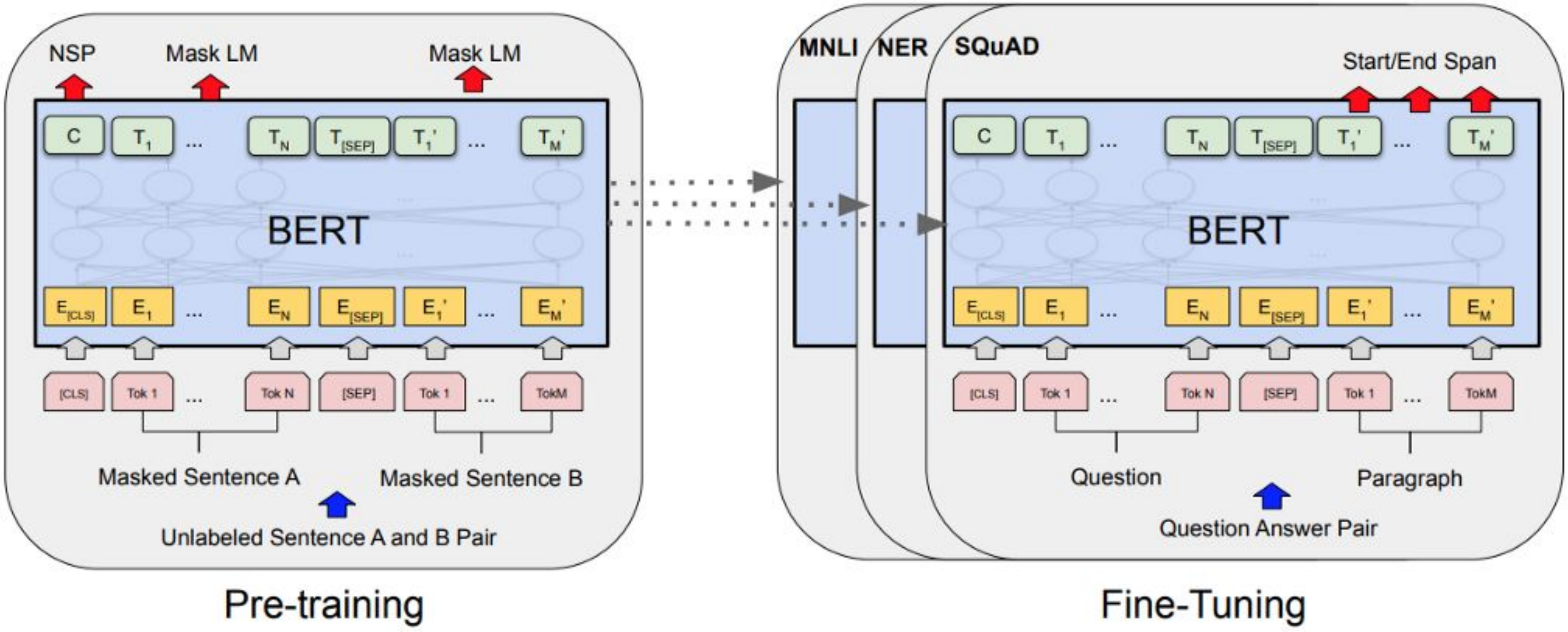


word “bank” has same embedding!

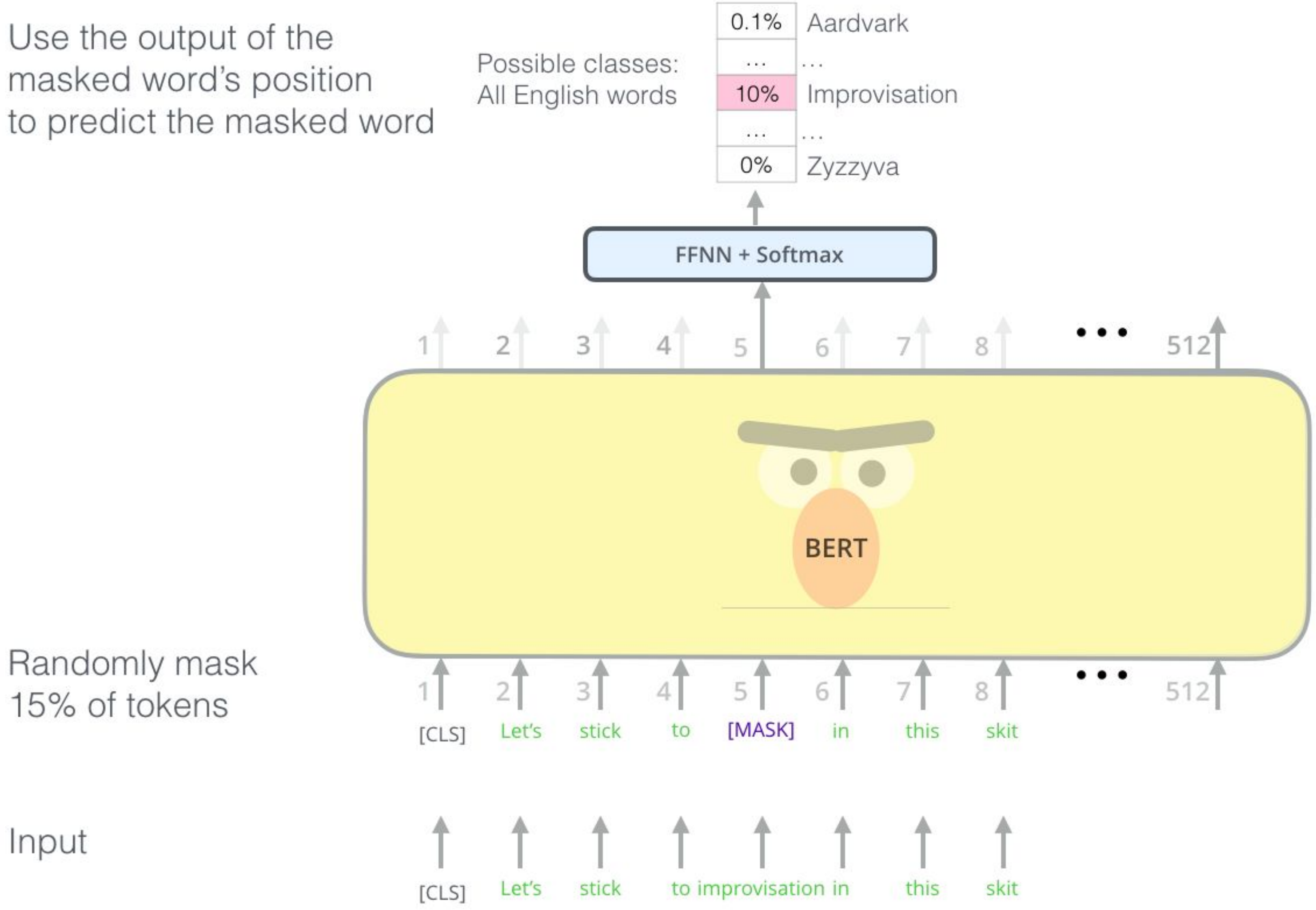
word “bank” has different embedding based on the context in which it is used.

- Parallel training with positional encodings
  - Sentences are processed as a whole, rather than word by word compared to typical RNN/LSTM models.
  - Positional embeddings: encode information related to a specific position of a token in a sentence.

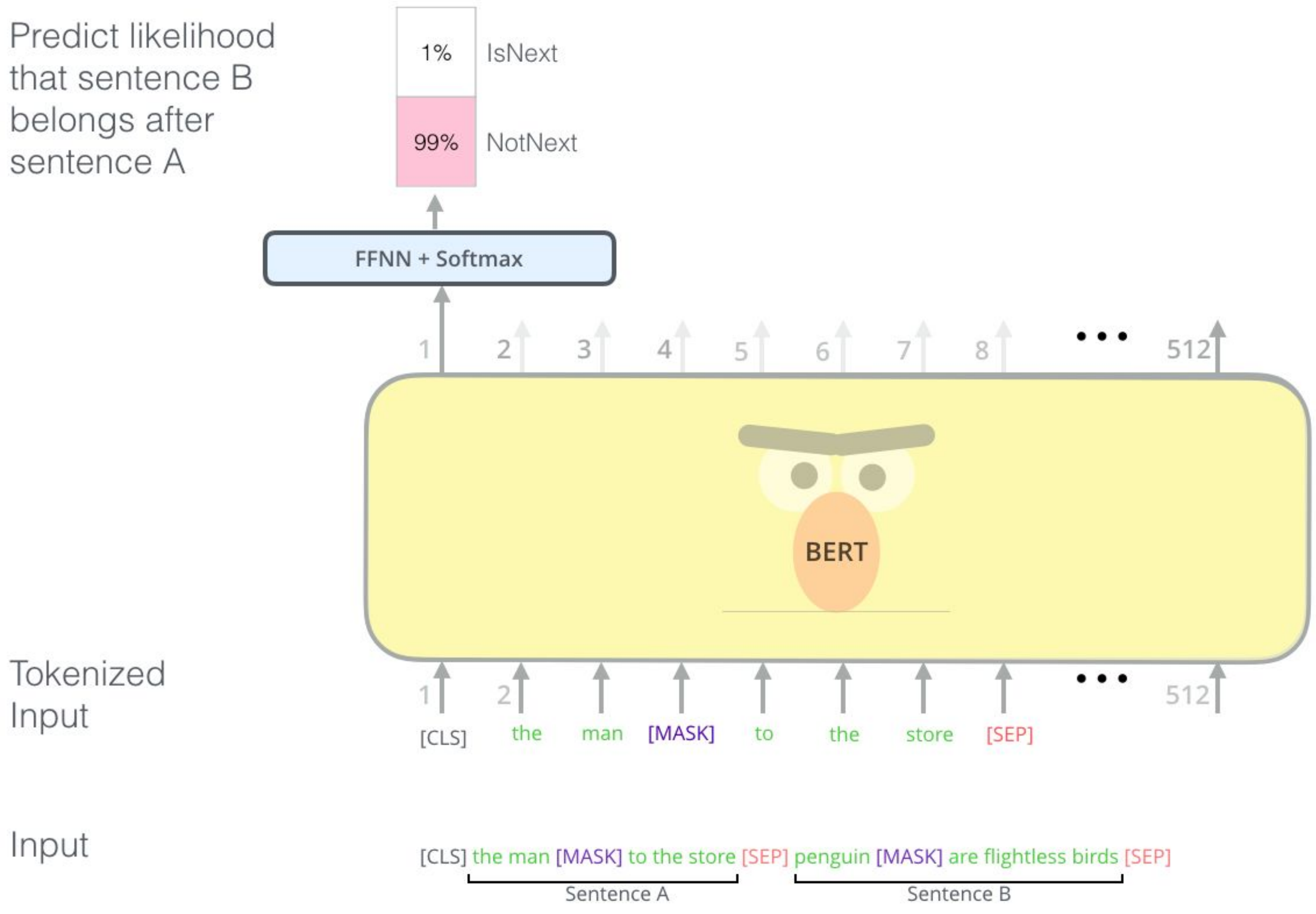
BERT : **B**idirectional **E**ncoder **R**epresentations from **T**ransformers



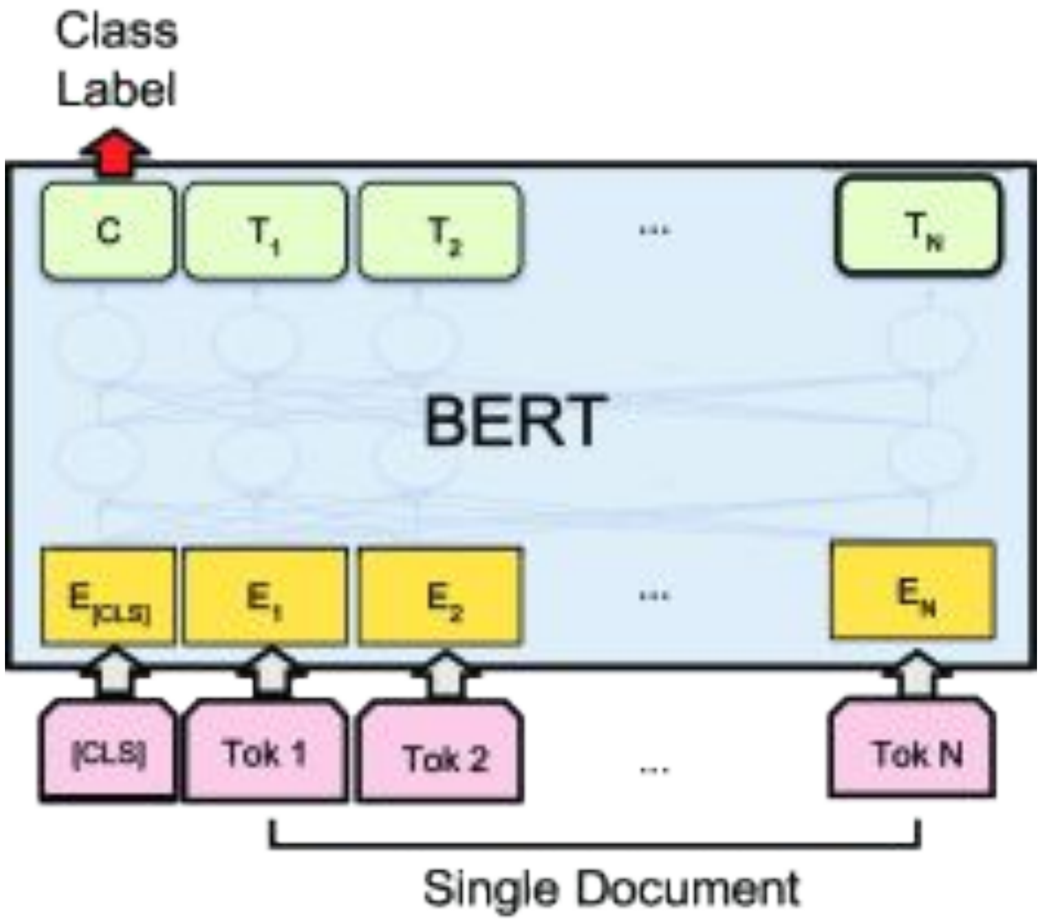
Masked Language Model



Next Sentence Prediction



BERT: CLS token for classification

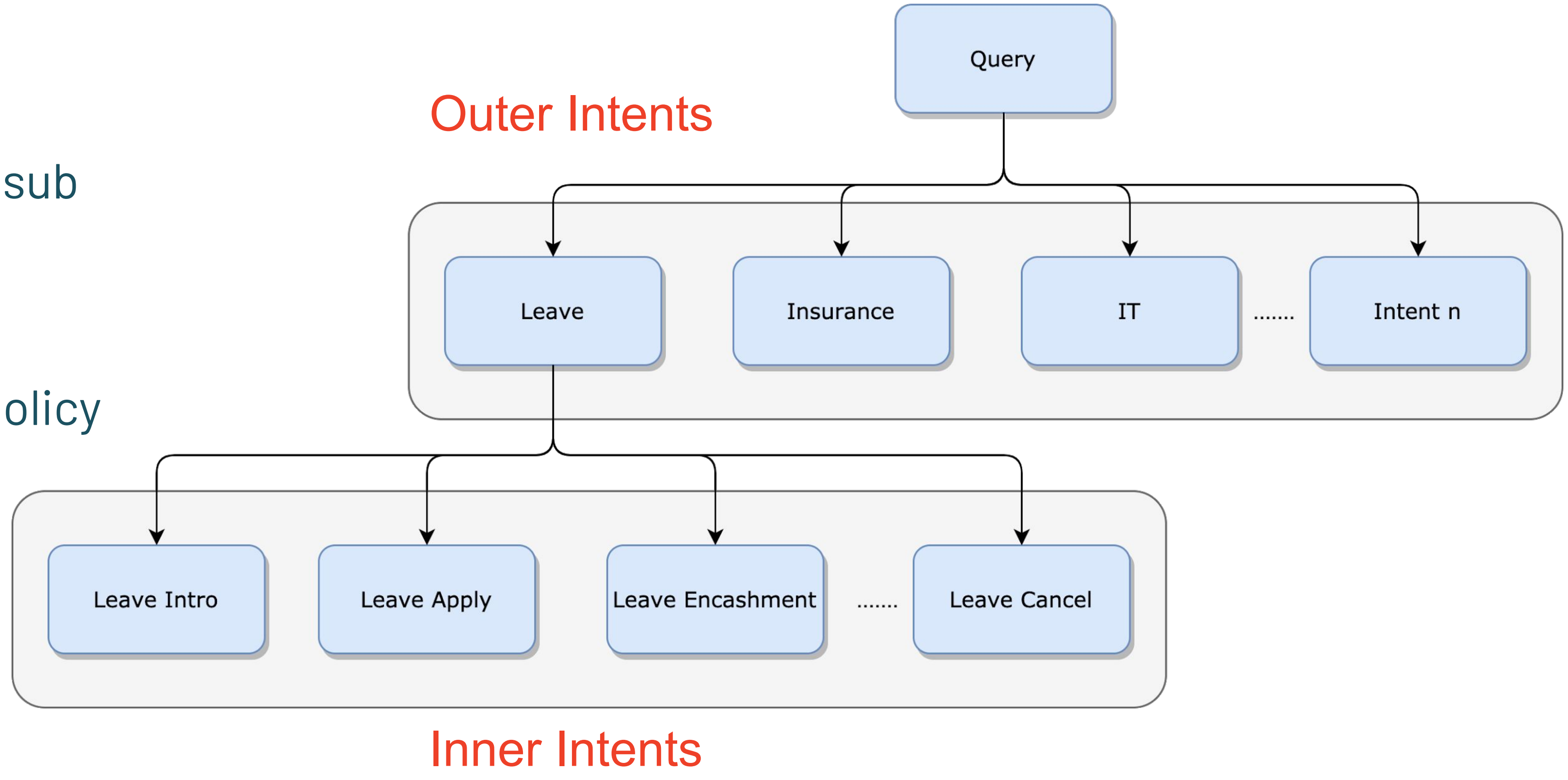


- Why just consider the embeddings of CLS token for classification ?
- Why not from other tokens ?

Different models with accuracy and size over time

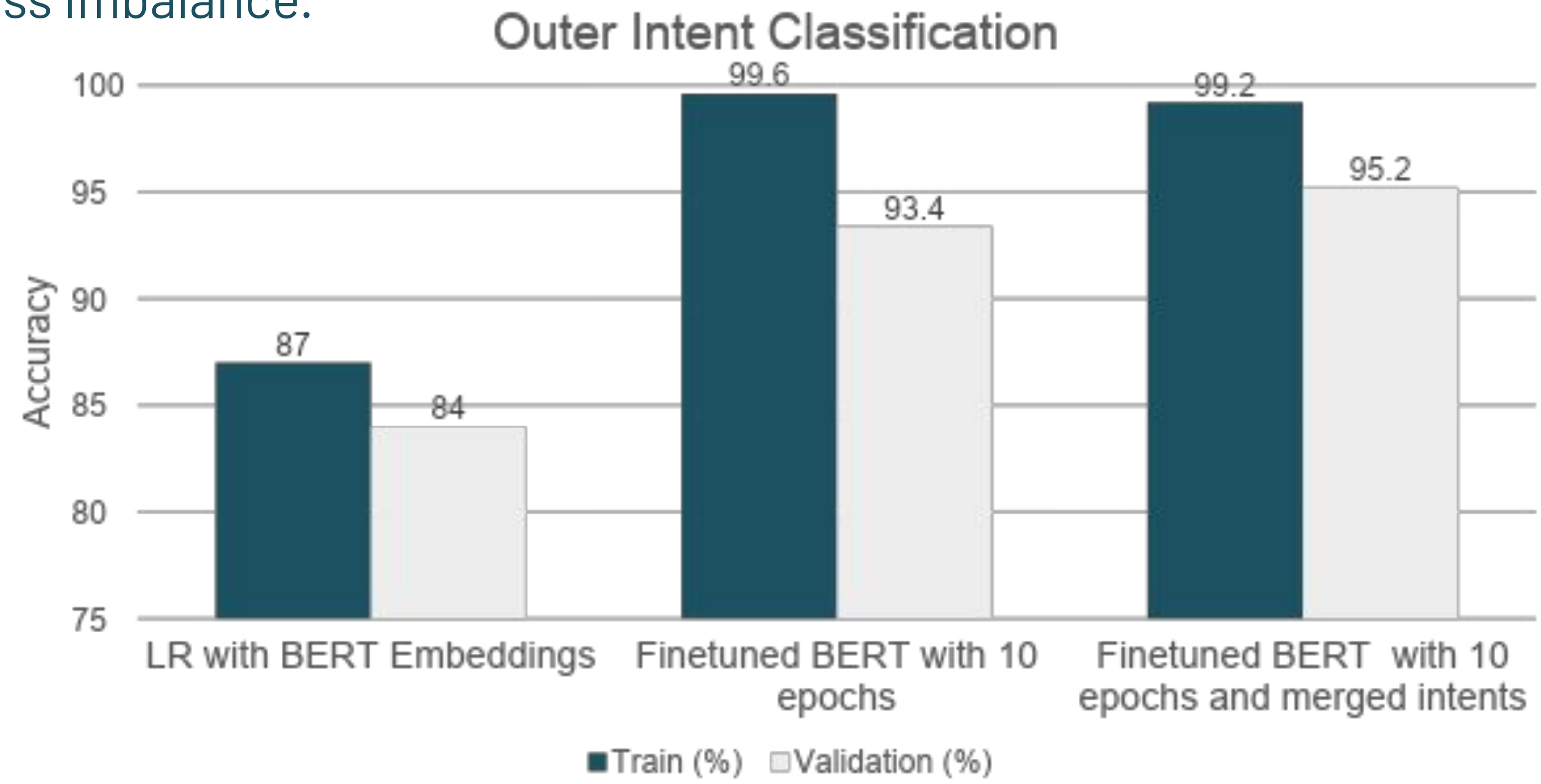
Use-case data summary

- Utterances tagged with 50+ policies.
- Each policy has multiple sub policies.
- Utterances tagged to a policy and a sub policy.



Model Training

- One model for outer policy classification.
- 50+ models for sub policy classification. One model for one policy.
- High class imbalance.



Efficient Model Inference

- Knowledge Distillation
- Quantization
- No padding, batch size == 1

Knowledge Distillation

DistilBERT* has 40 % less parameters than BERT base, 60 % faster with minimal loss in performance.

Ref: Sanh et al., 2019
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

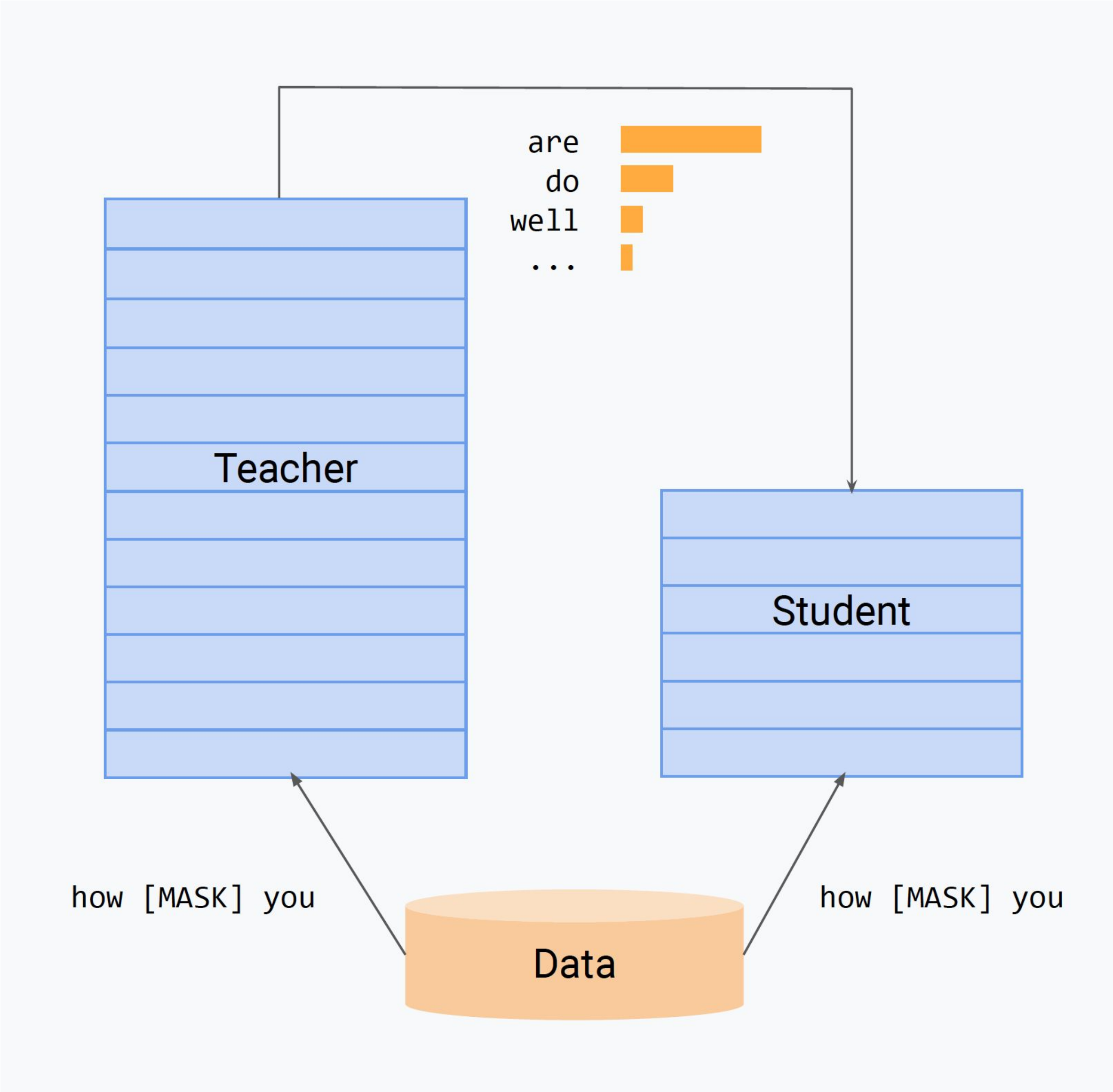
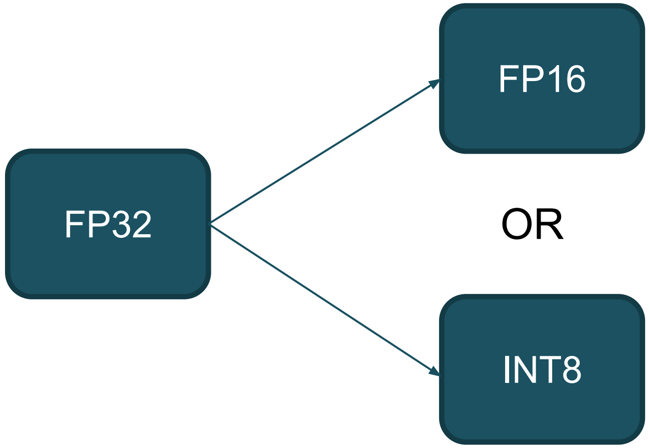


Image Source: High Performance NLP, EMNLP 2020

Quantization

- Modify the datatypes of weights in few layers from fp32 to static int8/fp16, post training.
- Dynamic quantization reduces the size of the model while having limited impact on model performance.




```
quantized_model = torch.quantization.quantize_dynamic(
    model, {torch.nn.Linear}, dtype=torch.qint8
)
print(quantized_model)
```

Image Source: https://pytorch.org/tutorials/intermediate/dynamic_quantization_bert_tutorial.html

No padding

- During training, DL model requires data to be in batches of 16, 32, 64 etc. for efficient training. Since text comes in variable length, we zero pad the tensors to a fixed size.

- During inference, zero padding is not necessary since batch size is 1. There is only one sentence in each batch.

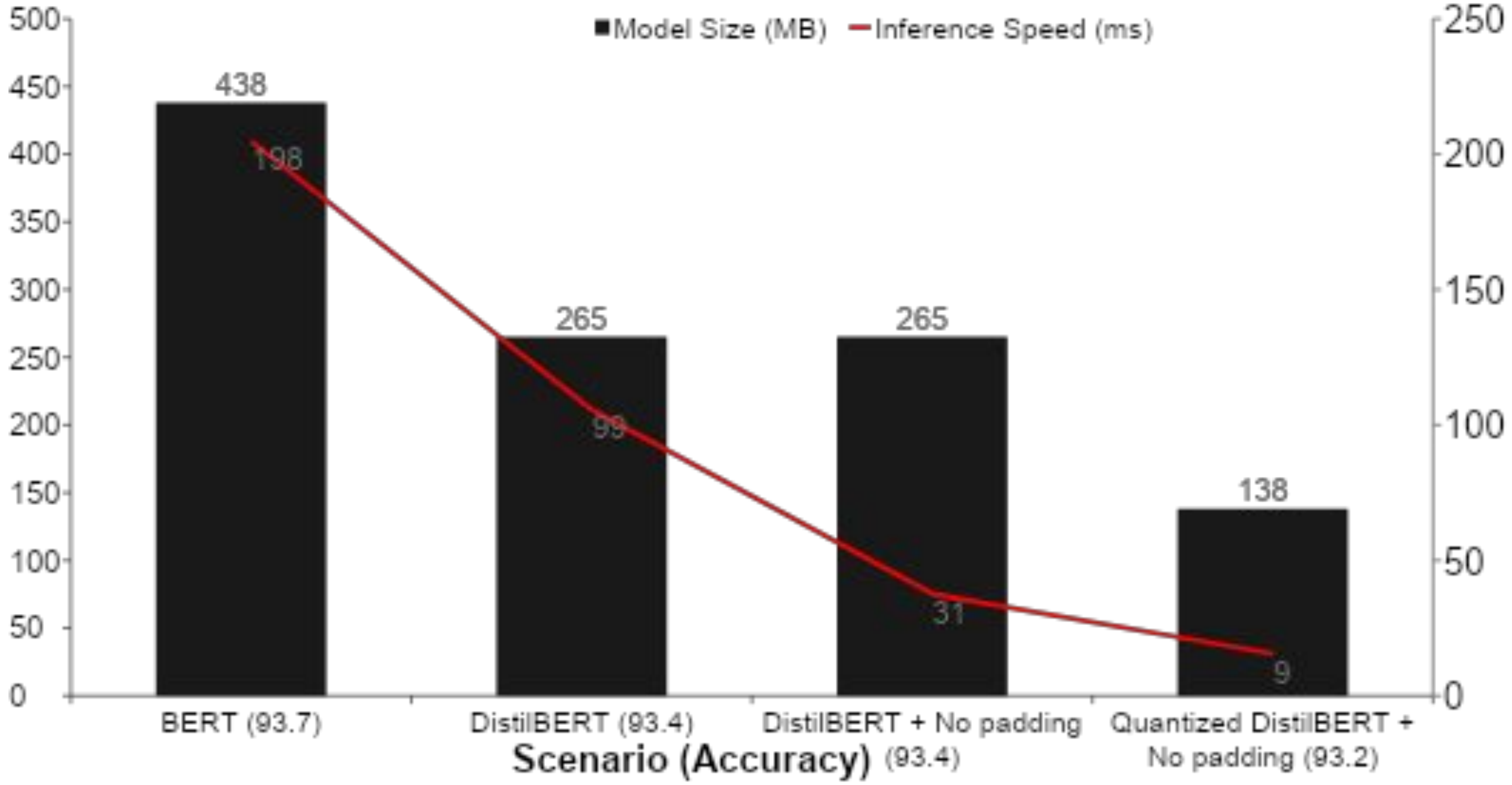
Padding to max length

Text	1	2	3	4	5	6	7	8	9	10
Adam likes scifi movies	100	250	260	135	144	0	0	0	0	0
BERT is so huge, it will never go into production	100	360	430	443	123	205	18 4	237	657	12
What can you do for me	100	234	431	257	426	123	0	0	0	0
Hello there!	100	200	0	0	0	0	0	0	0	0

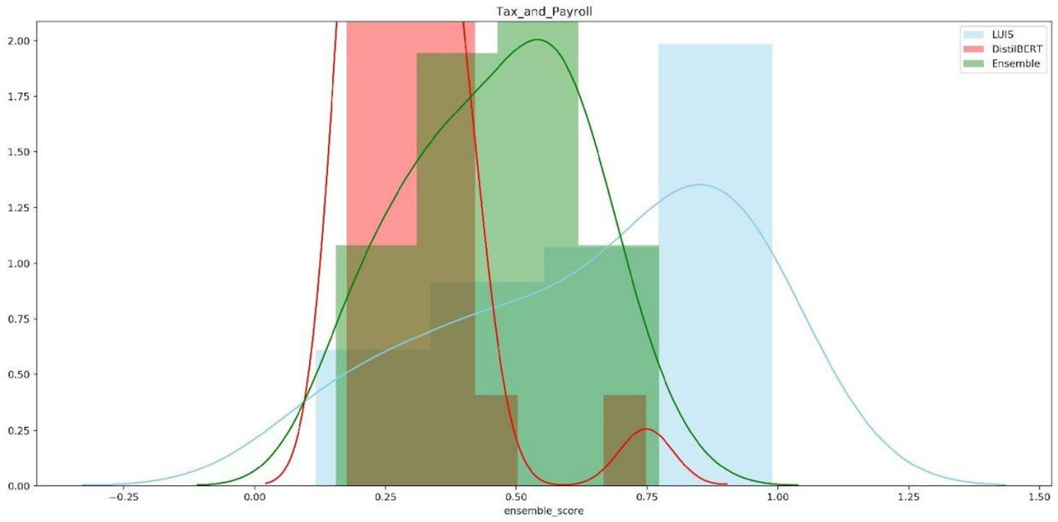
No padding

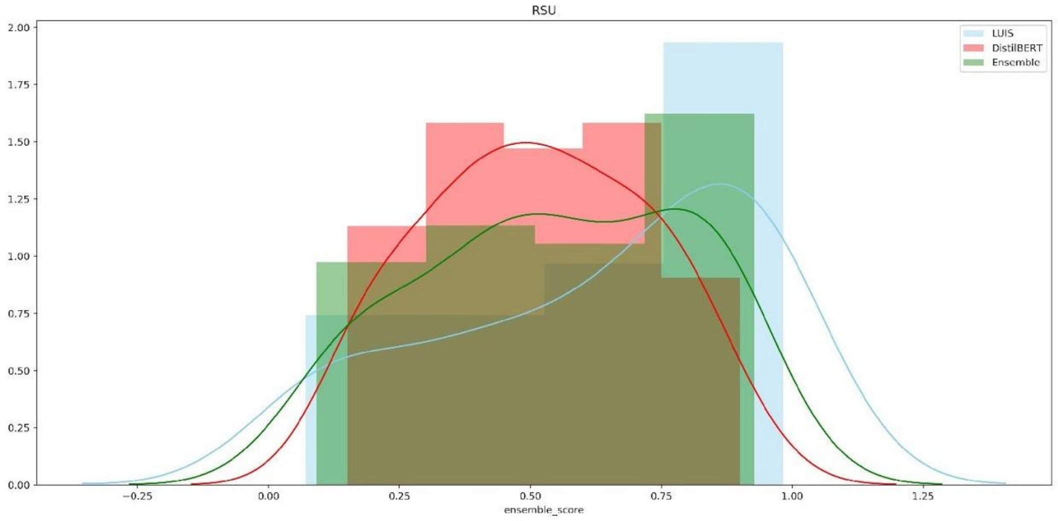
Text	1	2	3	4	5	6	7	8	9	10
Adam likes scifi movies	100	250	260	135	144					
BERT is so huge, it will never go into production	100	360	430	443	123	205	18 4	237	657	12
What can you do for me	100	234	431	257	426	123				
Hello there!	100	200								

Productizing BERT for CPU Inference



Ensembling LUIS and DistilBERT





Policy	No of records	LUIS Accuracy	DistilBERT Accuracy	Ensemble Accuracy	Improvement
Tax_and_Payroll	30	73.30%	70.00%	86.70%	13.40%
MIP	13	69.20%	69.20%	76.90%	7.70%
IJP	34	88.20%	82.40%	94.10%	5.90%
WFH	139	85.60%	92.10%	91.40%	5.80%
Insurance	288	87.20%	91.30%	91.70%	4.50%
Relocation	149	83.90%	89.30%	87.90%	4.00%
RSU	59	67.80%	61.00%	71.20%	3.40%
Leave	195	87.70%	86.20%	89.70%	2.00%
MobileAndInternet	50	86.00%	90.00%	88.00%	2.00%
COVID_Reimbursement	59	98.30%	100.00%	100.00%	1.70%

Ensembling improved by good margin when LUIS and DistilBERT are distinctive in predicting.

Team behind the project

Name	Designation
Dinesh Ladi	Data Scientist
Mainak Mitra	Senior Data Scientist
Rajesh Shreedhar Bhat	Senior Data Scientist
Ritish Menon	Senior Manager II – Data Science

Sample Code

https://bit.ly/3gMuKUc

Questions ??

Thank you !