NVIDIA

Overview of Large Language Models

Agenda

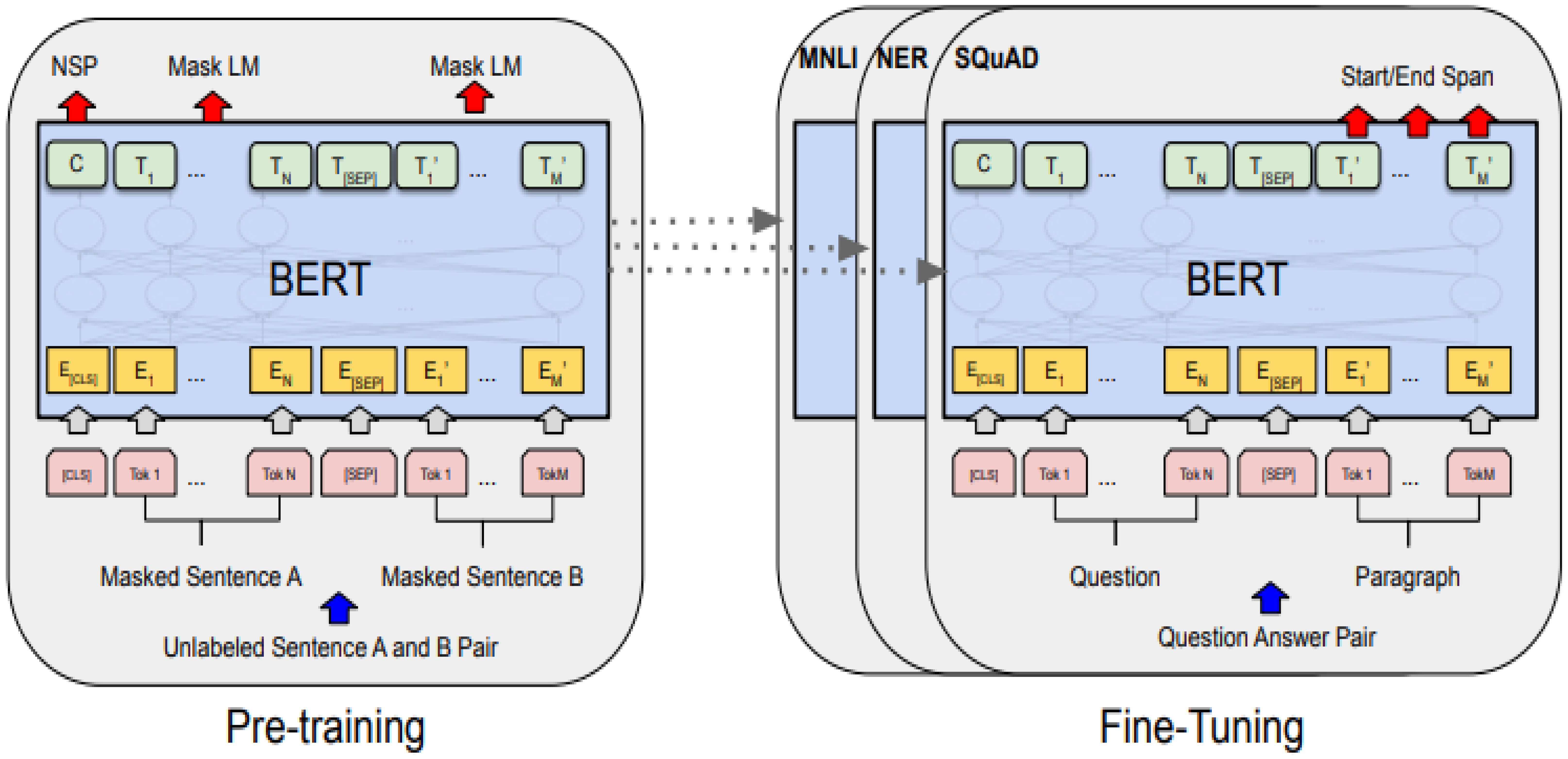
- Introduction to Question Answering (QA) Models and Architectures
- Overview of Question Answering Dataset

Introduction to Question Answering (QA) Models and Architectures

BERT

BERT Model for Question Answering

- **BERT** (Bidirectional Encoder Representations from Transformers) model is typically trained in two phases, **pre-training** and **fine-tuning.**
- BERT is pre-trained using two unsupervised tasks: Masked LM & Next Sentence Prediction (NSP)

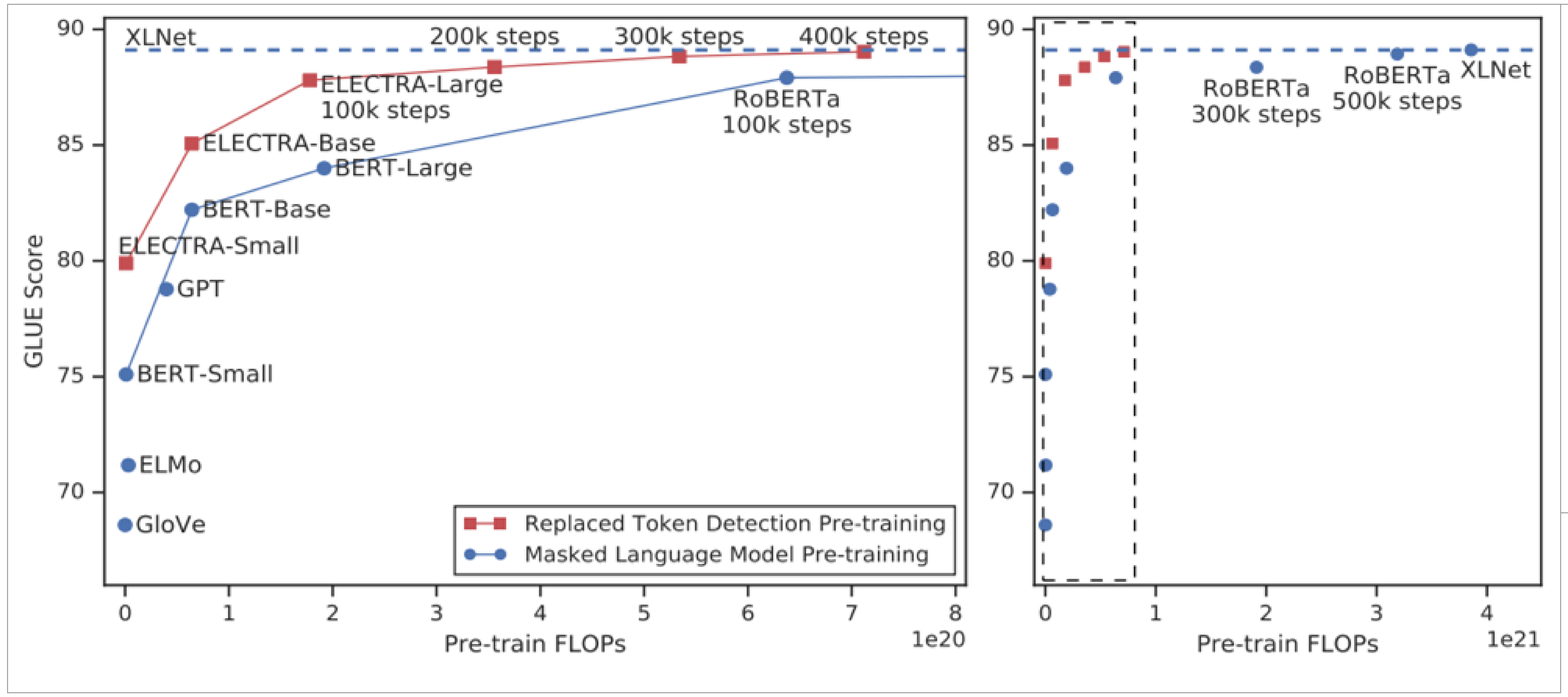


Source: [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]]

ELECTRA

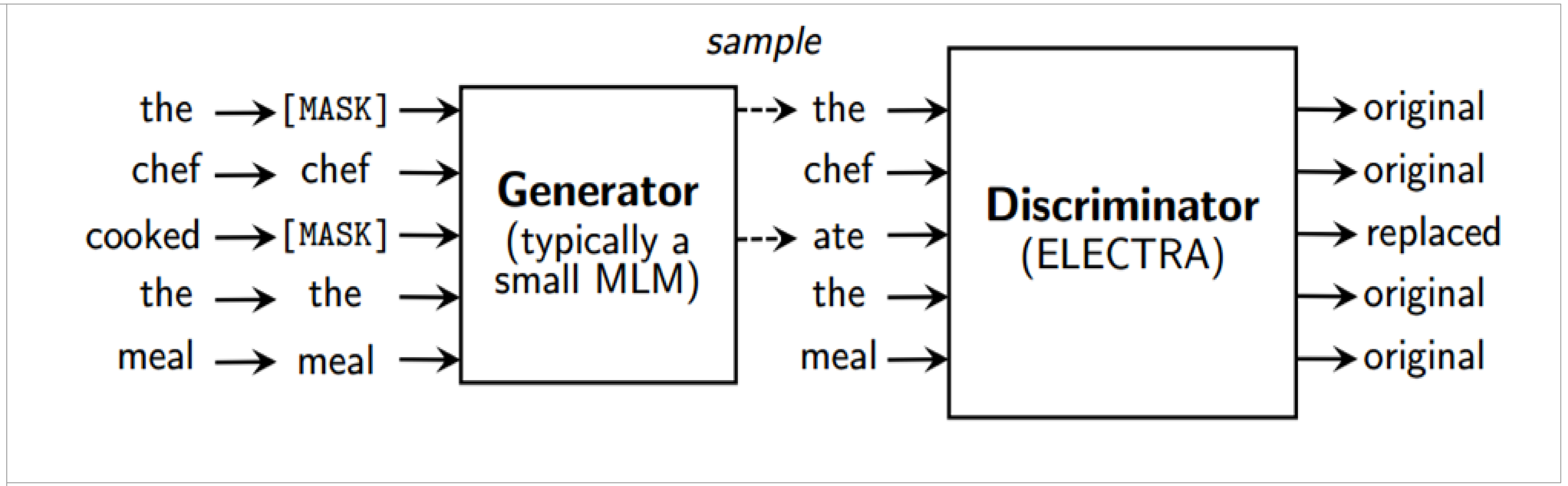
ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS

(A)



(A) Replaced token detection pre-training consistently outperforms masked language model pre-training given the same compute budget. The left graph is a zoomed-in view of the dashed box.

(B)



(B) An overview of replaced token detection. The generator can be any model that produces an output distribution over tokens, but a small masked language model that is trained jointly with the discriminator is used. Though the model is structured like GAN, the generator is trained with maximum likelihood. After pre-training, the generator is removed, and only the fine-tuning of the discriminator (ELECTRA model) on the downstream task is carried out.

Source: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, Kevin Clark, et al., ICLR, 2020

ELECTRA

Pre-train and Fine-tune Hyperparameters

Hyperparameter	Small	Base	Large
Number of layers	12	12	24
Hidden Size	256	768	1024
FFN inner hidden size	1024	3072	4096
Attention heads	4	12	16
Attention head size	64	64	64
Embedding Size	128	768	1024
Generator Size (multiplier for hidden-size, FFN-size, and num-attention-heads)	1/4	1/3	1/4
Mask percent	15	15	25
Learning Rate Decay	Linear	Linear	Linear
Warmup steps	10000	10000	10000
Learning Rate	5e-4	2e-4	2e-4
Adam ϵ	1e-6	1e-6	1e-6
Adam β_1	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999
Attention Dropout	0.1	0.1	0.1
Dropout	0.1	0.1	0.1
Weight Decay	0.01	0.01	0.01
Batch Size	128	256	2048
Train Steps (BERT/ELECTRA)	1.45M/1M	1M/766K	464K/400K

Pre-train hyperparameters

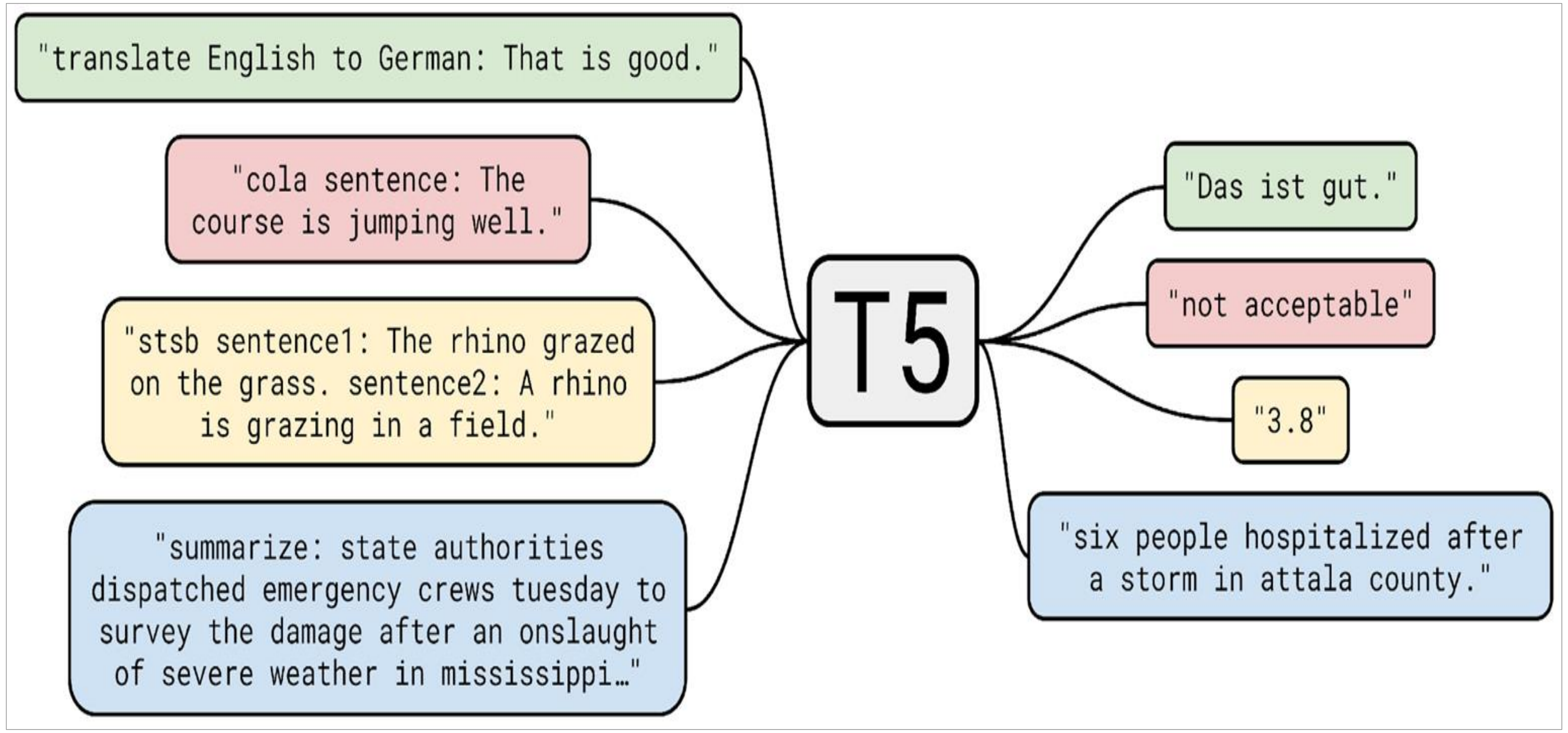
Hyperparameter	GLUE Value
Learning Rate	3e-4 for Small, 1e-4 for Base, 5e-5 for Large
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Layerwise LR decay	0.8 for Base/Small, 0.9 for Large
Learning rate decay	Linear
Warmup fraction	0.1
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0
Batch Size	32
Train Epochs	10 for RTE and STS, 2 for SQuAD, 3 for other tasks

Fine-tune hyperparameters

Source: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, Kevin Clark, et al., ICLR, 2020

T5 Model

T5 Model (Text-to-Text Transfer Transformer)

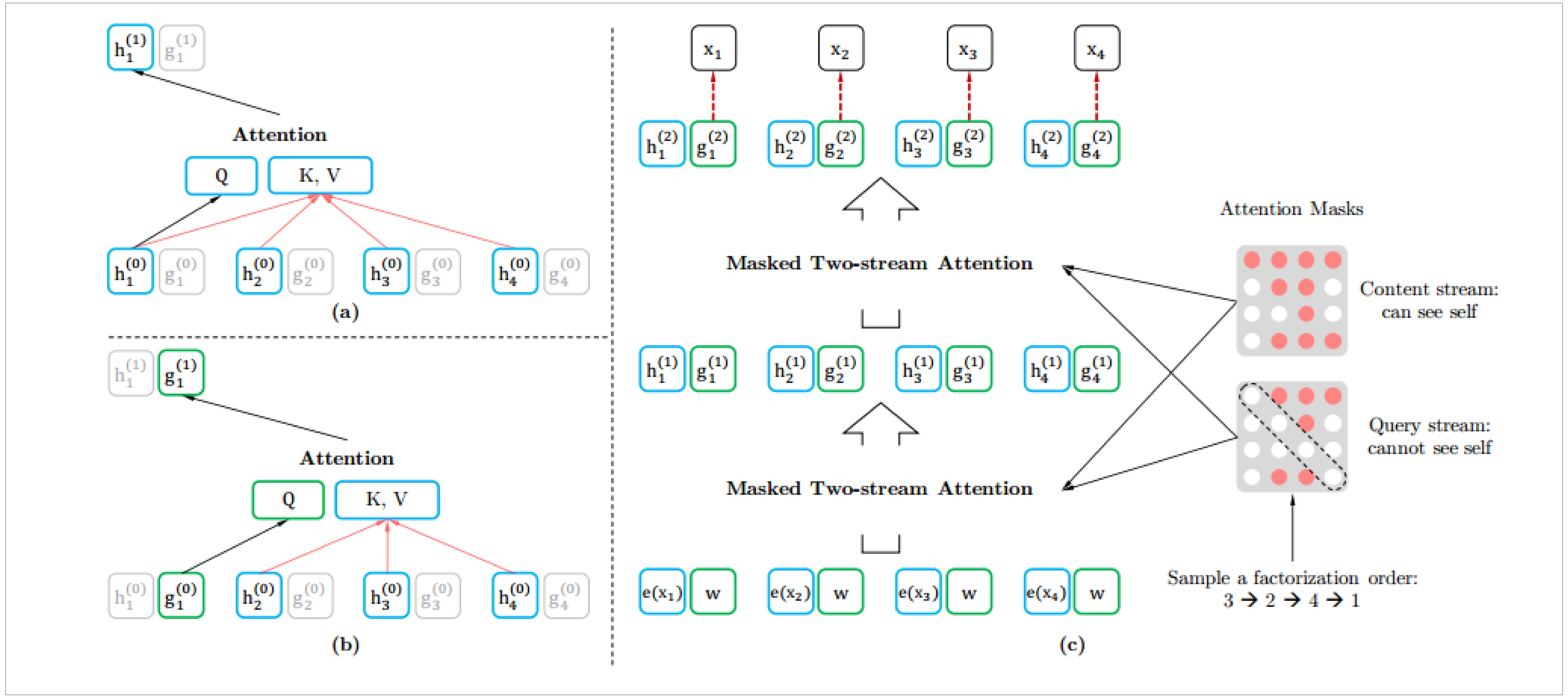


- T5 is all about reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings
- The T5 model, pre-trained on C4 dataset, achieves state-of-the-art results on many NLP benchmarks.

Source: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Raffel et al., 2020

XLNET

XLNet: Generalized Autoregressive Pretraining for Language Understanding



(a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access to information about the content x_{zt} . (c): Overview of the permutation language modeling training with two-stream attention.

- XLNet outperforms BERT on 20 tasks, often by a large margin, including question answering, natural language inference, sentiment analysis, and document ranking

Source: https://arxiv.org/pdf/1906.08237.pdf

XLNET

Hyperparameters

Hparam	Value
Number of layers	24
Hidden size	1024
Number of attention heads	16
Attention head size	64
FFN inner hidden size	4096
Hidden Dropout	0.1
GeLU Dropout	0.0
Attention dropout	0.1
Partial prediction K	6
Max sequence length	512
Batch size	8192
Learning rate	4e-4
Number of steps	500K
Warmup steps	40,000
Learning rate decay	linear
Adam epsilon	1e-6
Weight decay	0.01

Pre-train hyperparameters

Hparam	RACE	SQuAD	MNLI	Yelp-5
Dropout		0.1		
Attention dropout		0.1		
Max sequence length	512	512	128	512
Batch size	32	48	128	128
Learning rate	2e-5	3e-5	2e-5	1e-5
Number of steps	12K	8K	10K	10K
Learning rate decay		linear		
Weight decay		0.01		
Adam epsilon	1e-6	1e-6	1e-6	1e-6
Layer-wise lr decay	1.0	0.75	1.0	1.0

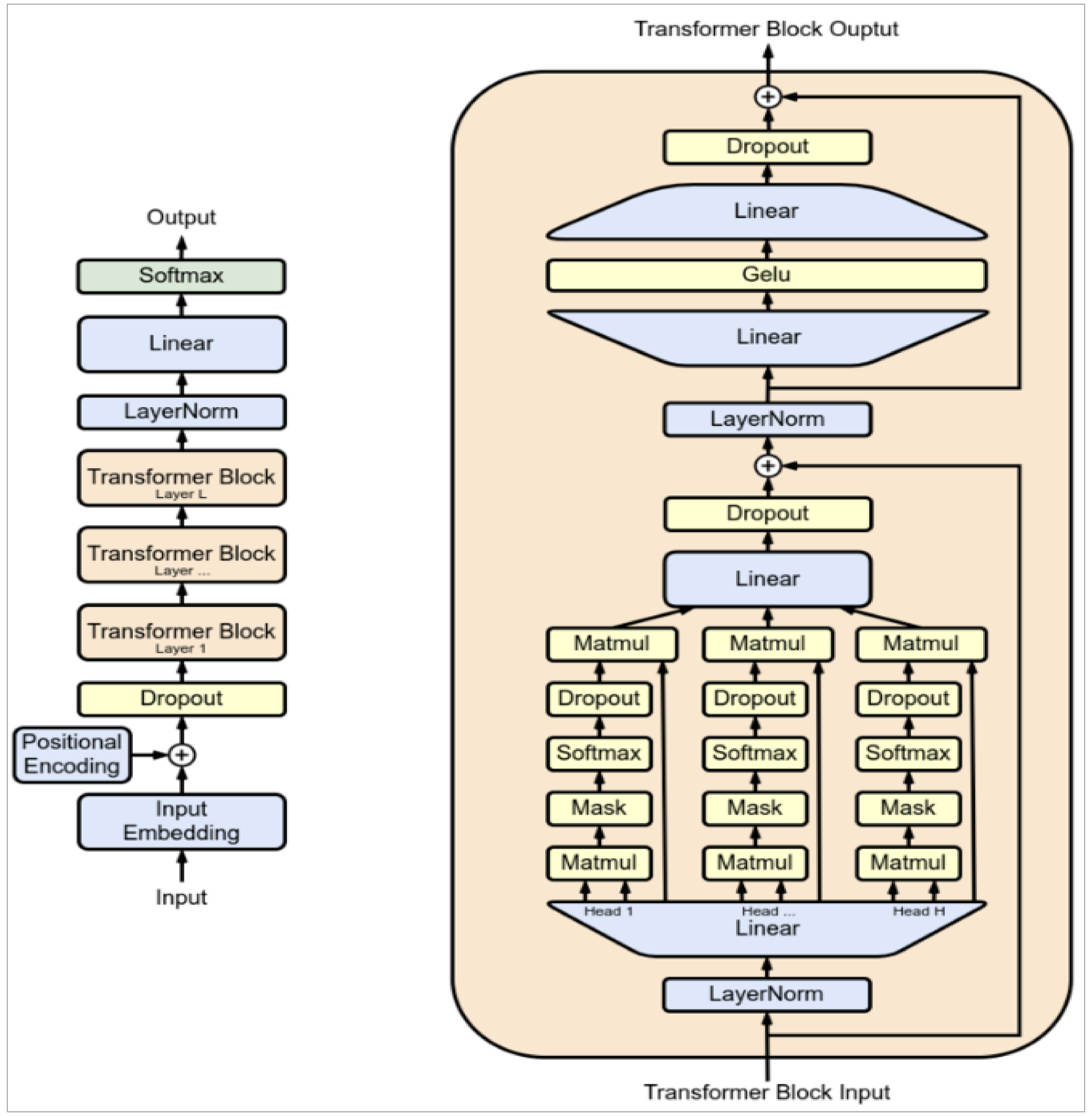
Fine-tune hyperparameters

- XLNet is fine-tuned on four datasets:
  - RACE
  - SQuAD
  - MNLI
  - Yelp-5

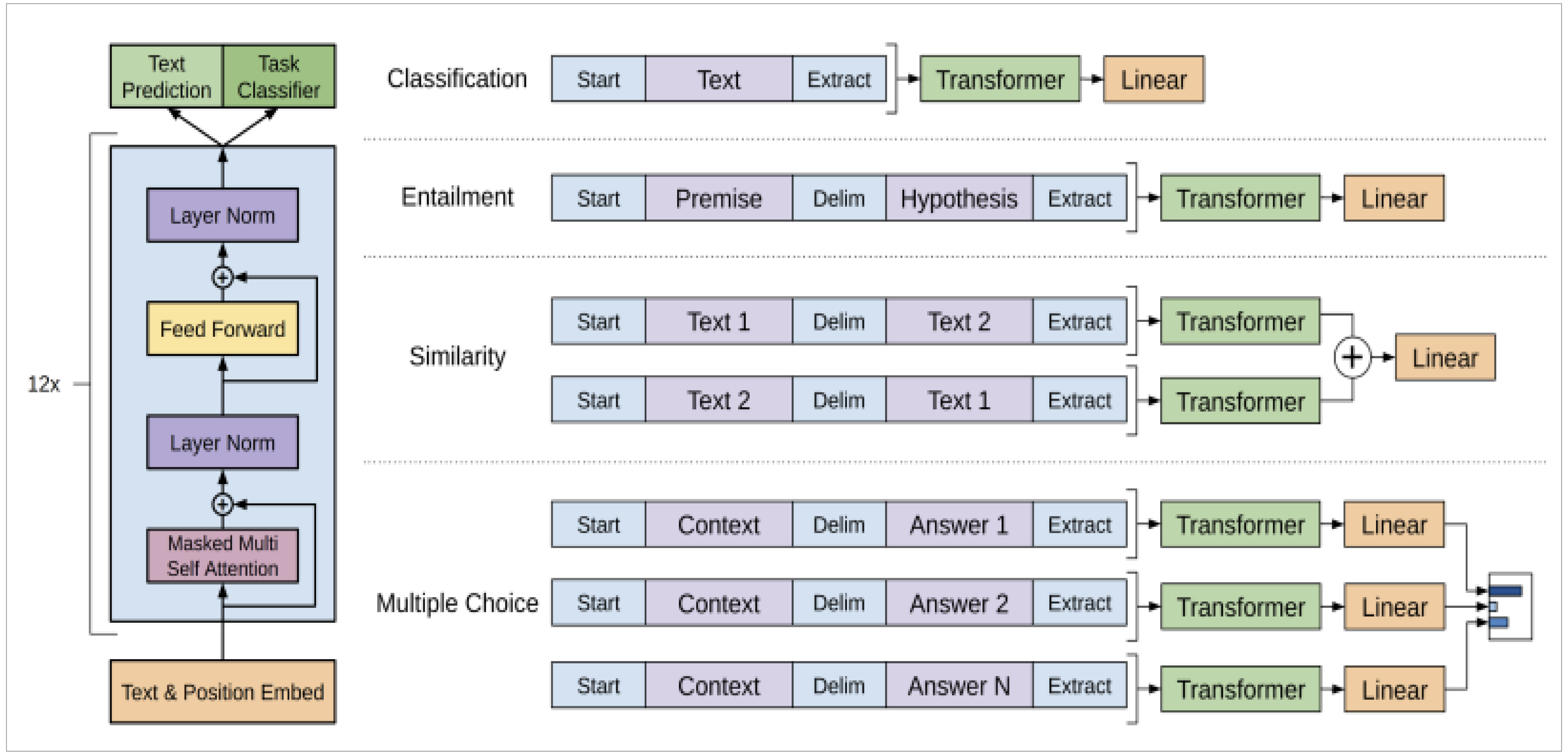
Source: https://arxiv.org/pdf/1906.08237.pdf

GPT Model

GPT (Generative Pre-trained Transformer) Architecture



Original GPT architecture



Improved GPT architecture by OpenAI

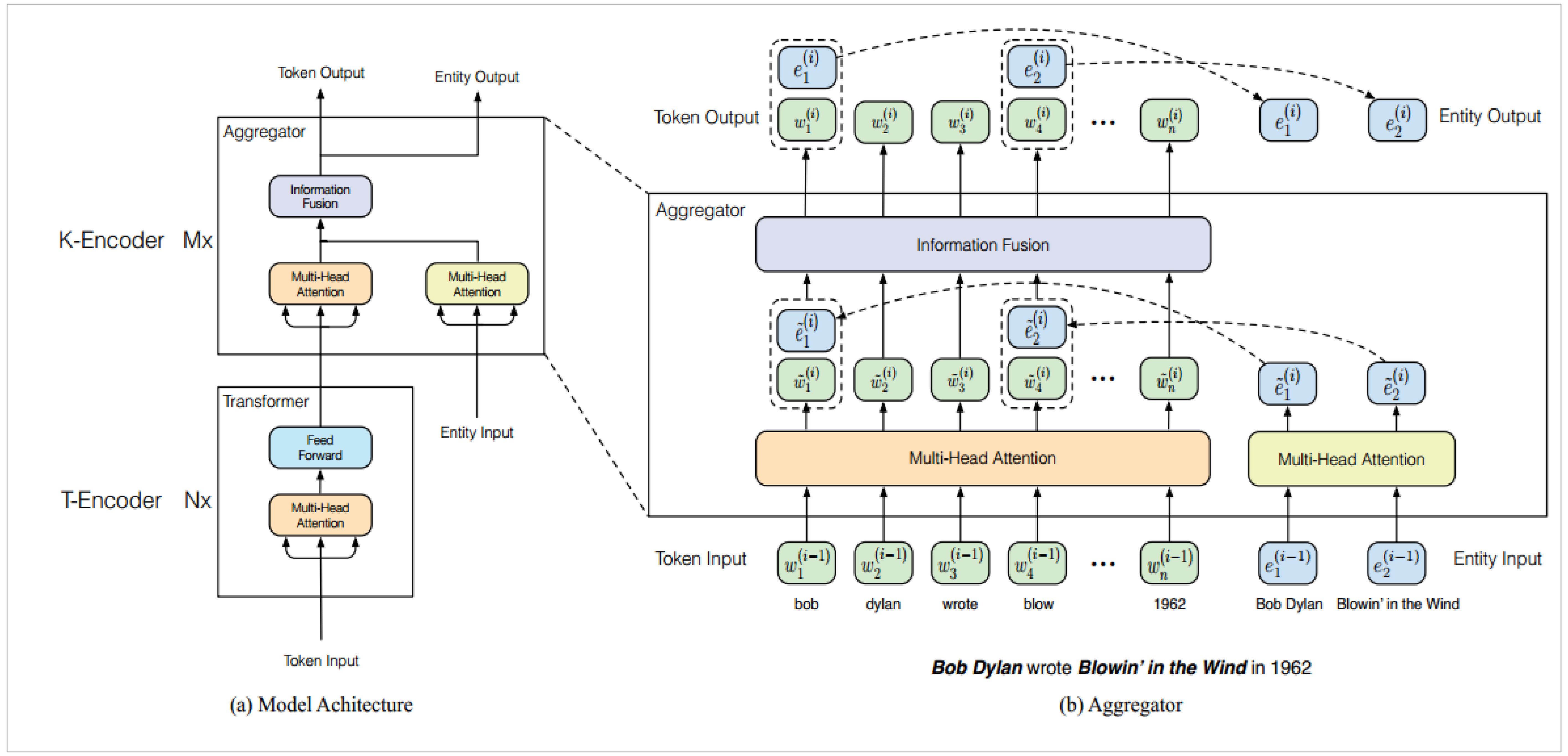
Source: https://en.wikipedia.org/wiki/GPT-2

Source:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

ERNIE Model

ERNIE: Enhanced Language Representation with Informative Entities

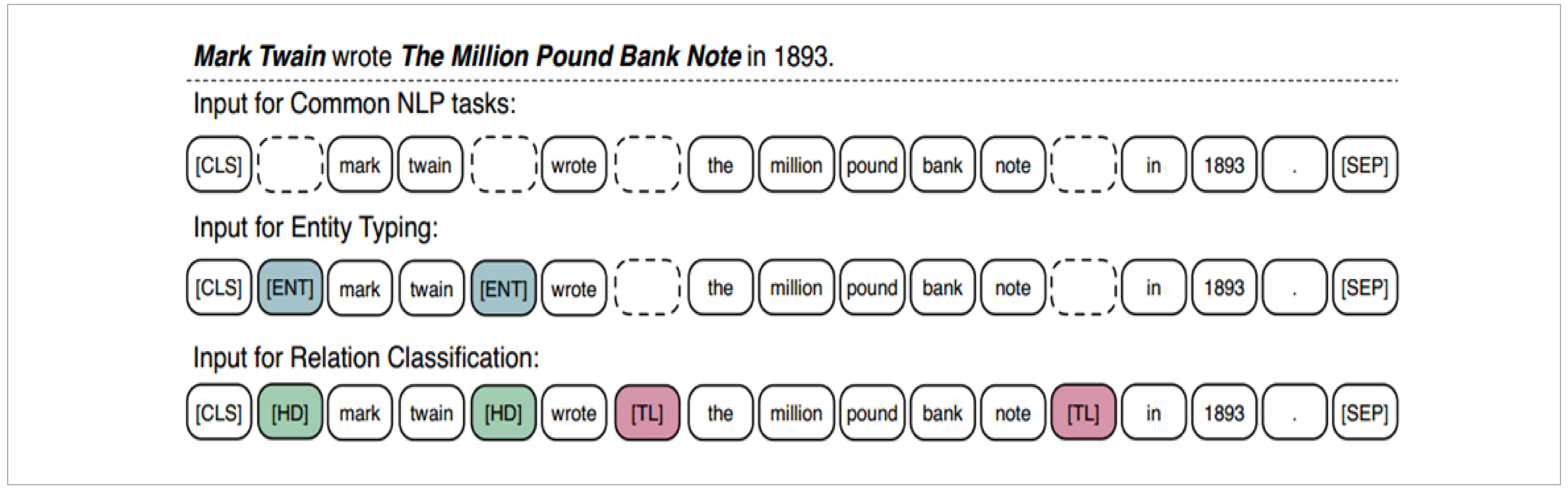
(A)



(a) Model Achitecture

(b) Aggregator

(B)



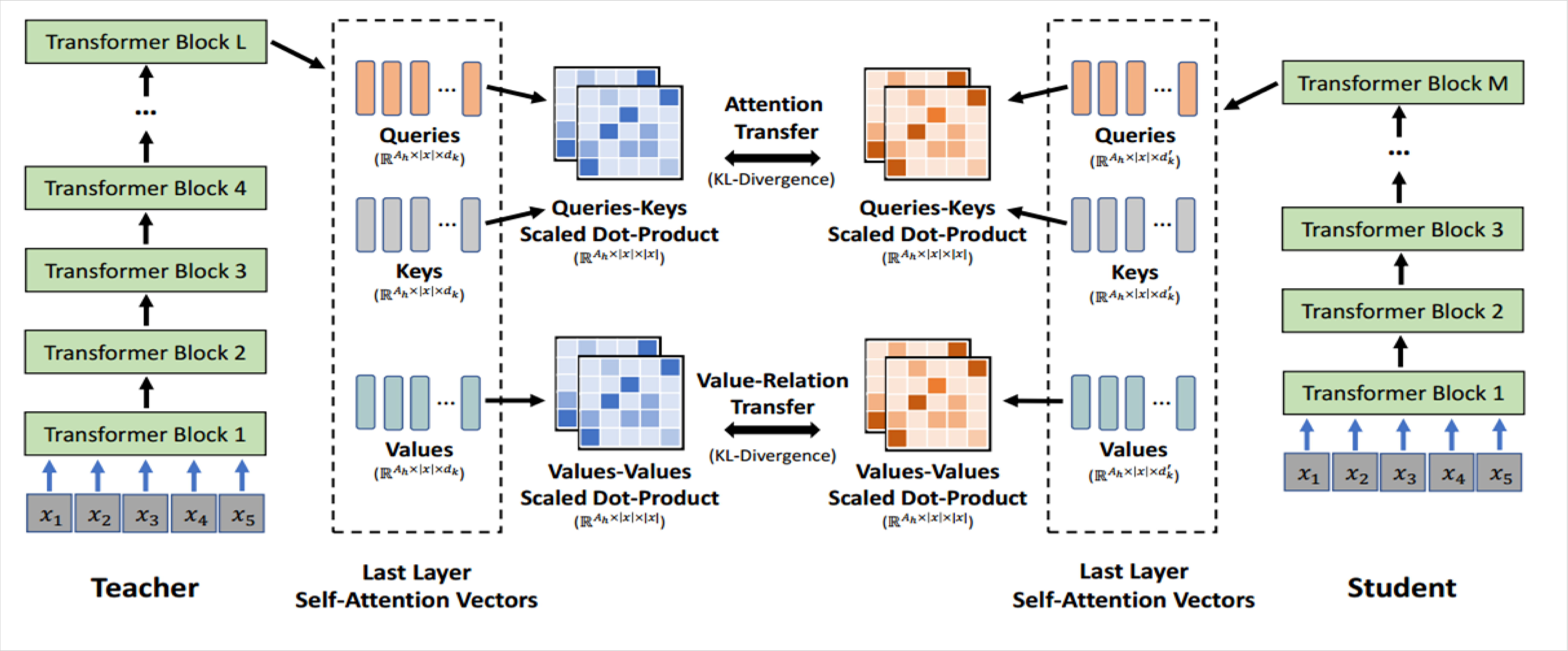
(B) Modifying the input sequence for the specific tasks. The dotted rectangles are used as placeholder to align tokens among different types of input. The coloured rectangles represent mark tokens.

(A) The left part is the architecture of ERNIE. On the right is the aggregator for the mutual integration of the input of tokens and entities. The information fusion layer takes two kinds of input: one is the token embedding, and the other one is the concatenation of the token embeddings and entity embedding. After information fusion, it outputs new token embeddings and entity embeddings for the next layer.

Source: https://arxiv.org/pdf/1905.07129.pdf

MINILM Model

MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers



Overview of Deep Self-Attention Distillation.

- The The student is trained by deeply mimicking the self-attention behavior of the last Transformer layer of the teacher.
- In addition to the self-attention distributions, the self-attention value-relation transfer is introduced to help the student achieve a deeper mimicry.
- The student models are named as MINILM

Source: https://arxiv.org/pdf/2002.10957.pdf

MINILM Model

The Teacher Model, Knowledge Distillation, and Use Cases

- The teacher model is trained using pre-training datasets which includes 160GB text corpora from English Wikipedia, BookCorpus, OpenWebText6, CC-News, and Stories.
- The teacher model is distilled into 12-layer and 6-layer models with 384 hidden sizes using the same corpora.
- The 12x384 model is used as the teacher assistant to train the 6x384 model.

- Knowledge distillation is a promising way to compress large models while maintaining accuracy.
- It transfers the knowledge of a large model or an ensemble of neural networks (teacher) to a single lightweight model (student).

- Use cases for MINILM include question generation, abstract summarization, multilingual, and extractive question answering.

Source: https://arxiv.org/pdf/2002.10957.pdf

Question Answering Model

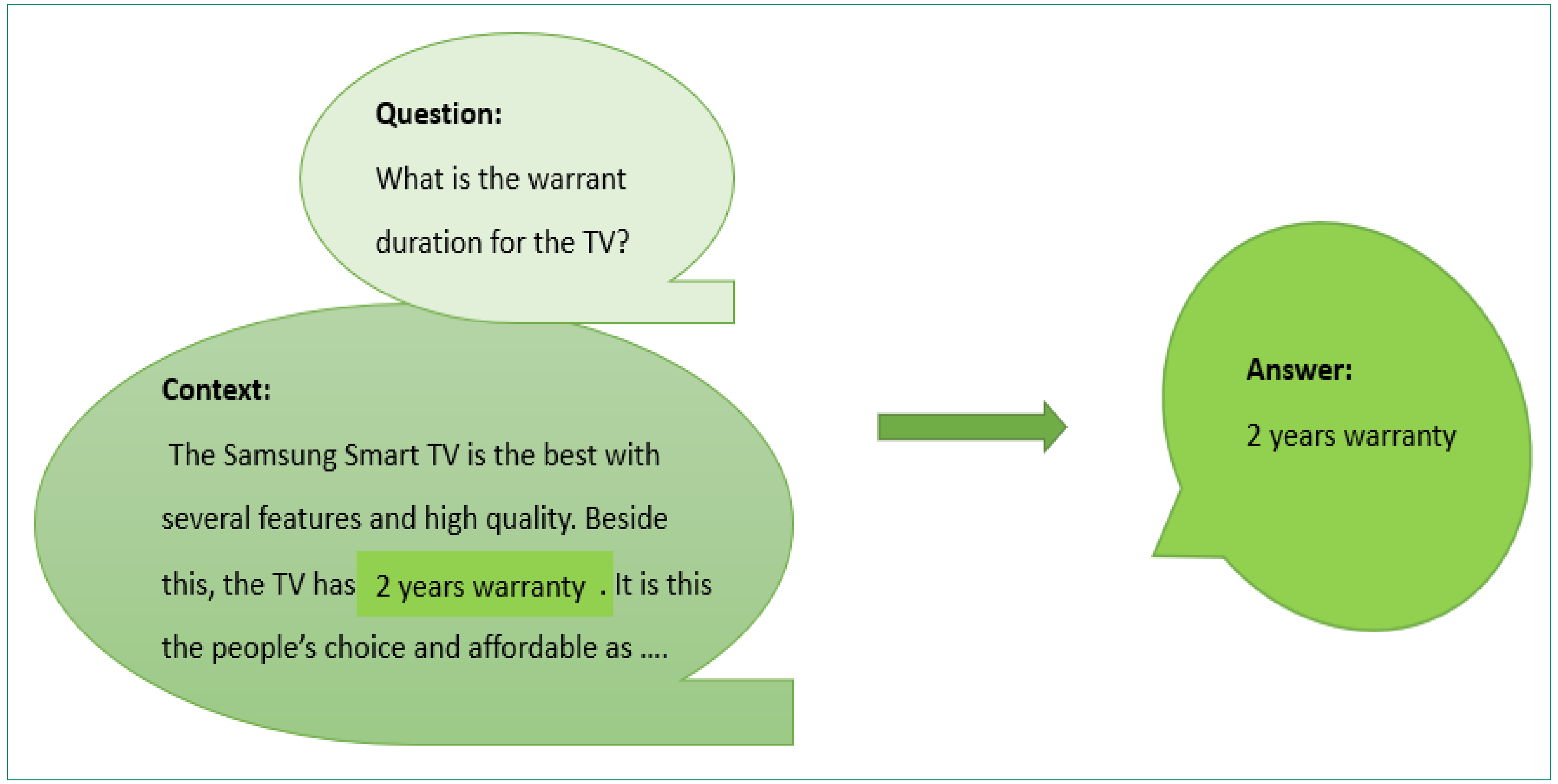
Other Models

Model Name	Paper link
RoBERTa	RoBERTa: A Robustly Optimized BERT Pretraining Approach Yinhan by Liu, et al., arXiv preprint, 2019.
DistilBERT	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter Victor sanh, et al., arXiv, 2019
ProQA	ProQA: Resource-efficient method for pretraining a dense corpus index for open-domain QA and IR. (2020)
GPT-4	GPT-4 Technical Report by OpenAI, 2023
DiffusionBERT	DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models by Zhengfu et al., 2022

Overview of Question Answering Dataset

Question Answering

Introduction



QA Definition

- Question Answering (QA) is about information retrieval whereby a question is posed to the system and a corresponding answer is replied in return.
- The QA system does this by retrieving the answer from a given context such as text or document.

Question Answering

Different Types of QA

- Based on the inputs and output pattern, there are 3 different types of QA:
  - **Extractive QA** - which extracts answers from a text or document referred to as context.
  - **Open Generative QA** - that generates direct text using the context given
  - **Closed Generative QA** - generates answers without any given context
- Our focus would be on Extractive QA including examples of such datasets, and how to build custom datasets for extractive QA

Question Answering Dataset

Stanford Question Answering Dataset (SQuAD)

Extract from SQuAD 2.0 dataset

```
{'qas': [{'question': 'What fraction of New Yorkers in the private sector are employed by foreign companies?',
   'id': '56cf4722aab44d1400b88f06',
   'answers': [{'text': 'One out of ten', 'answer_start': 113}],
   'is_impossible': False},
  {'question': 'What publication ranked New York first in the 2013 American Cities of the Future rankings?',
   'id': '56cf4722aab44d1400b88f07',
   'answers': [{'text': 'FDi Magazine', 'answer_start': 372}],
   'is_impossible': False}],
 'context': 'Many Fortune 500 corporations are headquartered in New York City, as are a large number of foreign corporations. One out of ten private sector jobs in the city is with a foreign company. New York City has been ranked first among cities across the globe in attracting capital, business, and tourists. This ability to attract foreign investment helped New York City top the FDi Magazine American Cities of the Future ranking for 2013.'}
```

```
{
  'version':
    'data': [
              {
                        'title':
                    'paragraphs': [
                                    { 'qas': [ { }, { }, { }, ..., { } ], 'context': }, { 'qas': [ ... ], 'context': }, { } ... { }
                                  ]
              },
              { },
              { },
              .....
              {
                        'title': ,
                    'paragraphs': [
                                    { 'qas': [ { }, { }, { }, ..., { } ], 'context': }, { 'qas': [ ... ], 'context': }, { } ... { }
                                  ]
              }
            ]
}
```

```
{
        'question' :
              'id' : ,
        'answers' : [ {'text': , 'answer_start': } ],
    'is_impossible' :
}
```

Simplified SQUAD JSON Format

- SQuAD is a reading comprehension dataset that contains questions posed by crowd workers on a set of Wikipedia articles.
- These questions are answerable within a text paragraph known as context.
- The data format include:
  - version
  - data
  - title
  - paragraphs
  - qas
  - context
- There are 442 topics/domains and 442 paragraphs covered in the SQuAD json dataset

Question Answering Dataset

Natural Questions (NQ)

Google AI Blog

The latest from Google Research

Natural Questions: a New Corpus and Challenge for Question Answering Research

Wednesday, January 23, 2019

Posted by Tom Kwiatkowski and Michael Collins, Research Scientists, Google AI Language

Open-domain question answering (QA) is a benchmark task in natural language understanding (NLU) that aims to emulate how people look for information, finding answers to questions by reading and understanding entire documents. Given a question expressed in natural language (*"Why is the sky blue?"*), a QA system should be able to read the web (such as this Wikipedia page) and return the correct answer, even if the answer is somewhat complicated and long. However, there are currently no large, publicly available sources of *naturally occurring* questions (i.e. questions asked by a person seeking information) and answers that can be used to train and evaluate QA models. This is because assembling a high-quality dataset for question answering requires a large source of real questions and significant human effort in finding correct answers.

- The Natural Questions is a large-scale corpus dataset from Google that target open-domain question answering system.
- It contains questions issued to Google search engines and long and short answers that were annotated from Wikipedia pages.
- The full dataset is 42GB including HTML of Wikipedia pages, and contains 307k training examples, 8k examples each for testing and development respectively.
- The simplified version of NQ training dataset is 4GB

Question Answering Dataset

Natural Questions (NQ) Format


```
{'document_text': "Email marketing - Wikipedia <H1> Email marketing </H1> Jump to : navigation , search <Table> <Tr> <Td>
</Td> <Td> ( hide ) This article has multiple issues . Please help improve it or discuss these issues on the talk page .
( Learn how and when to remove these template messages ) <Table> <Tr> <Td> </Td> <Td> This article needs additional
citations for verification . Please help improve this article by adding citations to reliable sources . Unsourced
material may be challenged and removed . ( September 2014 ) ( Learn how and when to remove this template message ) </Td>
</Tr> </Table> <Table> <Tr> <Td> </Td> <Td> This article possibly contains original research . Please improve it by
verifying the claims made and adding inline citations . Statements consisting only of original research should be removed
. ( January 2015 ) ( Learn how and when to remove this template message ) </Td> </Tr> </Table> ( Learn how and when to
remove this template message ) </Td> </Tr> </Table> <Table> <Tr> <Td> Part of a series on </Td> </Tr> <Tr> <Th> Internet
marketing </Th> </Tr> <Tr> <Td> <Ul> <Li> Search engine optimization </Li> <Li> Local search engine optimisation </Li>
<Li> Social media marketing </Li>........
This email resulted in $13 million worth of sales in DEC products , and highlighted the potential of marketing through
mass emails . However , as email marketing developed as an effective means of direct communication , users began blocking
out content from emails with filters and blocking programs . In order to effectively communicate a message through email
, marketers had to develop a way of pushing content through to the end user , without being cut out by automatic filters
and spam removing software .....
</Li> <Li> </Li> <Li> </Li> <Li> </Li> <Li> </Li> <Li> </Li> <Li> </Li> </Ul> <Ul> <Li> </Li> <Li> </Li> </Ul>",

'long_answer_candidates': [{'start_token': 14, 'top_level': True, 'end_token': 170}, {'start_token': 15, 'top_level':
False, 'end_token': 169}, {'start_token': 52, 'top_level': False, 'end_token': 103}, {'start_token': 53, 'top_level':
False, 'end_token': 102}, {'start_token': 103, 'top_level': False, 'end_token': 156}, {'start_token': 104, 'top_level':
False, 'end_token': 155}, {'start_token': 170, 'top_level': True, 'end_token': 321}, {'start_token': 171, 'top_level':
False, 'end_token': 180}, {'start_token': 180, 'top_level': False, 'end_token': 186}, {'start_token': 186, 'top_level':
False, 'end_token': 224}, {'start_token': 188, 'top_level': False, 'end_token': 222}, {'start_token': 189, 'top_level':
False,.... }],

'question_text': 'which is the most common use of opt-in e-mail marketing',
'annotations': [{'yes_no_answer': 'NONE', 'long_answer': {'start_token': 1952, 'candidate_index': 54, 'end_token': 2019},
'short_answers': [{'start_token': 1960, 'end_token': 1969}], 'annotation_id': 593165450220027640}],

'document_url': 'https://en.wikipedia.org//w/index.php?title=Email_marketing&amp;oldid=814071202',

'example_id': 5655493461695504401}
```


Extract from NQ Dataset

- Each example of NQ contains:
  - a document paragraph (document_text),
  - long answer candidates (long_answer_candidates),
  - question (question_text),
  - annotations,
  - document_url , and
  - example_id.
- Training examples from the simplified version (v1.0-simplified_simplified-nq-train.jsonl.gz) are shown image on the left-side

Question Answering Dataset

Conversational Question Answering (CoQA)

CoQA: A Conversational Question Answering Challenge

Siva Reddy* **Danqi Chen*** **Christopher D. Manning**
Computer Science Department
Stanford University
{sivar,danqi,manning}@cs.stanford.edu

Abstract

Humans gather information through conversations involving a series of interconnected questions and answers. For machines to assist in information gathering, it is therefore essential to enable them to answer conversational questions. We introduce CoQA, a novel dataset for building **Co**nversational **Q**uestion **A**nswering systems.[1] Our dataset contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. The questions are conversational, and the answers are free-form text with their corresponding evidence highlighted in the passage. We analyze CoQA in depth and show that conversational questions have challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning. We evaluate strong dialogue and reading comprehension models on CoQA. The best system obtains an F1 score of 65.4%, which is 23.4 points behind human performance (88.8%), indicating there is ample room for improvement. We present CoQA as a challenge to the community at https://stanfordnlp.github.io/coqa.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i)

Extract from CoQA paper

- Conversational Question Answering (CoQA) is a large-scale dataset for building conversational question-answering systems.
- The goal is to have a dataset that can measure the ability of machines to comprehend a text passage and correctly respond to a series of interconnected questions within a conversation.

Question Answering Dataset

CoQA Format


```
{'source': 'wikipedia',

 'id': '3zotghdk5ibi9cex97fepx7jetpso7',

 'filename': 'Vatican_Library.txt',

 'story': 'The Vatican Apostolic Library (), more commonly called the Vatican Library or simply the Vat, is the library of the Holy See, located
in Vatican City. Formally established in 1475, although it is much older, it is one of the oldest libraries in the world and contains one of the
most significant collections of historical texts. It has 75,000 codices from throughout history, as well as 1.1 million printed books, which
include some 8,500 incunabula. \n\nThe Vatican Library is a research library for history, law, philosophy, science and theology. The Vatican
Library is open to anyone who can document their qualifications and research needs. Photocopies for private study of pages from books published
between 1801 and 1990 can be requested in person or by mail. \n\nIn March 2014, the Vatican Library began an initial four-year project of
digitising its collection of manuscripts, to be made available online. \n\nThe Vatican Secret Archives were separated from the library at the
beginning of the 17th century; they contain another 150,000 items. \n\nScholars have traditionally divided the history of the library into five
periods, Pre-Lateran, Lateran, Avignon, Pre-Vatican and Vatican. \n\nThe Pre-Lateran period, comprising the initial days of the library, dated
from the earliest days of the Church. Only a handful of volumes survive from this period, though some are very significant.',
 'questions': [{'input_text': 'When was the Vat formally opened?',
   'turn_id': 1},
  {'input_text': 'what is the library for?', 'turn_id': 2},
  {'input_text': 'for what subjects?', 'turn_id': 3},
  {'input_text': 'and?', 'turn_id': 4},
  {'input_text': 'what was started in 2014?', 'turn_id': 5},
  {'input_text': 'how do scholars divide the library?', 'turn_id': 6},
  ...........................................................................
  {'input_text': 'what will this allow?', 'turn_id': 20}],

 'answers': [{'span_start': 151,
   'span_end': 179,
   'span_text': 'Formally established in 1475',
   'input_text': 'It was formally established in 1475',
   'turn_id': 1},
  {'span_start': 454,
   'span_end': 494,
   'span_text': 'he Vatican Library is a research library',
   'input_text': 'research',
   'turn_id': 2},
  {'span_start': 457,
   'span_end': 511,
   'span_text': 'Vatican Library is a research library for history, law',
   'input_text': 'history, and law',
   'turn_id': 3},
  {'span_start': 457,
   'span_end': 545,
   'span_text': 'Vatican Library is a research library for history, law, philosophy, science and theology',
   'input_text': 'philosophy, science and theology',
   'turn_id': 4},
 {'span_start': 769,
  'span_end': 879,
  'span_text': 'March 2014, the Vatican Library began an initial four-year project of digitising its collection of manuscripts',
  'input_text': 'a  project',
  'turn_id': 5},
 {'span_start': 1048,
  'span_end': 1127,
  'span_text': 'Scholars have traditionally divided the history of the library into five period',
  'input_text': 'into periods',
  'turn_id': 6},
.............................................................................................
 {'span_start': 868,
  'span_end': 910,
  'span_text': 'manuscripts, to be made available online. ',
  'input_text': 'them to be viewed online.',
  'turn_id': 20}],

'name': 'Vatican_Library.txt'}
```


- CoQA data format contains:
  - Source
  - Id
  - Filename
  - Story
  - Questions
  - Answers
  - Name

Question Answering Dataset

Other Datasets

QA Dataset Name	Download Link	Paper link
Explain Like I'm Five (ELI5)	https://github.com/facebookresearch/ELI5	Long Form Question Answering
TriviaQA	http://nlp.cs.washington.edu/triviaqa/	TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension:
Question Answering in Context (QuAC)	https://quac.ai/	Question Answering in Context:
TWEETQA	https://aclanthology.org/P19-1496/	TWEETQA: A Social Media Focused Question Answering Dataset:

- For more on large and small Question Answering dataset, see:
  - *10 Question-Answering Datasets To Build Robust Chatbot System* by Ambika Choudhury, 2019 and
  - *University of Freiburg: Algorithms and Data Structures Group* large-qa-datasets GitHub page .

Q & A

THANK YOU