Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

Natural Language Processing with Deep Learning

Bài 11: MÔ HÌNH OPENAI GPT

https://csc.edu.vn/data-science-machine-learning/natural-language-processing-with-deep-learning_293

2023

MÔ HÌNH **OPENAI GPT**

I. Masked Self-Attention

II. Language Modeling

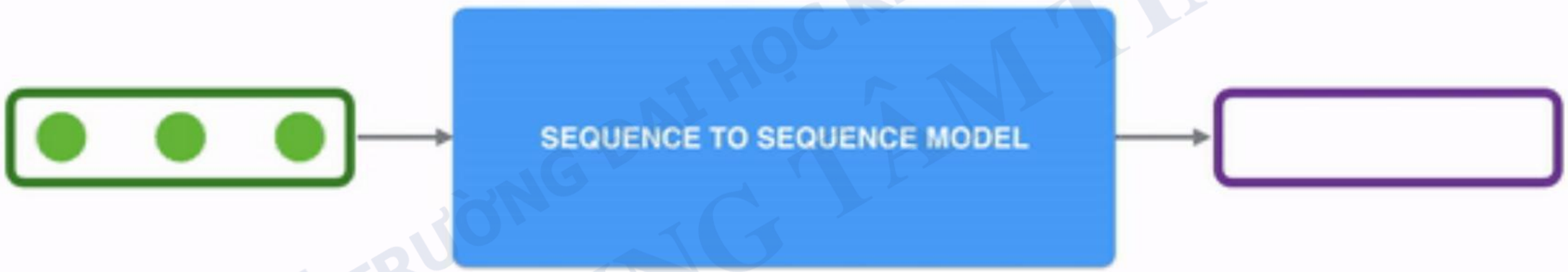
III. GPT-1

IV. GPT-2

V. GPT-3

Tổng quan Masked-Self Attention

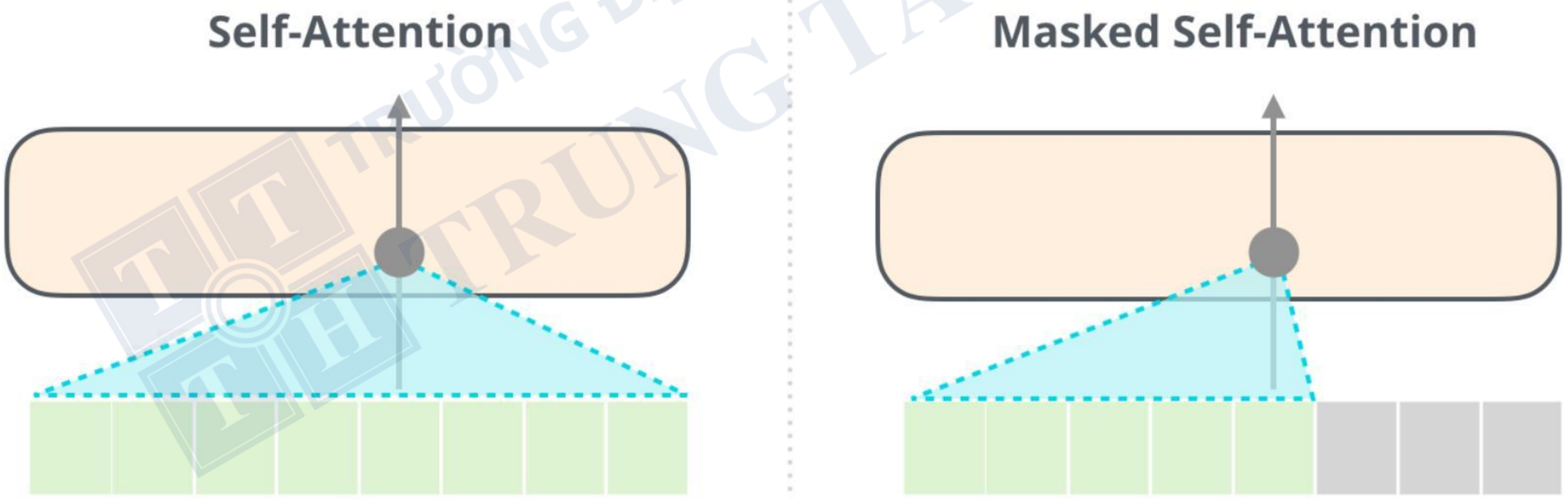
Self-attention là cơ chế phân chia tỉ trọng theo mức độ quan trọng của các từ khác nhau trong một chuỗi.



Masked-Self Attention dự đoán từ tiếp theo trong một chuỗi bằng cách xem xét tất cả các từ quan trọng trước đó.

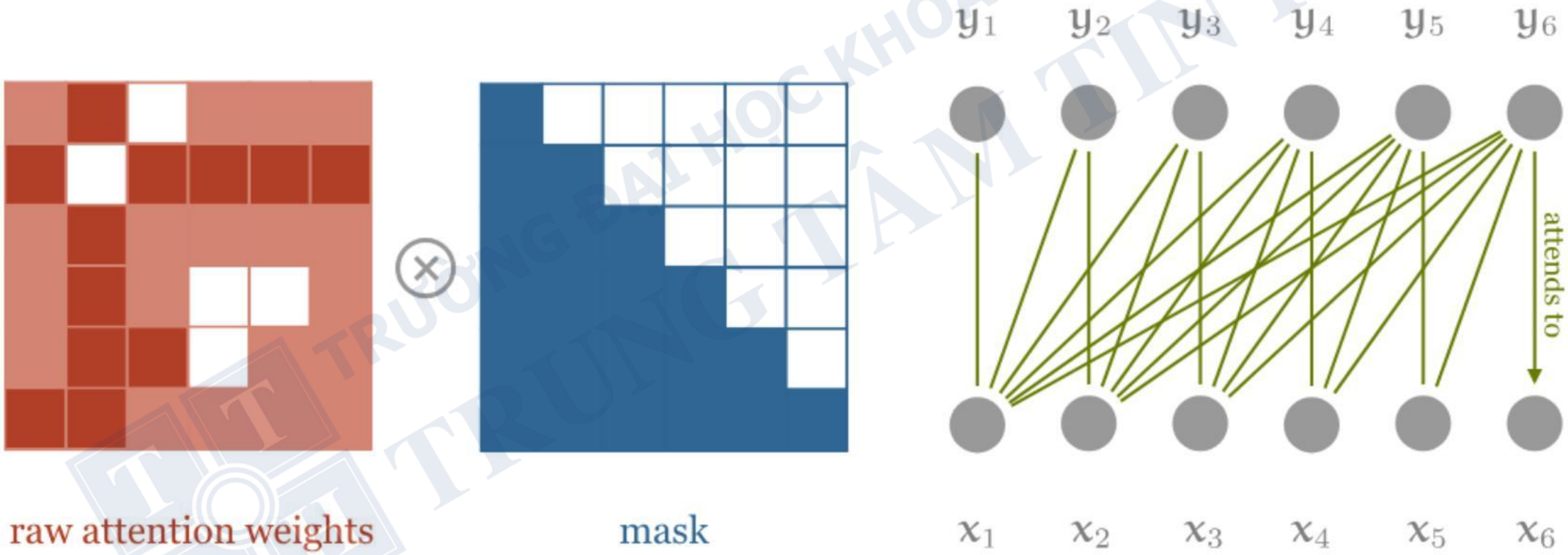
Tổng quan Masked-Self Attention

Masked-Self Attention thay thế các từ không quan trọng trong input bằng các **"mask"** token.
→ Mô hình tập trung vào ngữ cảnh liên quan và dự đoán các từ bị thiếu dựa trên các thông tin quan trọng.

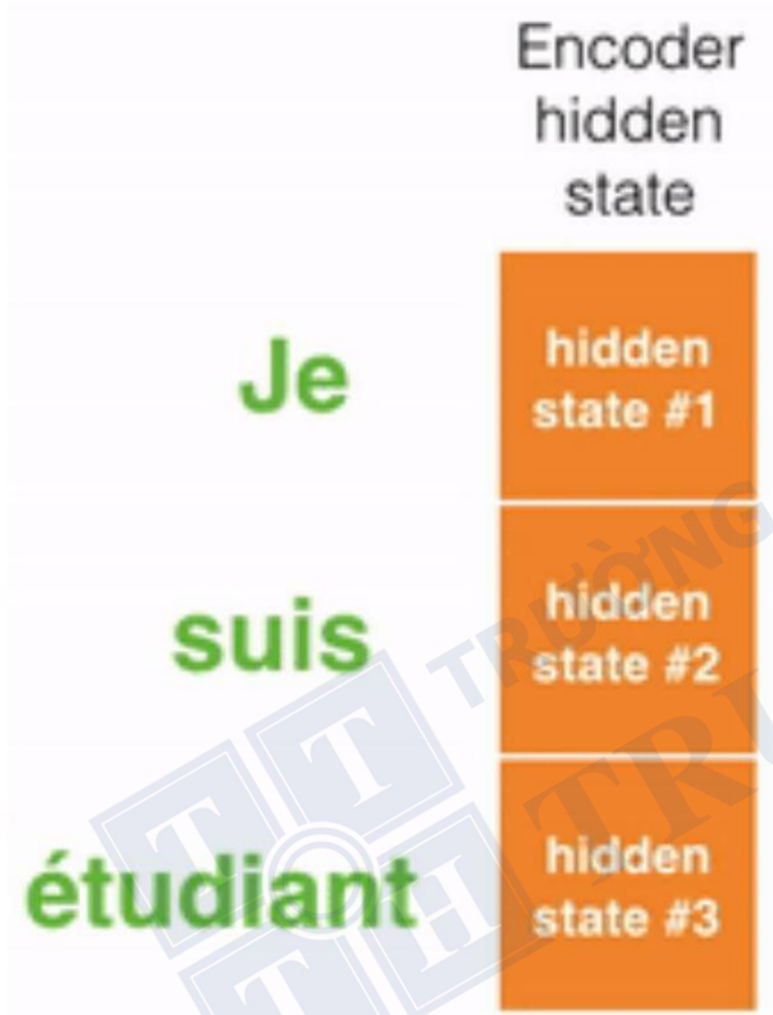


Tổng quan Masked-Self Attention

Masked-Self Attention thay thế các từ không quan trọng trong input bằng các **"mask"** token.



Tổng quan Masked-Self Attention



Tổng quan Masked-Self Attention

Tổng quan Masked-Self Attention

Query, key và value đến từ cùng một câu.

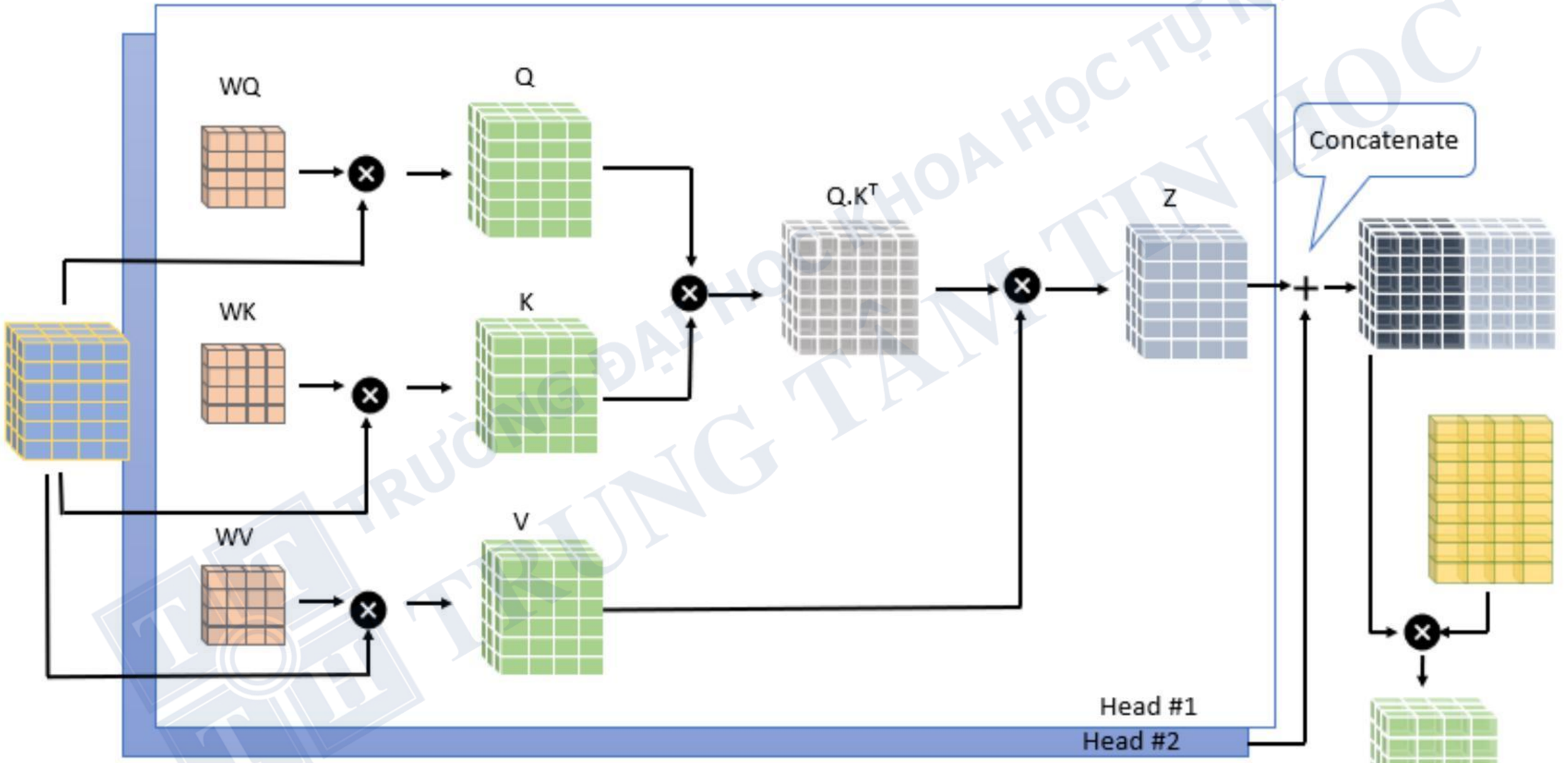
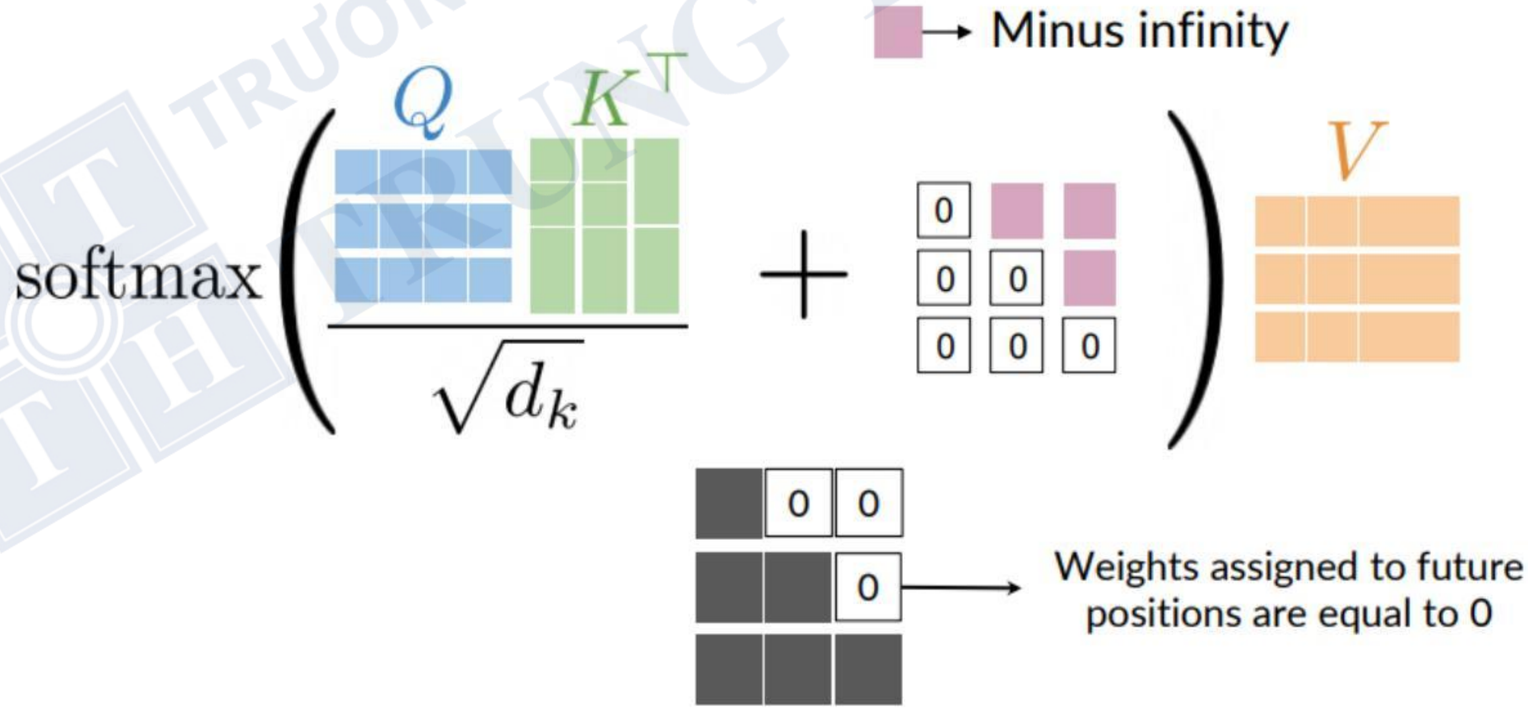


Fig 15: All steps involved in Multi-head self attention with number of heads = 2

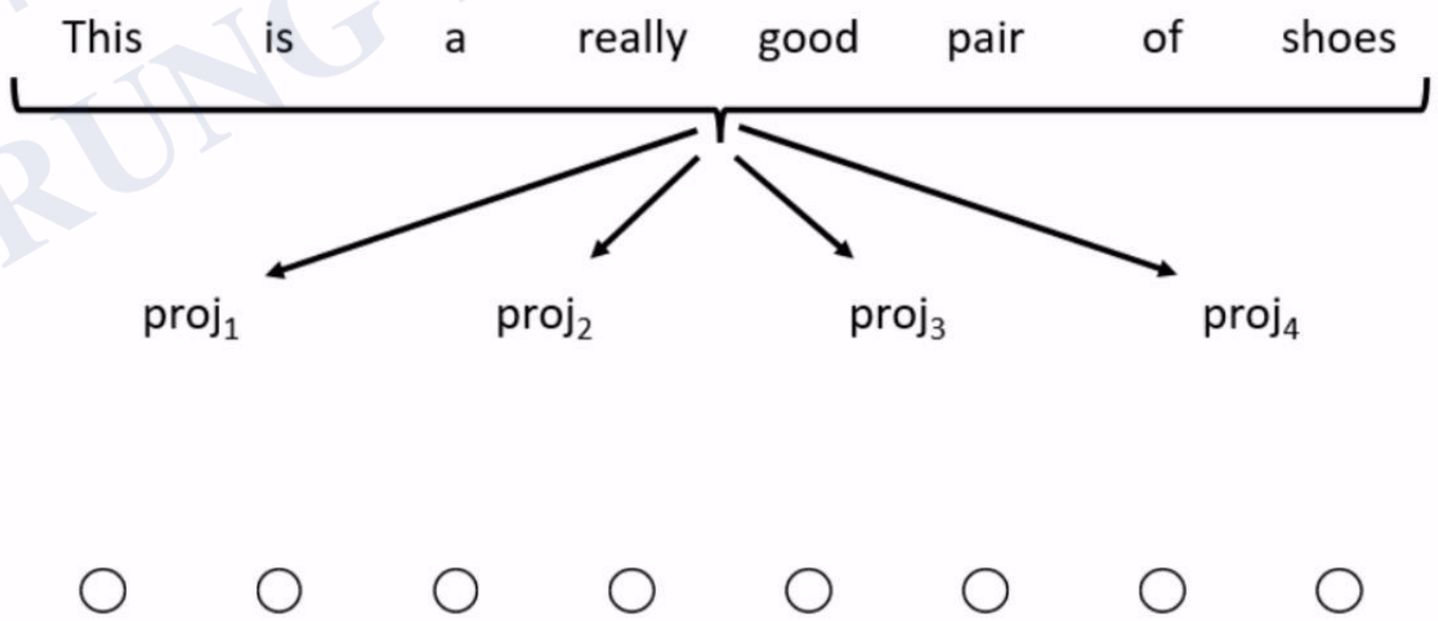
Masked Self-Attention

Weight matrix



Masked Self-Attention

- **Có 3 loại Attention: Encoder/decoder, Self-attention và Masked Self-attention.**
  - Input: Query, key, value.
  - Query và key được lấy từ 1 câu.
  - Masked self-attention không thể tham dự vào tương lai.



MÔ HÌNH **OPENAI GPT**

I. Masked Self-Attention

II. Language Modeling

III. GPT-1

IV. GPT-2

V. GPT-3

Language Modeling

là quá trình dự đoán từ tiếp theo trong một chuỗi ngôn ngữ dựa trên thông tin từ các từ trước đó.

1. **Auto-regressive**

   Language Modeling

2. **Auto-encoding**

   Language Modeling

Auto-Regressive Language Modeling

Dự đoán từ tiếp theo trong một chuỗi theo tuần tự sử dụng xác suất.

- **Forward** (trái → phải) **prediction**

Paris is a beautiful __.

- **Backward** (phải → trái) **prediction**

__. I love Paris

Auto-Encoder Language Modeling

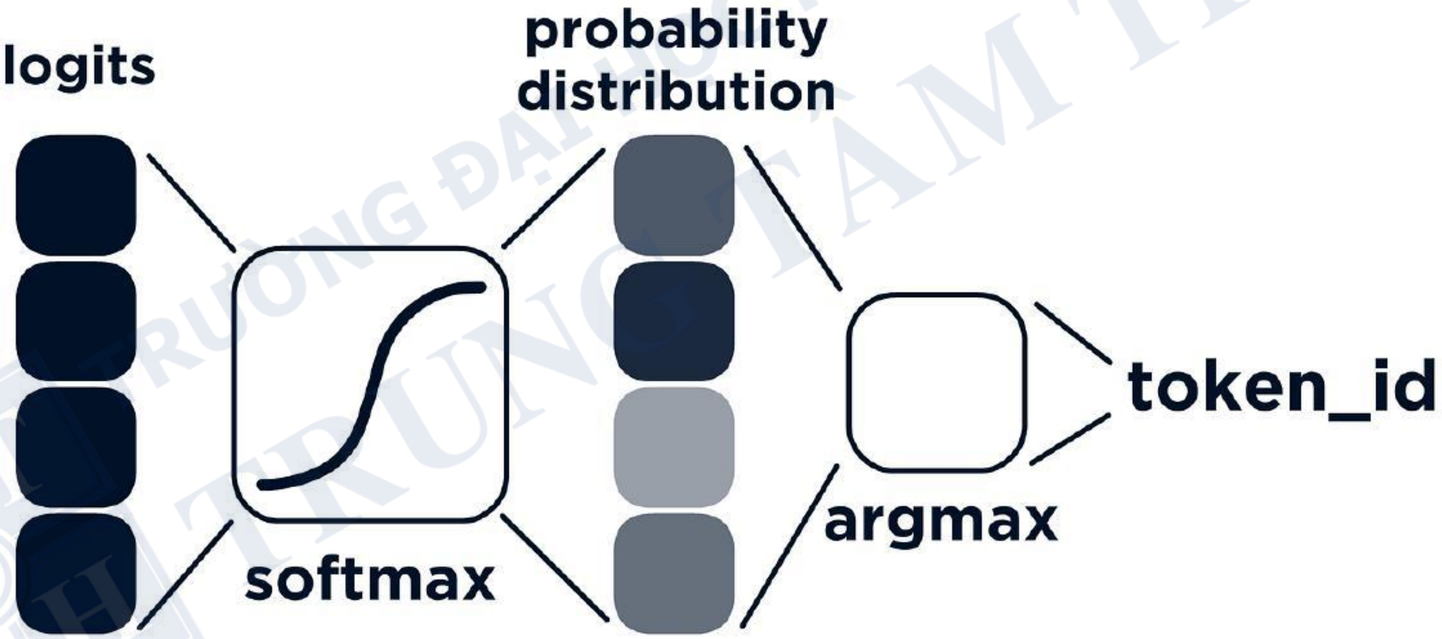
Mô hình được train để tái tạo lại input ban đầu thông qua encoding và decoding dữ liệu ngôn ngữ.

Paris is a beautiful __. I love Paris

➔ Tạo ra một **language model** có khả năng tự động sinh ra các câu mới.

Masked Language Modeling

Là mô hình ngôn ngữ với một số từ trong câu được ẩn đi (*mask*), được train để dự đoán các từ bị che này dựa trên ngữ cảnh của câu.



➔ Tạo ra một **language model** có khả năng hiểu và sinh ra các từ phù hợp còn thiếu trong câu.

Masked Language Modeling

Masked language model sử dụng **mask** token.

Masked Language Modeling

Ví dụ **mask token**

Replaced Token Detection

the chef cooked the meal

MÔ HÌNH **OPENAI GPT**

I. Masked Self-Attention

II. Language Modeling

III. GPT-1

IV. GPT-2

V. GPT-3

GPT-1 Model

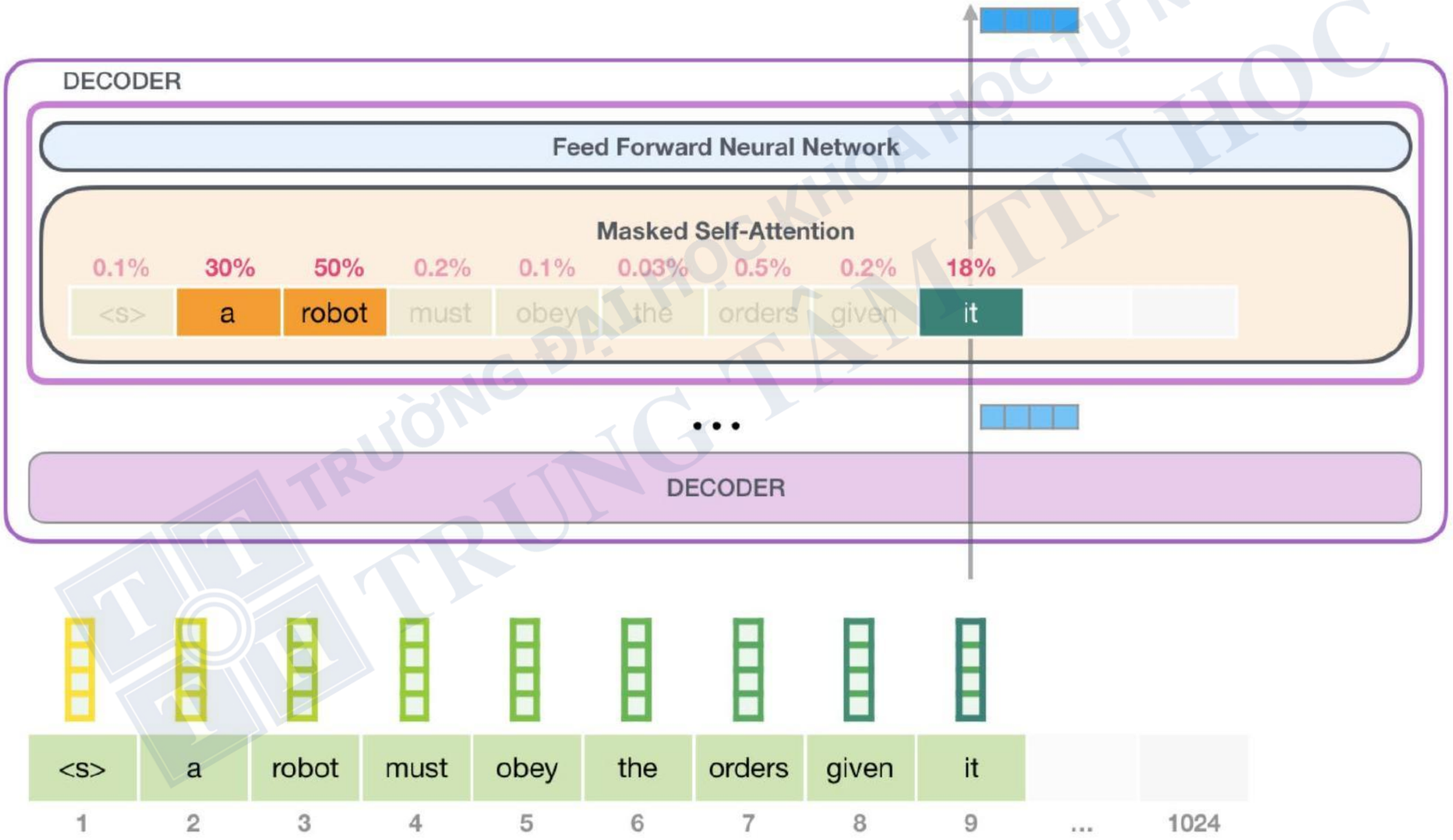
Generative Pre-trained Transformer 1 là một transformer model được train trên một lượng lớn dữ liệu ngôn ngữ tự nhiên.

- Stack A Bunch Transformer Decoders
- Semi-Supervised Pretraining
- Finetune trên nhiều tác vụ khác nhau

➔ Có thể tự động sinh ra văn bản, hiểu các mối quan hệ ngữ nghĩa trong câu. Tiến bộ hơn đáng kể trong nhiều tác vụ xử lý ngôn ngữ tự nhiên.

GPT-1 Model

1. Stack A Bunch Transformer Decoders



GPT-1 Model

1. Stack A Bunch Transformer Decoders

	Features				Labels
	position: 1	2	3	4	
Example: 1	robot	must	obey	orders	must
2	robot	must	obey	orders	obey
3	robot	must	obey	orders	orders
4	robot	must	obey	orders	<eos>

1. Stack A Bunch Transformer Decoders

Queries

robot	must	obey	orders

X

Keys

robot	must	obey	orders
robot	must	obey	orders
robot	must	obey	orders
robot	must	obey	orders

=

Scores (before softmax)

0.11	0.00	0.81	0.79
0.19	0.50	0.30	0.48
0.53	0.98	0.95	0.14
0.81	0.86	0.38	0.90

Scores (before softmax)

0.11	0.00	0.81	0.79
0.19	0.50	0.30	0.48
0.53	0.98	0.95	0.14
0.81	0.86	0.38	0.90

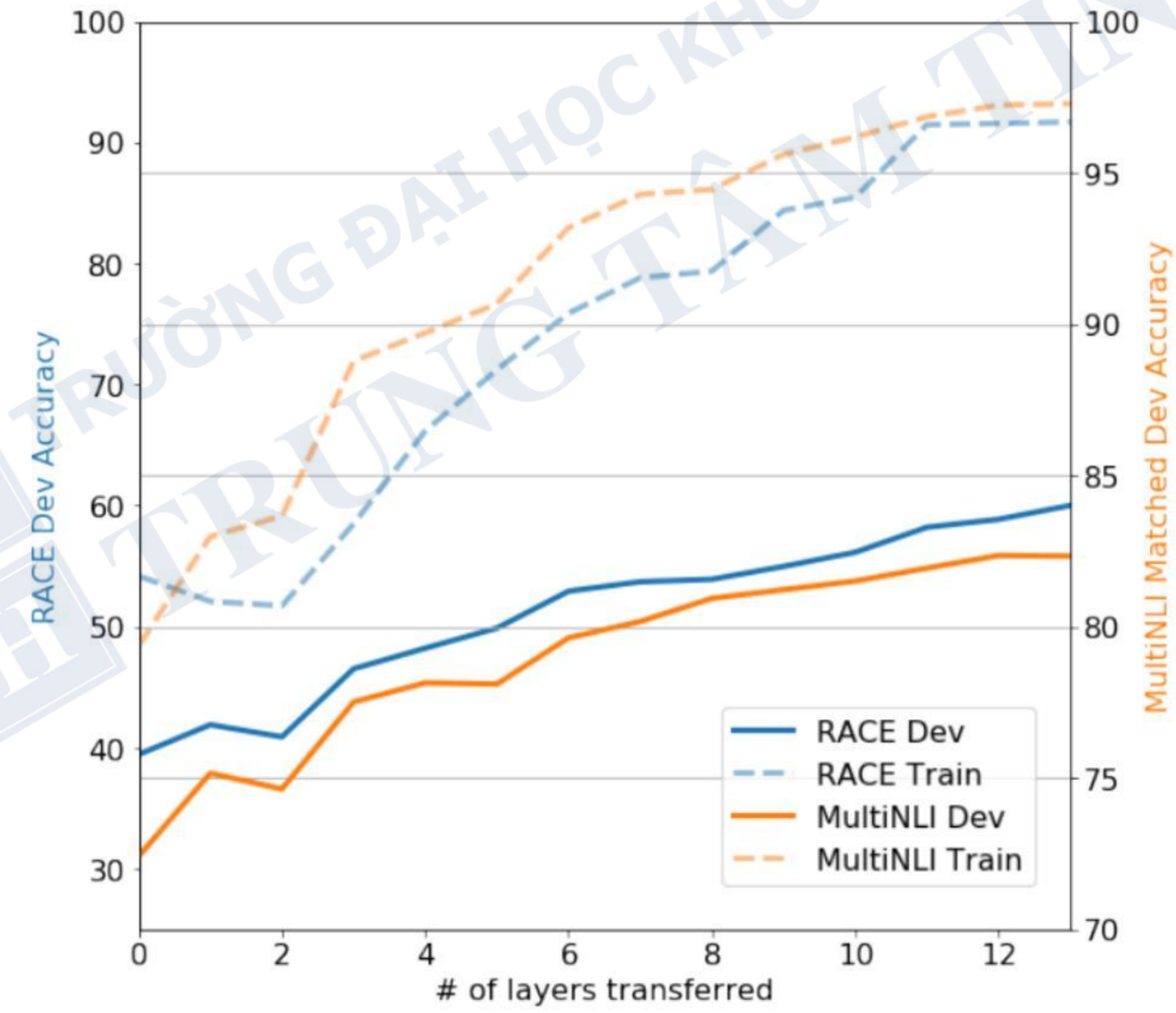
Apply Attention Mask →

Masked Scores (before softmax)

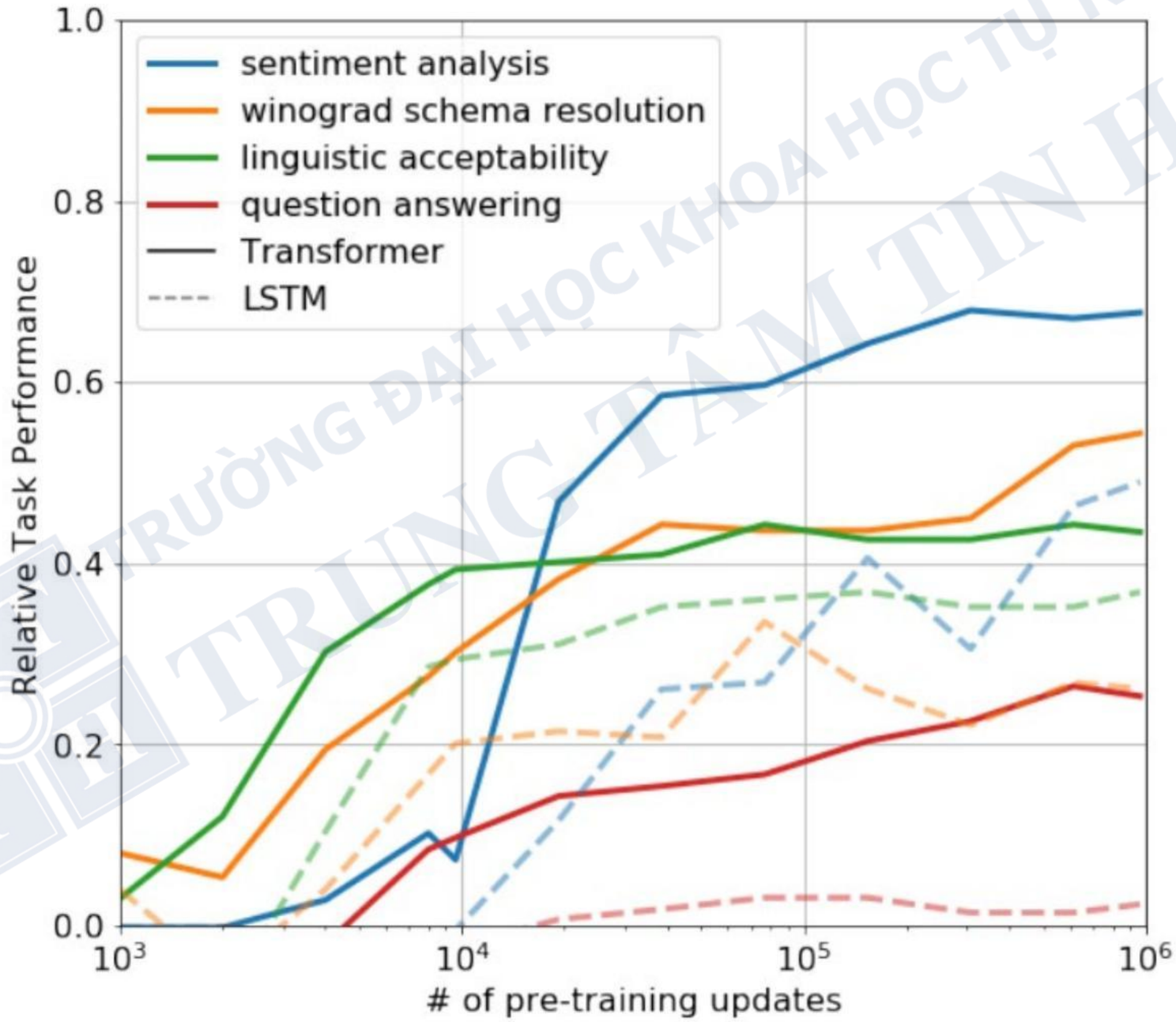
0.11	-inf	-inf	-inf
0.19	0.50	-inf	-inf
0.53	0.98	0.95	-inf
0.81	0.86	0.38	0.90

2. Semi-Supervised Pretraining

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

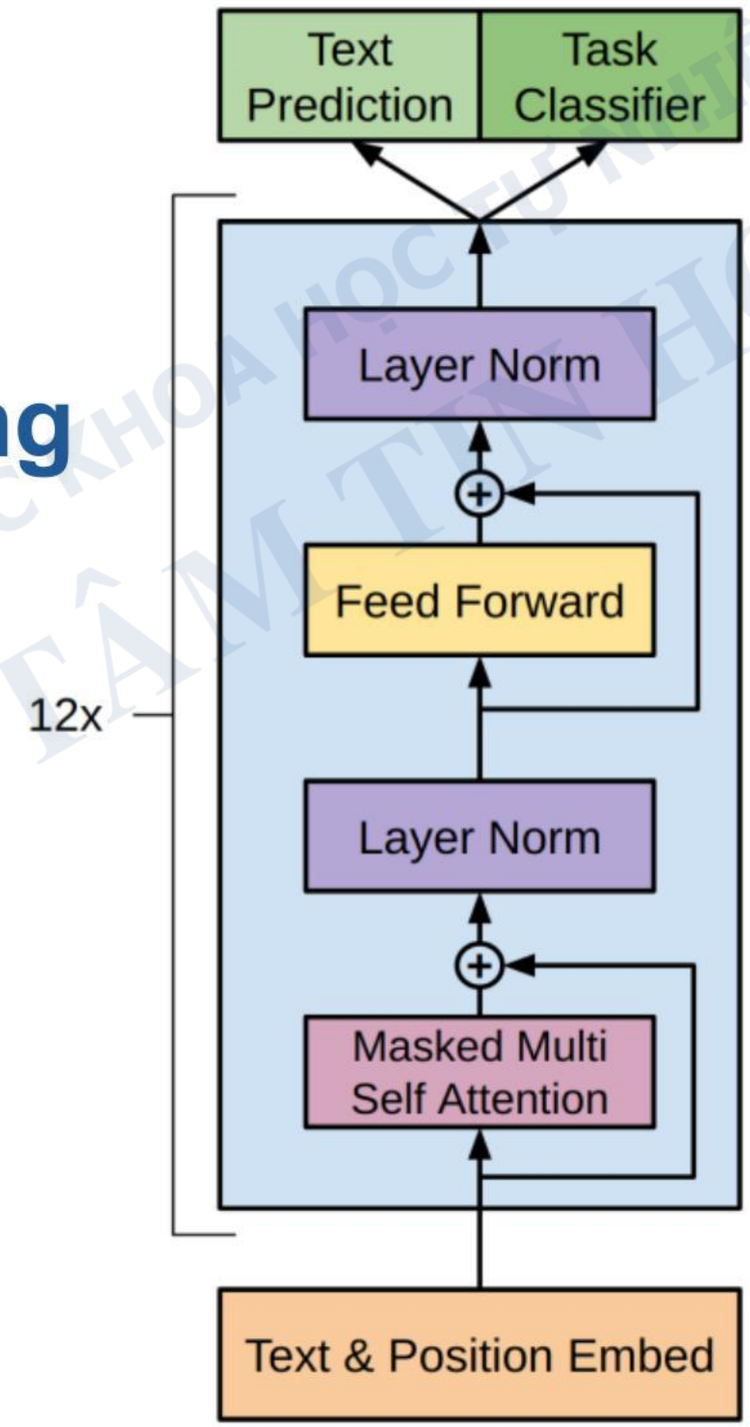


2. Semi-Supervised Pretraining



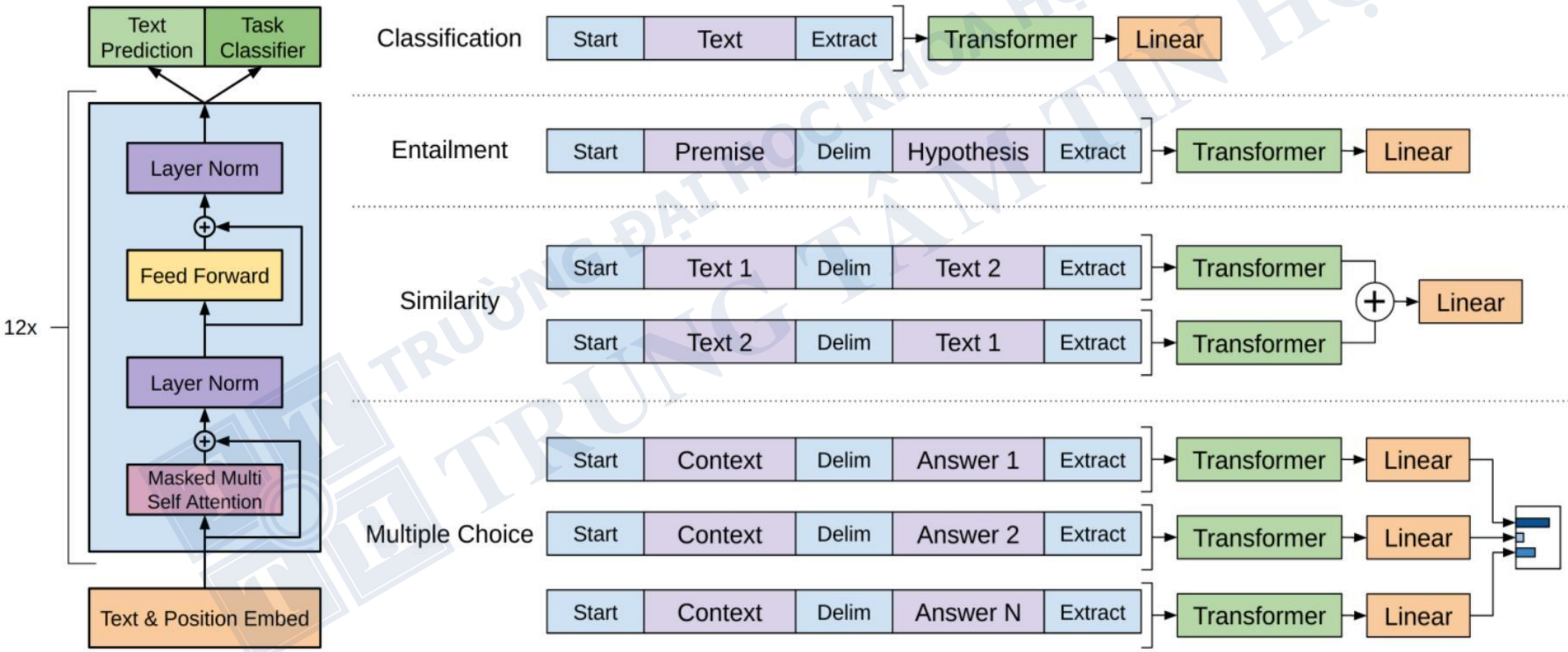
GPT-1 Model

3. Multiple tasks **finetuning**



GPT-1 Model

3. Multiple tasks **finetuning**



MÔ HÌNH **OPENAI GPT**

I. Masked Self-Attention

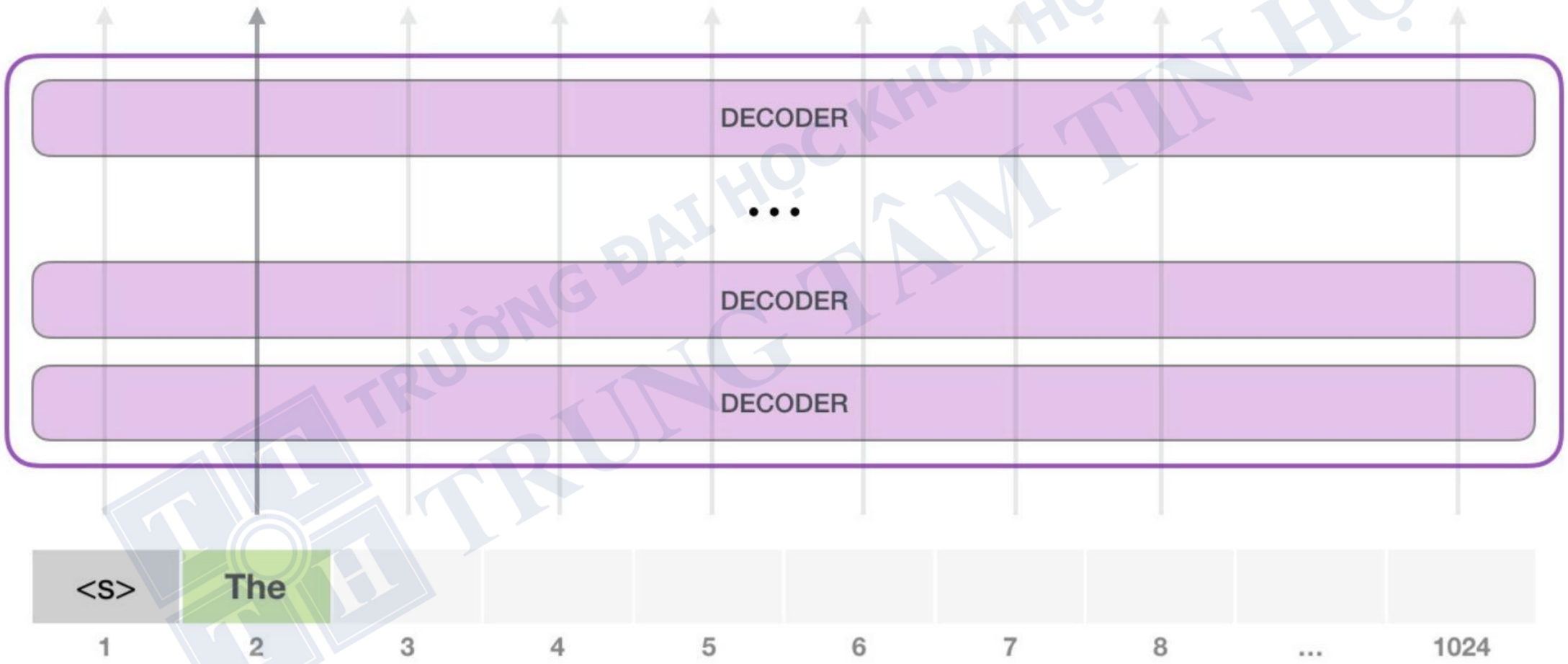
II. Language Modeling

III. GPT-1

IV. GPT-2

V. GPT-3

Các loại **Decoder**



- Greedy Decoding
- Random Sampling
- Beam Search

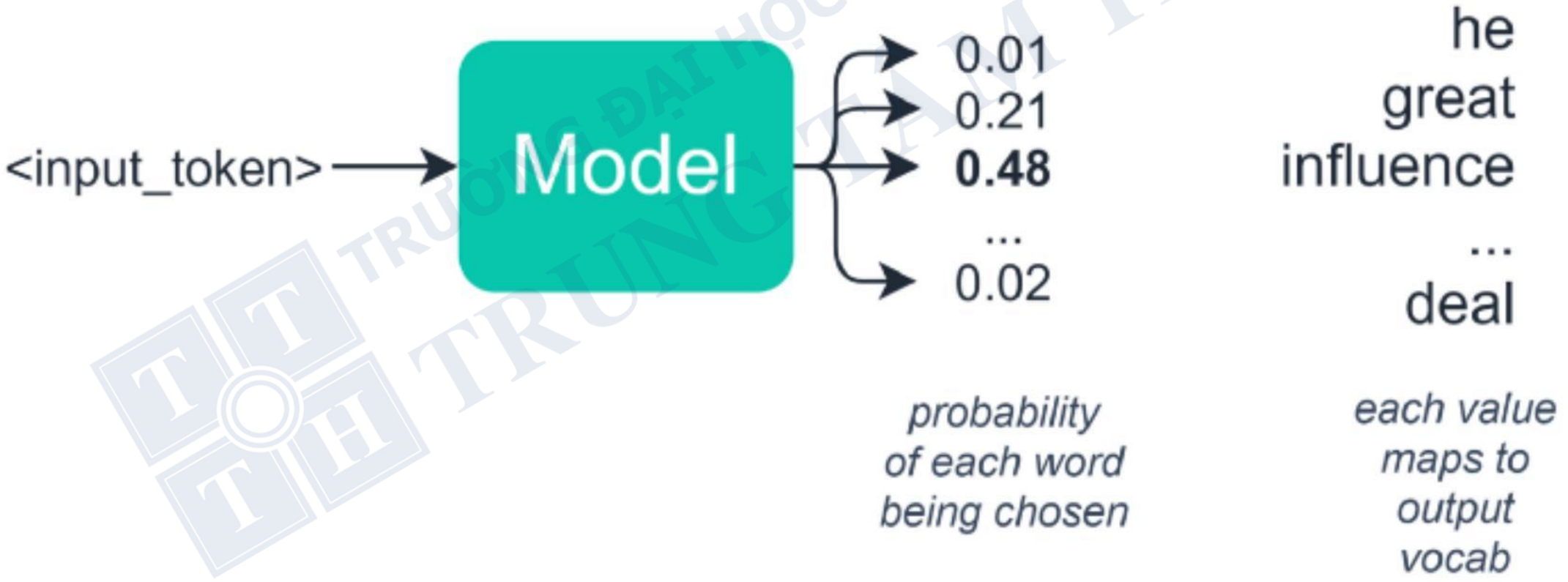
Kiến thức bổ trợ

1. **Greedy** Decoder

```
He began his premiership by forming a five-man war cabinet which included
Chamerlain as Lord President of the Council, Labour leader Clement Attlee as
Lord Privy Seal (later as Deputy Prime Minister), Halifax as Foreign Secretary
and Labour's Arthur Greenwood as a minister without portfolio. In practice,
+ the cabinet was divided into three parts: the Cabinet of Ministers, the
+ Cabinet of Ministers of the Crown, and the Cabinet of Ministers of the Crown.
+ The Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers was the most important part of the government. The
+ Cabinet of Ministers...
```

2. Random sampling

Chọn ngẫu nhiên một từ dựa trên xác suất của nó được mô hình chỉ định trước đó.



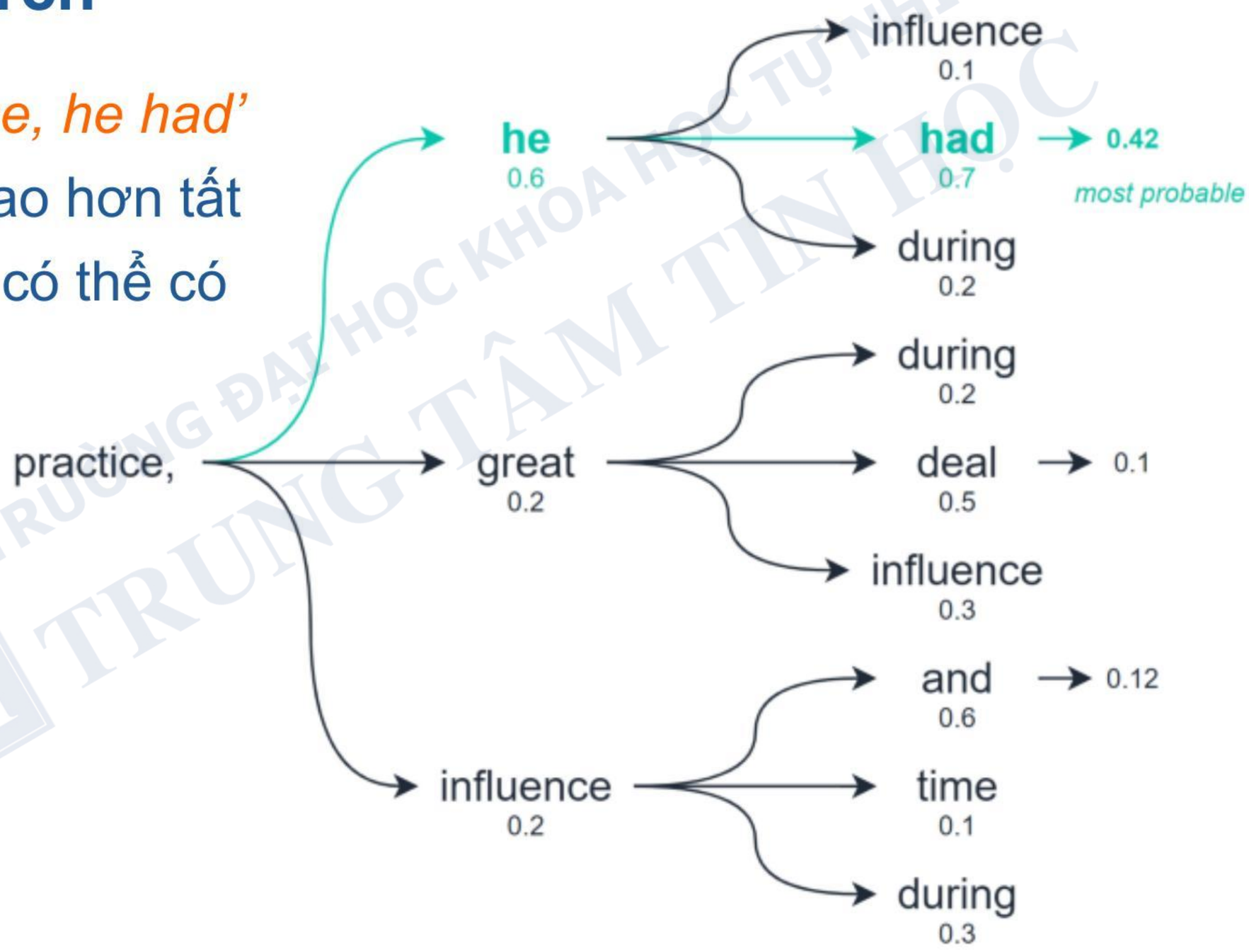
2. Random sampling

```
He began his premiership by forming a five-man war cabinet which included
Chamerlain as Lord President of the Council, Labour leader Clement Attlee as
Lord Privy Seal (later as Deputy Prime Minister), Halifax as Foreign Secretary
and Labour's Arthur Greenwood as a minister without portfolio. In practice,
+ vernacular spelling errors and short comments, with a particular emphasis
+ on Chameleon and the French words 'vie', were common place in the Prime
+ Minister's office during his reign.
+
+ In the 1980s after Chameleon became Prime Minister, more than 40,000 copies
+ of this booklet, including many containing Chameleon memorabilia, disappeared
+ along with countless copies of the game. In 1992, a new edition of The
+ Chameleon Paradox was published. It was the most popular title of the book,
+ a compilation from 1987 to 1994 which included a selection of new games,
+ such as R&R and The Bigger Picture, including some original games, including
+ the original R&...
```

Kiến thức bổ trợ

3. Beam search

Chuỗi *'practice, he had'* có xác suất cao hơn tất cả các chuỗi có thể có



3. Beam search

```
He began his premiership by forming a five-man war cabinet which included
Chamerlain as Lord President of the Council, Labour leader Clement Attlee as
Lord Privy Seal (later as Deputy Prime Minister), Halifax as Foreign Secretary
and Labour's Arthur Greenwood as a minister without portfolio. In practice,
+ vernacular terms such as "prime minister" and "foreign secretary" were used
+ to describe him.\n\nIn February 1967, he was appointed as prime minister by
+ Malcolm Fraser.\n\nAt the time of his appointment, vernacular terms such as
+ "prime minister" and "foreign secretary" were used to describe him.
+
+ In July 1967, vernacular terms such as "prime minister" and "foreign
+ secretary" were used to describe him.
+
+ In September 1967, vernacular terms such as "prime minister" and "foreign
+ secretary" were used to describe him.
+
+ In September 1967, vernacular terms such as "prime minister" and "foreign...
```

Kiến thức bổ trợ

Perplexity là một thước đo được sử dụng để đánh giá mức độ tốt của mô hình ngôn ngữ qua: Test set inverse probability và Cross-entropy.

1. Inverse probability: $$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Test Set

"Yesterday I went to the cinema"

"Hello, how are you?"

"The dog was wagging its tail"

High probability
Low perplexity

Fake/incorrect sentences

"Can you does it?"

"For wall a driving"

"She said me this"

Low probability
High perplexity

2. Cross-Entropy

Là số bit trung bình cần thiết để mã hóa một từ.

Perplexity là số từ có thể được mã hóa bằng số bit đó:

$$PP(W) = 2^{H(W)} = 2^{-\frac{1}{N}\log_2 P(w_1, w_2, \ldots, w_N)}$$

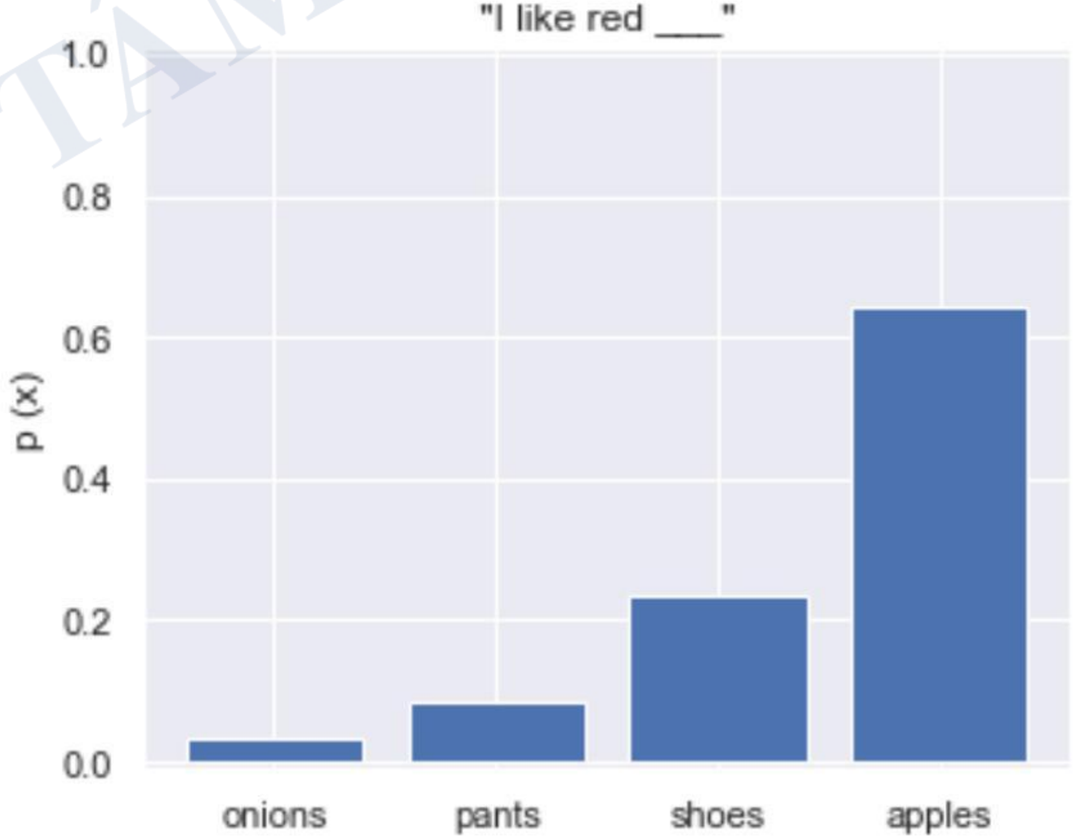
➔ Perplexity là 100 có nghĩa là bất cứ khi nào mô hình cố đoán từ tiếp theo, nó sẽ phải chọn **giữa 100 từ**.

Kiến thức bổ trợ

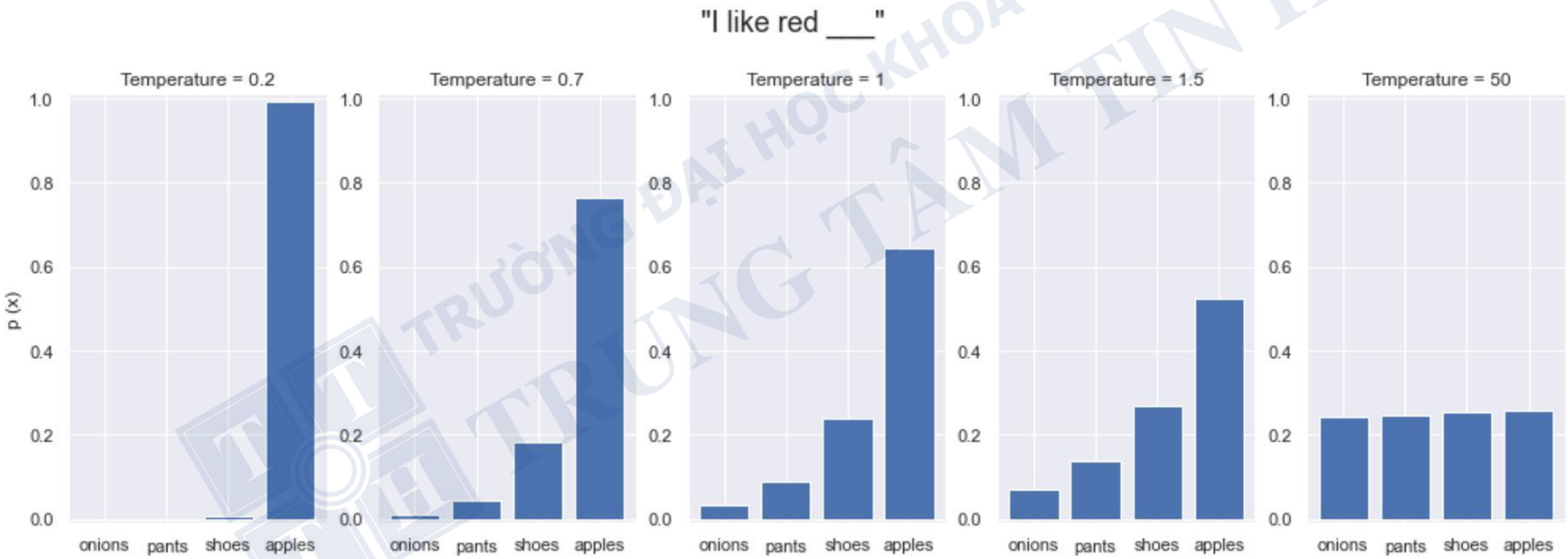
Temperature là một công cụ điều chỉnh phân bố xác suất của các từ.

$$p_i = \frac{1}{Q} e^{-\varepsilon_i/(kT)} = \frac{e^{-\varepsilon_i/(kT)}}{\sum_{j=1}^{M} e^{-\varepsilon_j/(kT)}}$$

➔ Temperature thấp thì mô hình có tính quyết định cao và temperature cao thì mô hình có tính quyết định ít hơn.

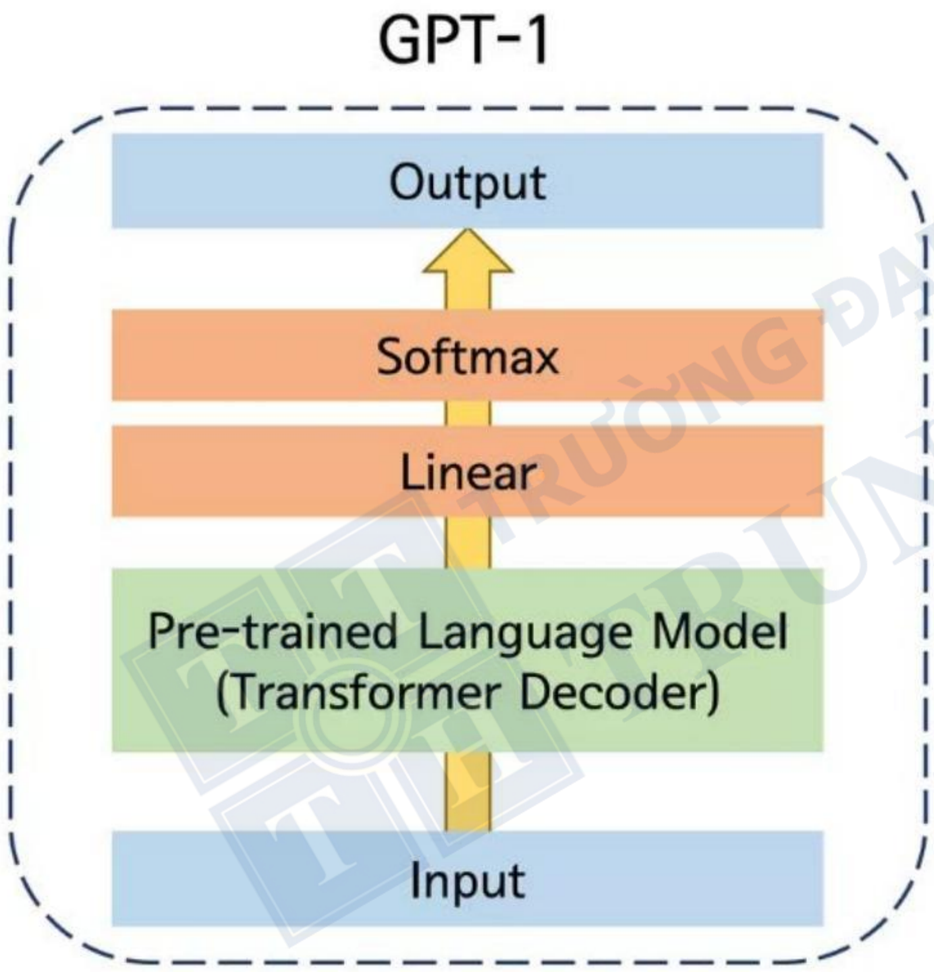


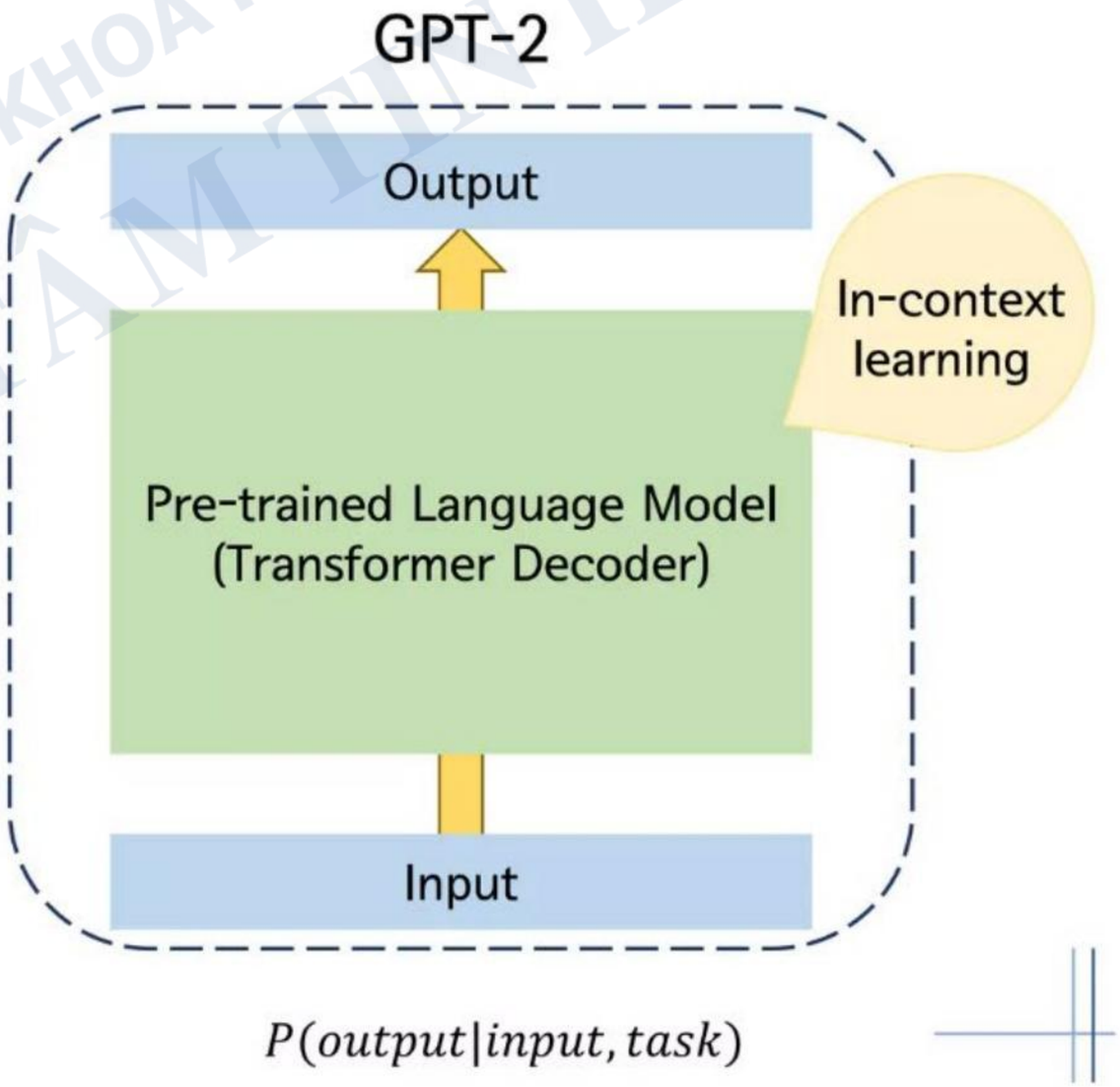
Ví dụ về **Temperature**



GPT-2 Model

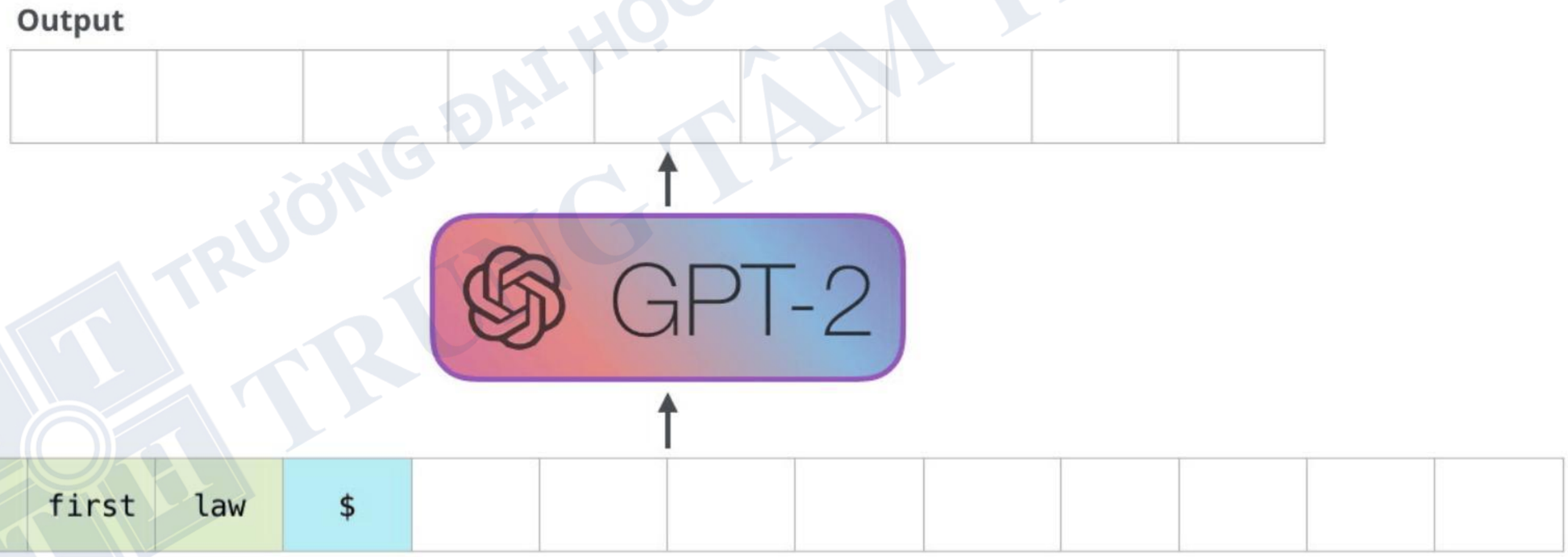
- Kích thước lớn hơn so với GPT-1
- **Zero-shot NLP.**



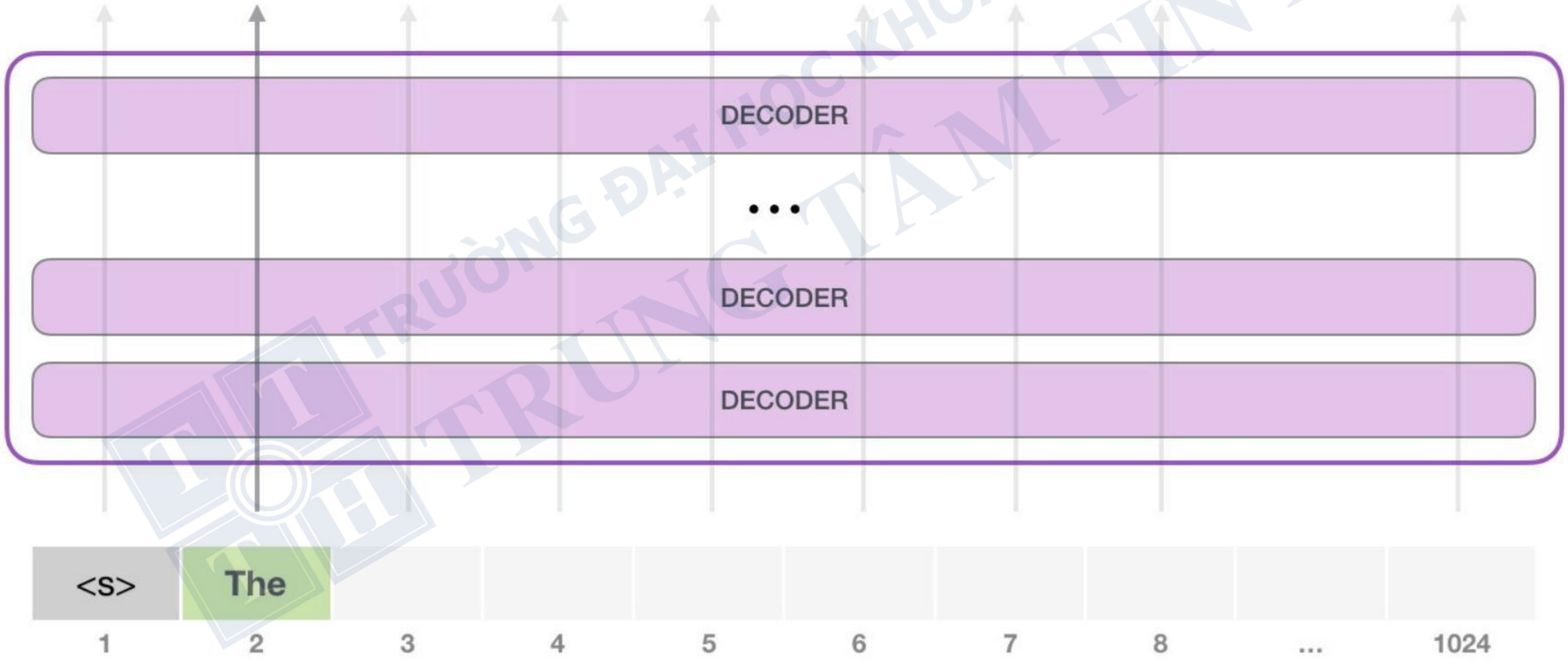


GPT-2 Model

- Kích thước lớn hơn so với GPT-1
- **Zero-shot NLP.**



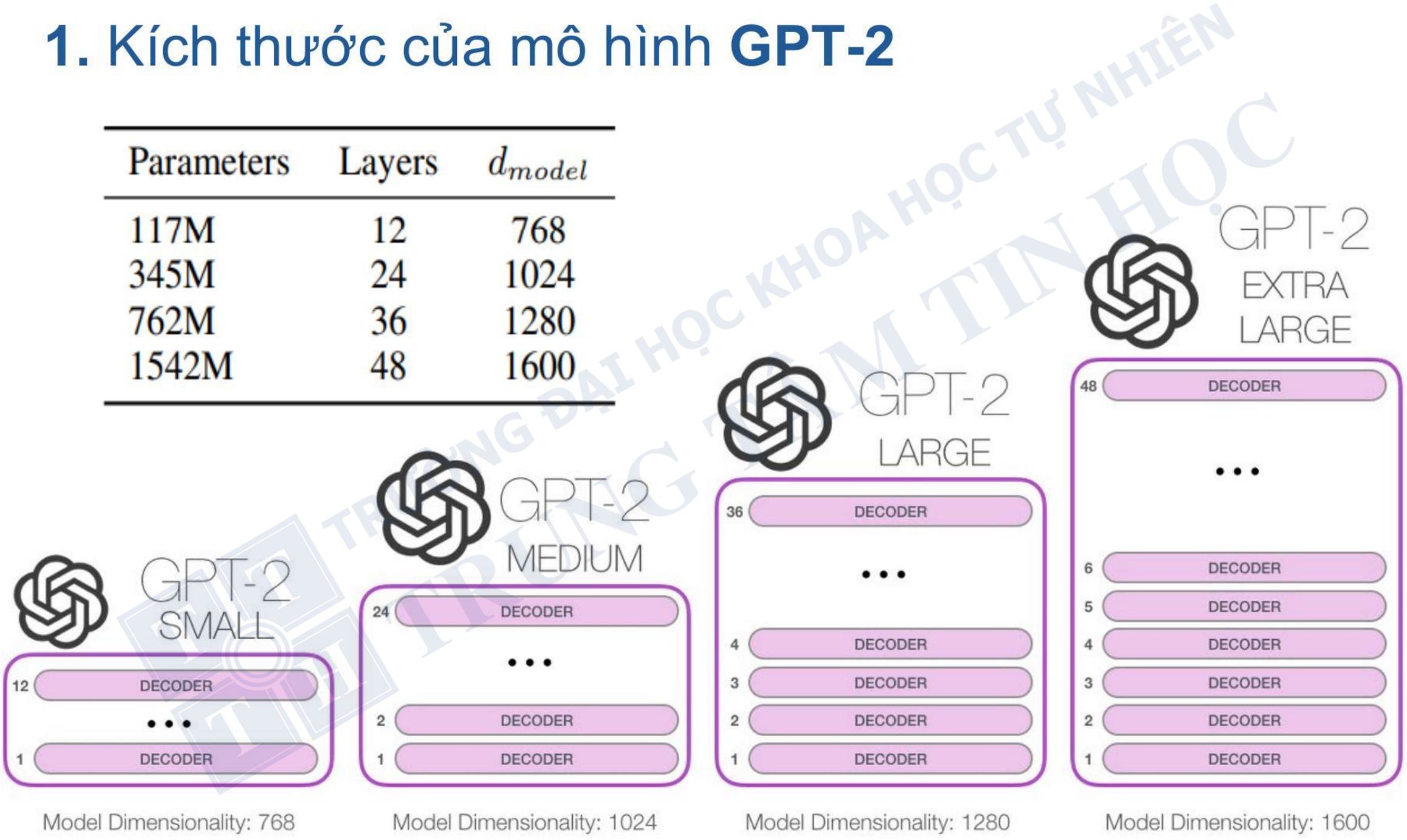
GPT-2 encoder – decoder



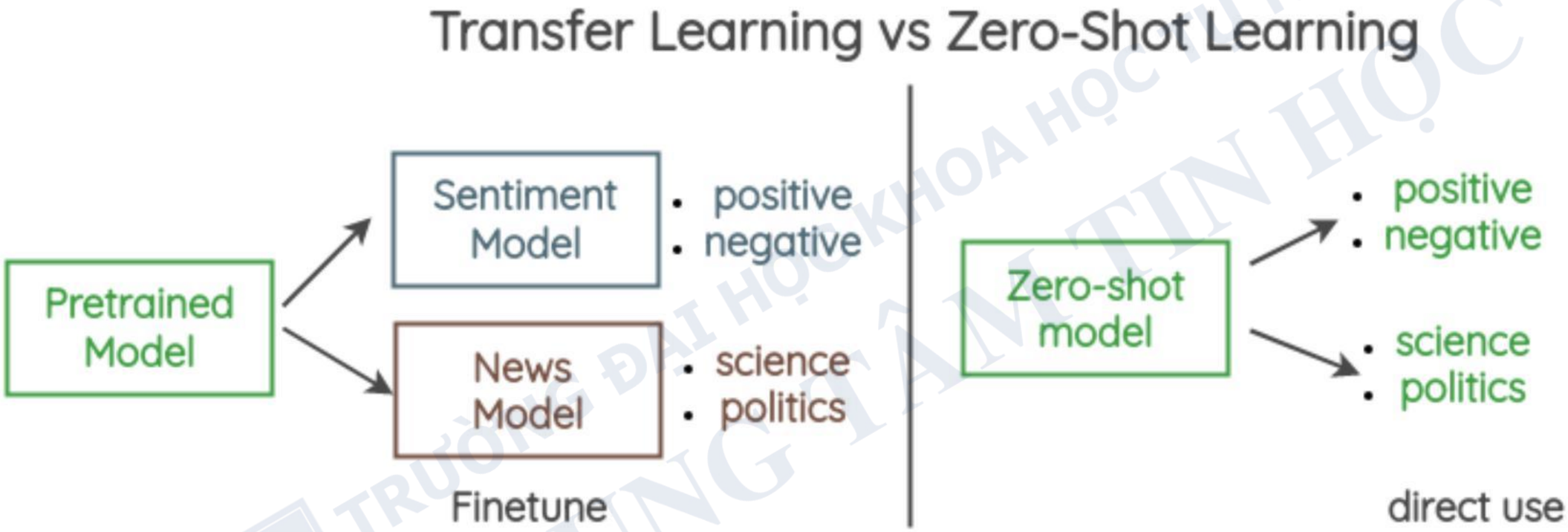
GPT-2 Model

1. Kích thước của mô hình **GPT-2**

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600



2. Zero-shot NLP



Mô hình được train để thực hiện các tác vụ ngôn ngữ chưa được train trước đó.

➔ Khả năng xử lý và hiểu các tác vụ mới mà **không cần train** cho từng tác vụ.

GPT-2 Model

Các tác vụ **Zero-shot NLP**

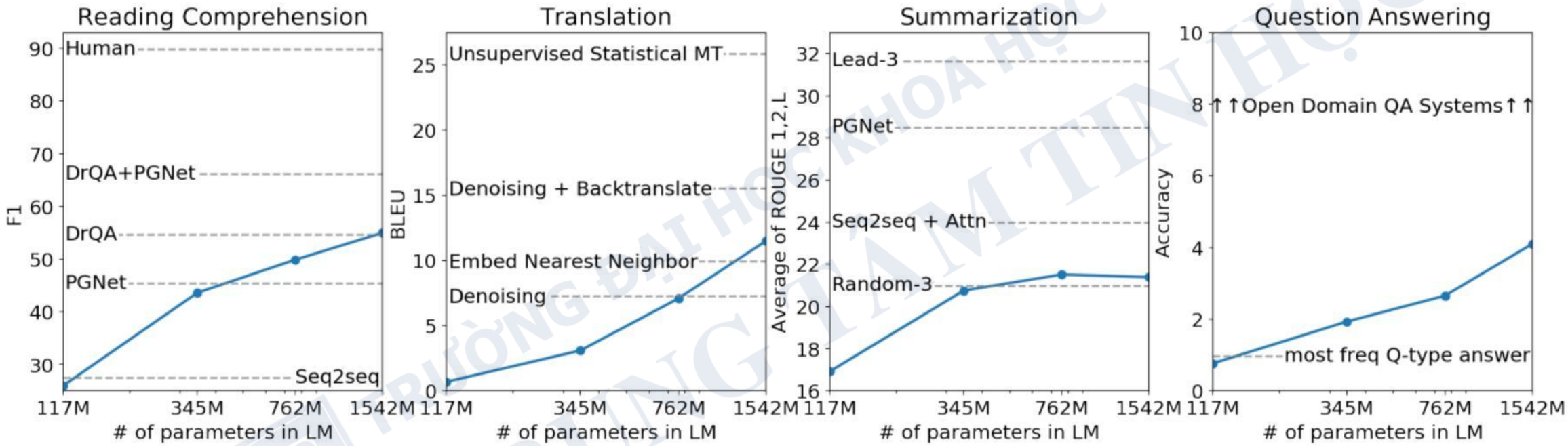
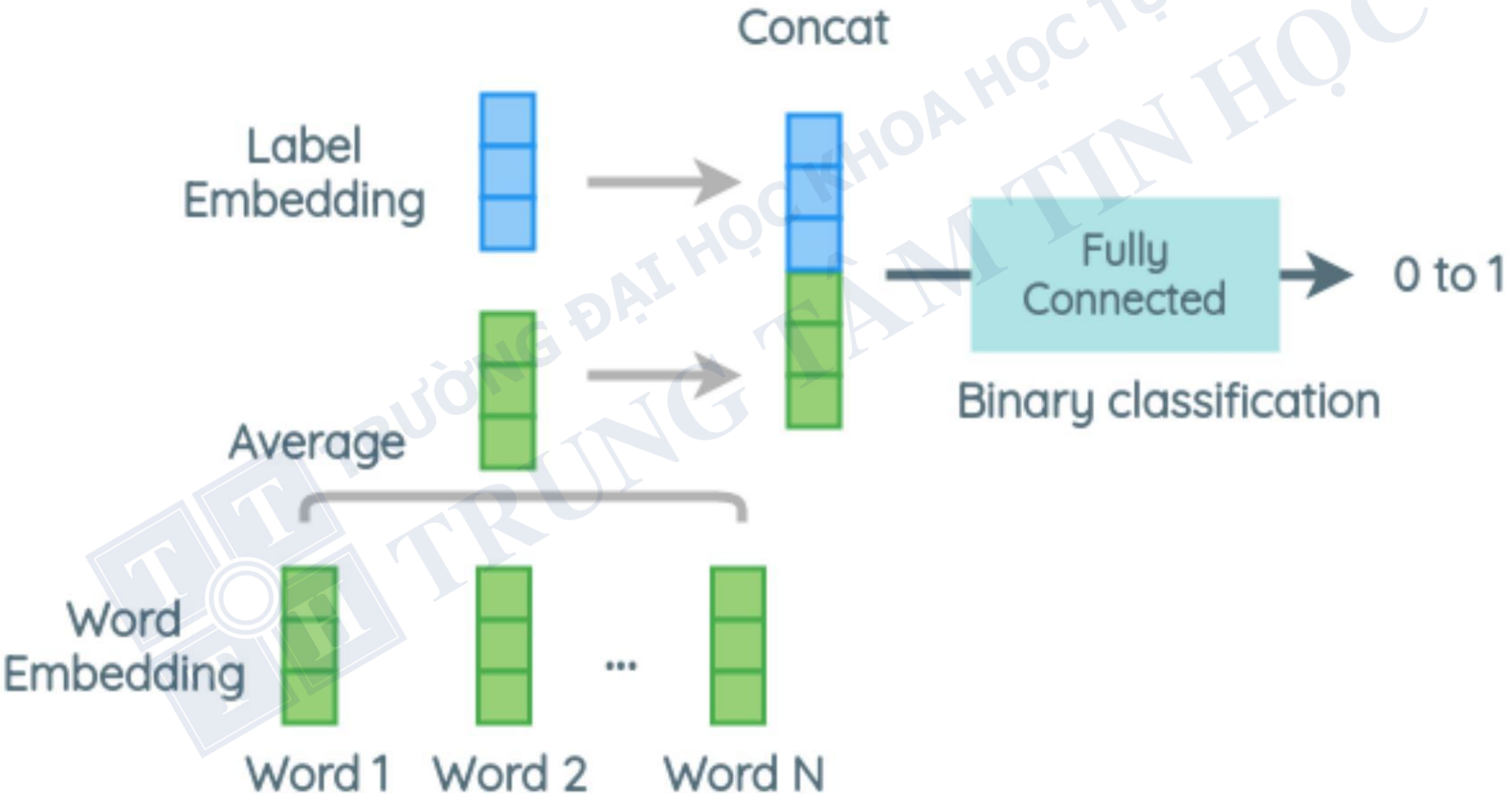


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

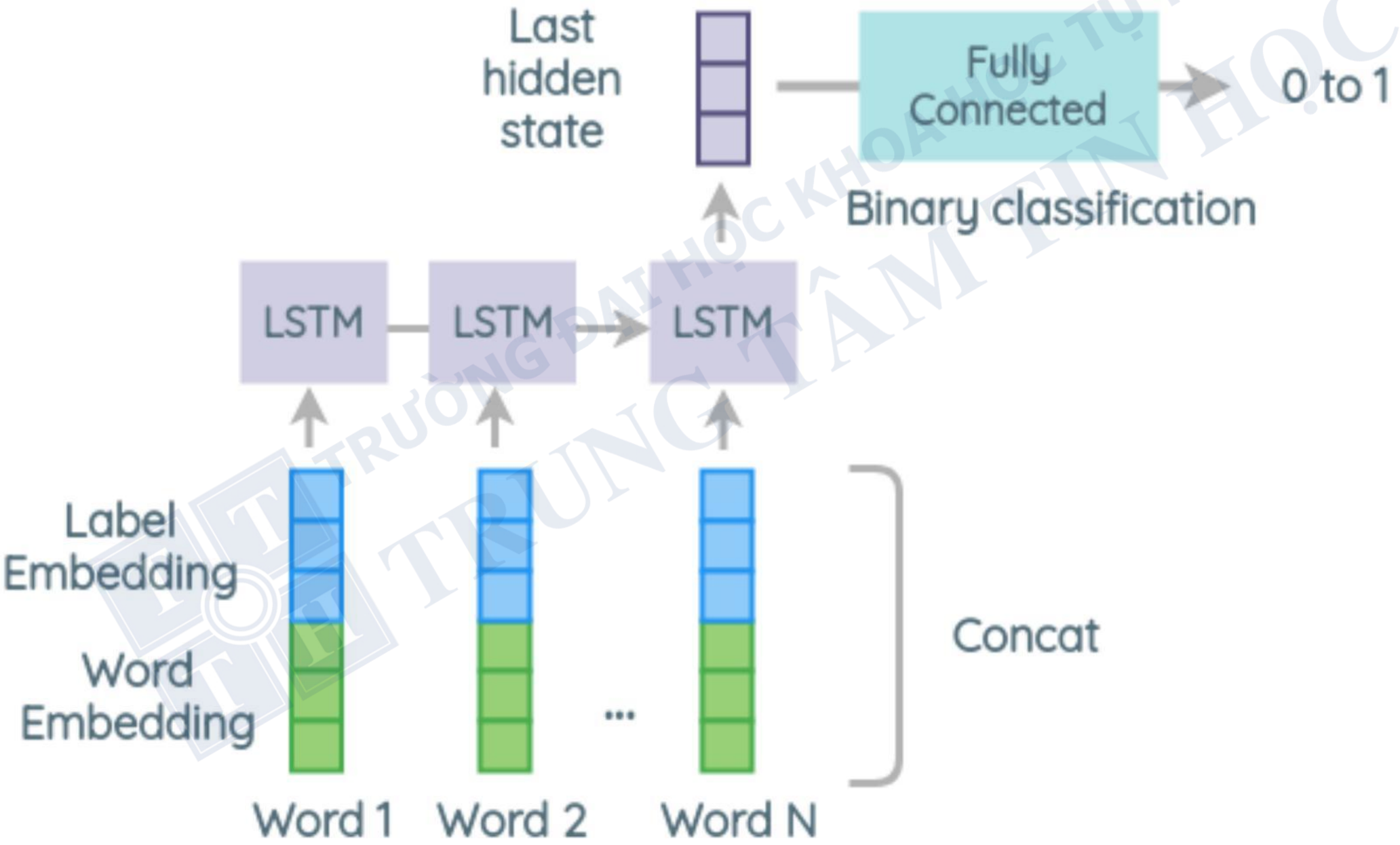
GPT-2 Model

Dạng I: GTBT

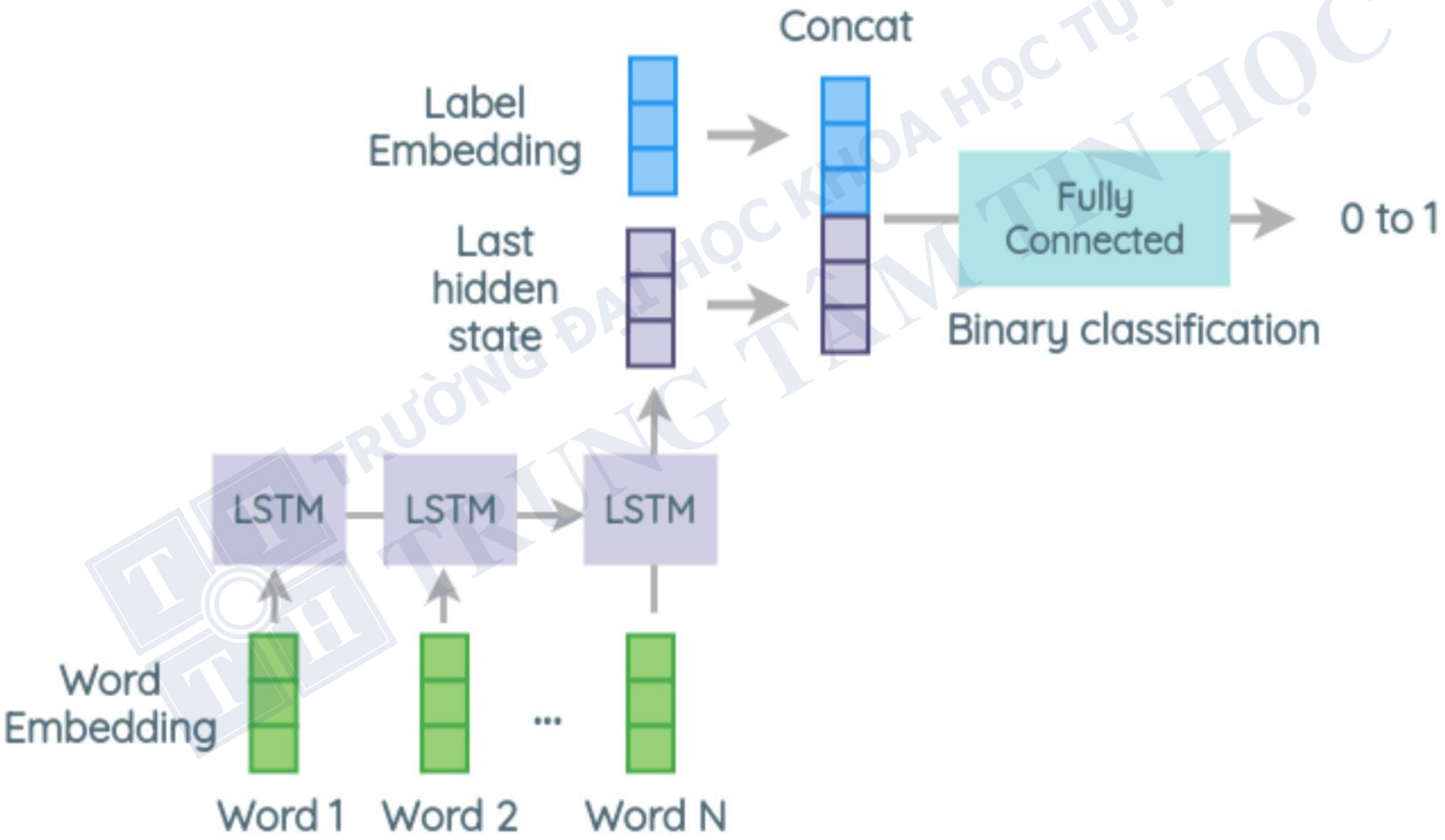


GPT-2 Model

Dạng II: LSTM Models



Dạng III: LSTM Models + Label embedding



GPT-2 Model

Large Models

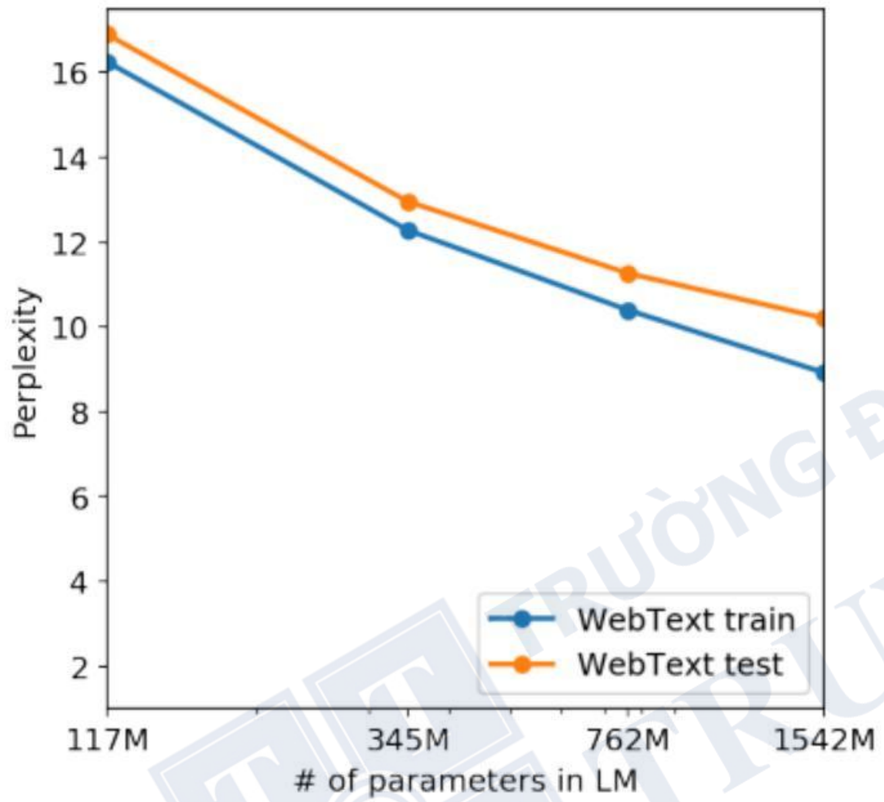
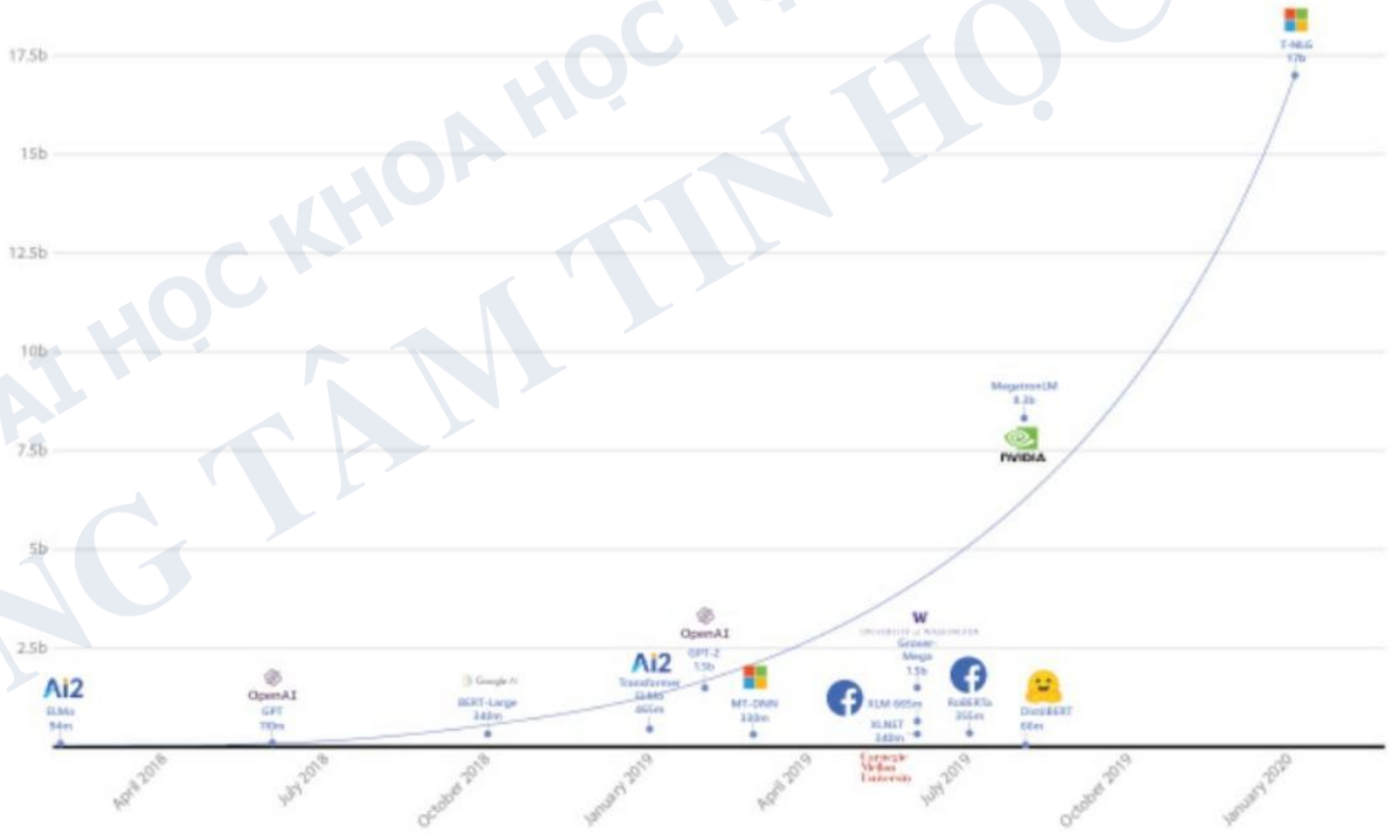


Figure 4. The performance of LMs trained on WebText as a function of model size.



	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

MÔ HÌNH **OPENAI GPT**

I. Masked Self-Attention

II. Language Modeling

III. GPT-1

IV. GPT-2

V. GPT-3

GPT-3 Model

- Kích thước lớn hơn so với GPT-1, GPT-2
- **Zero-Shot, One-Shot & Few-Shot** learners.

Input Prompt: `Recite the first law of robotics`

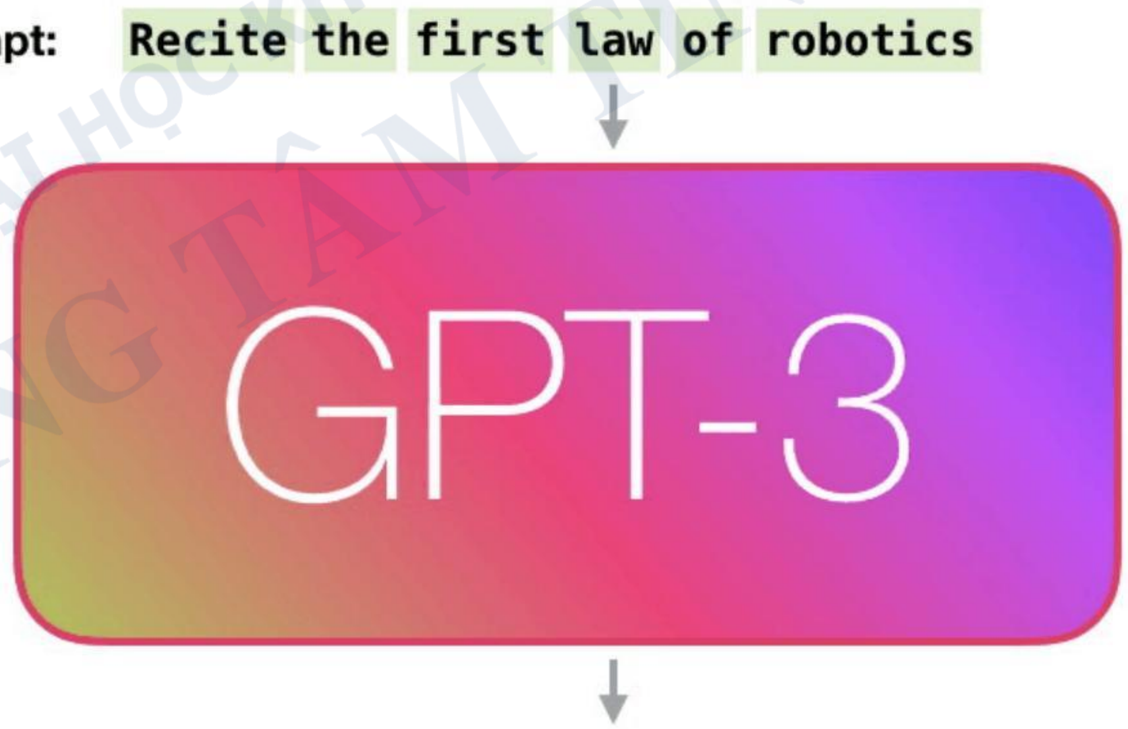
Output là câu

Output:

GPT-3 Model

- Kích thước lớn hơn so với GPT-1, GPT-2
- **Zero-Shot, One-Shot & Few-Shot** learners.

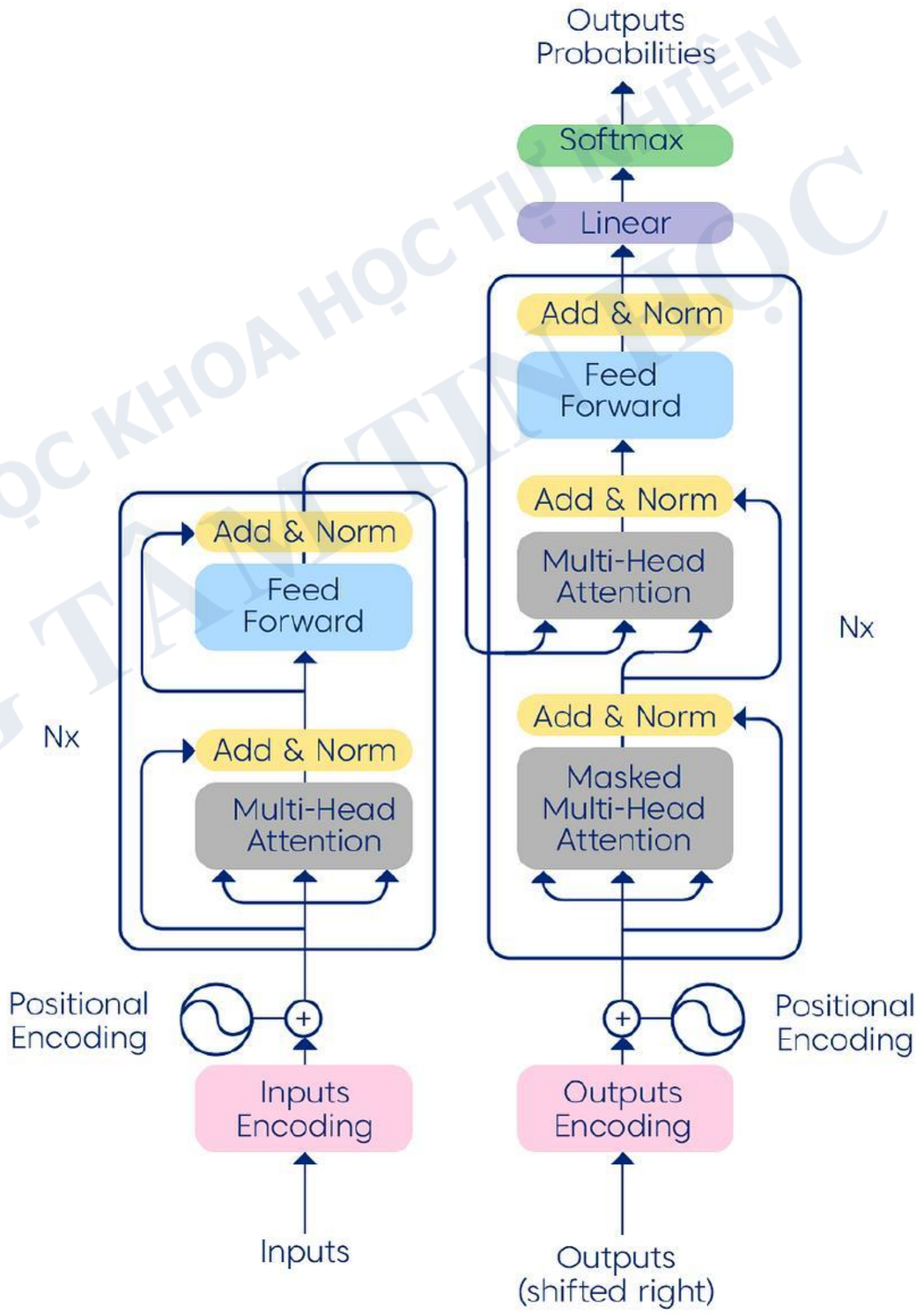
Output là từng từ



GPT-3 Model

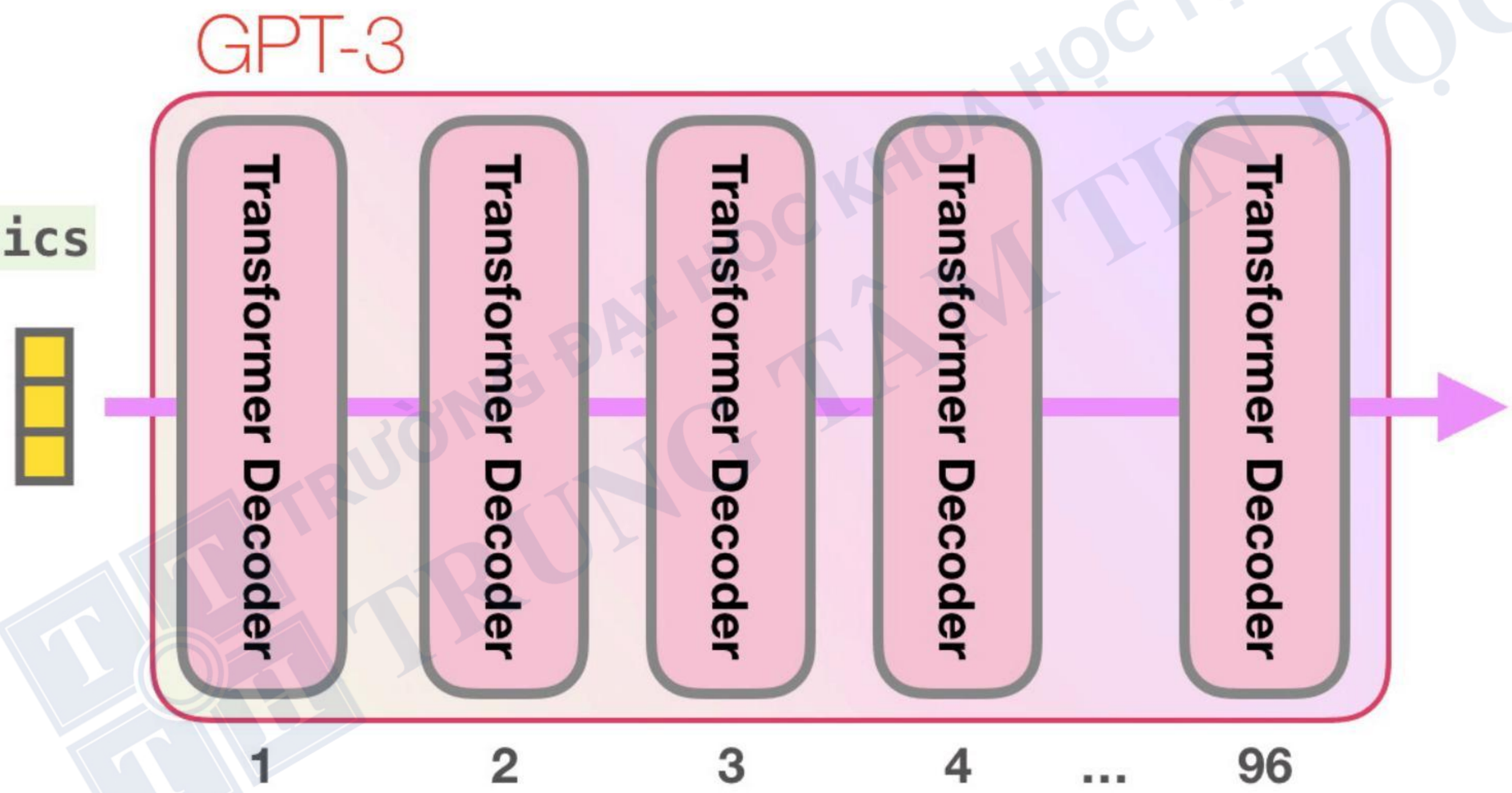
Cấu trúc mô hình

GPT-3



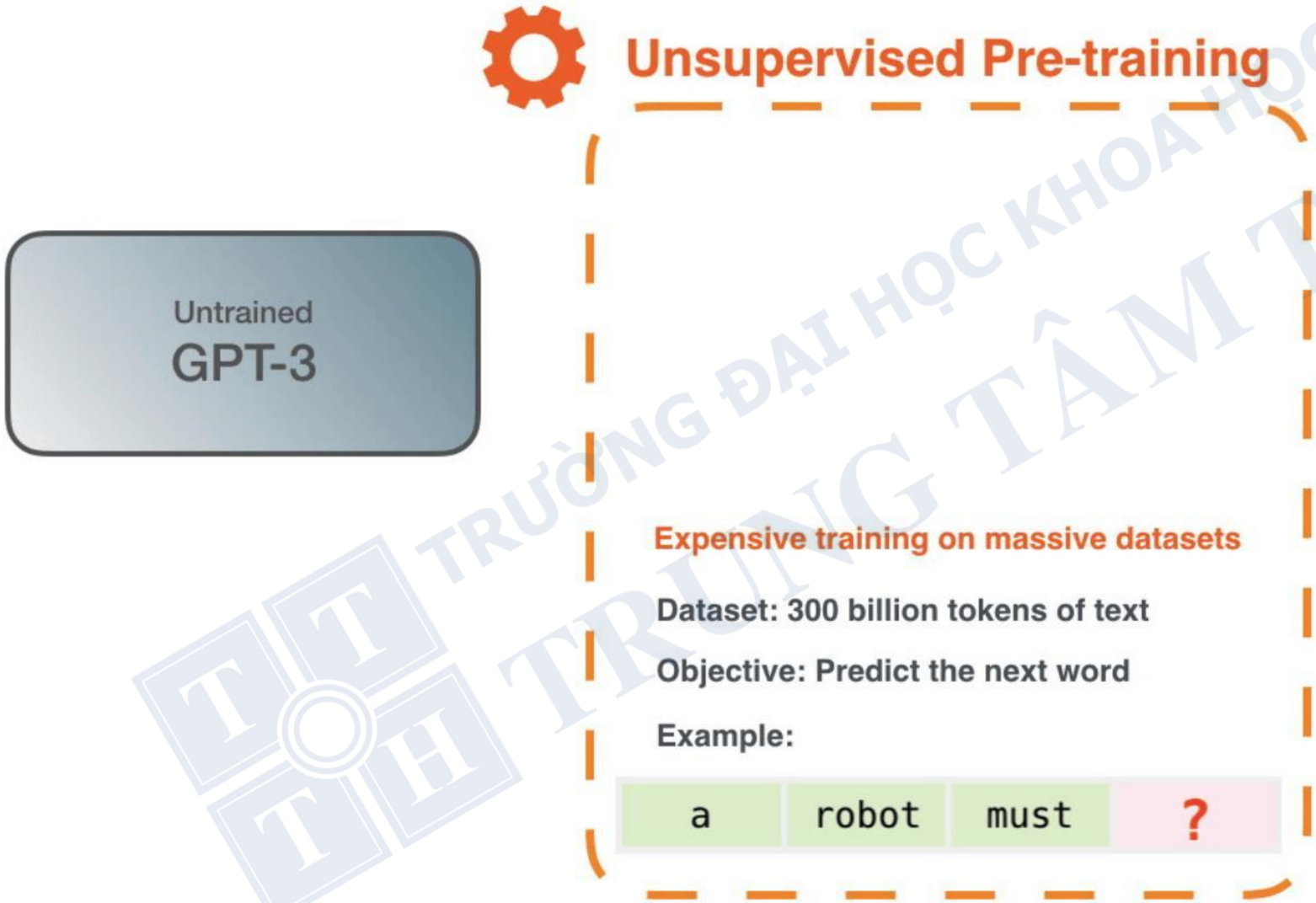
GPT-3 Model

GPT-3 Decoder



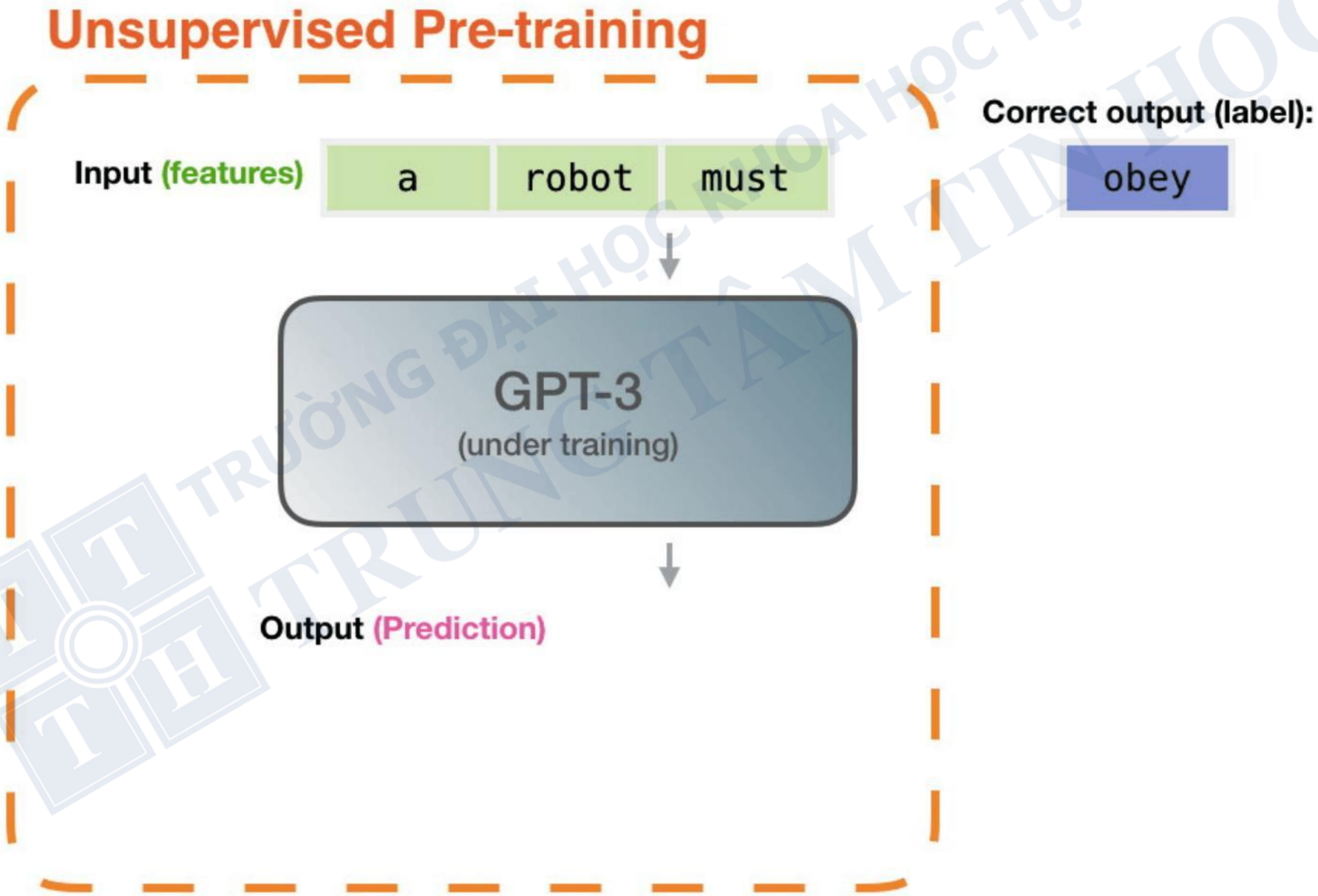
GPT-3 Model

GPT-3 **Pre-training**



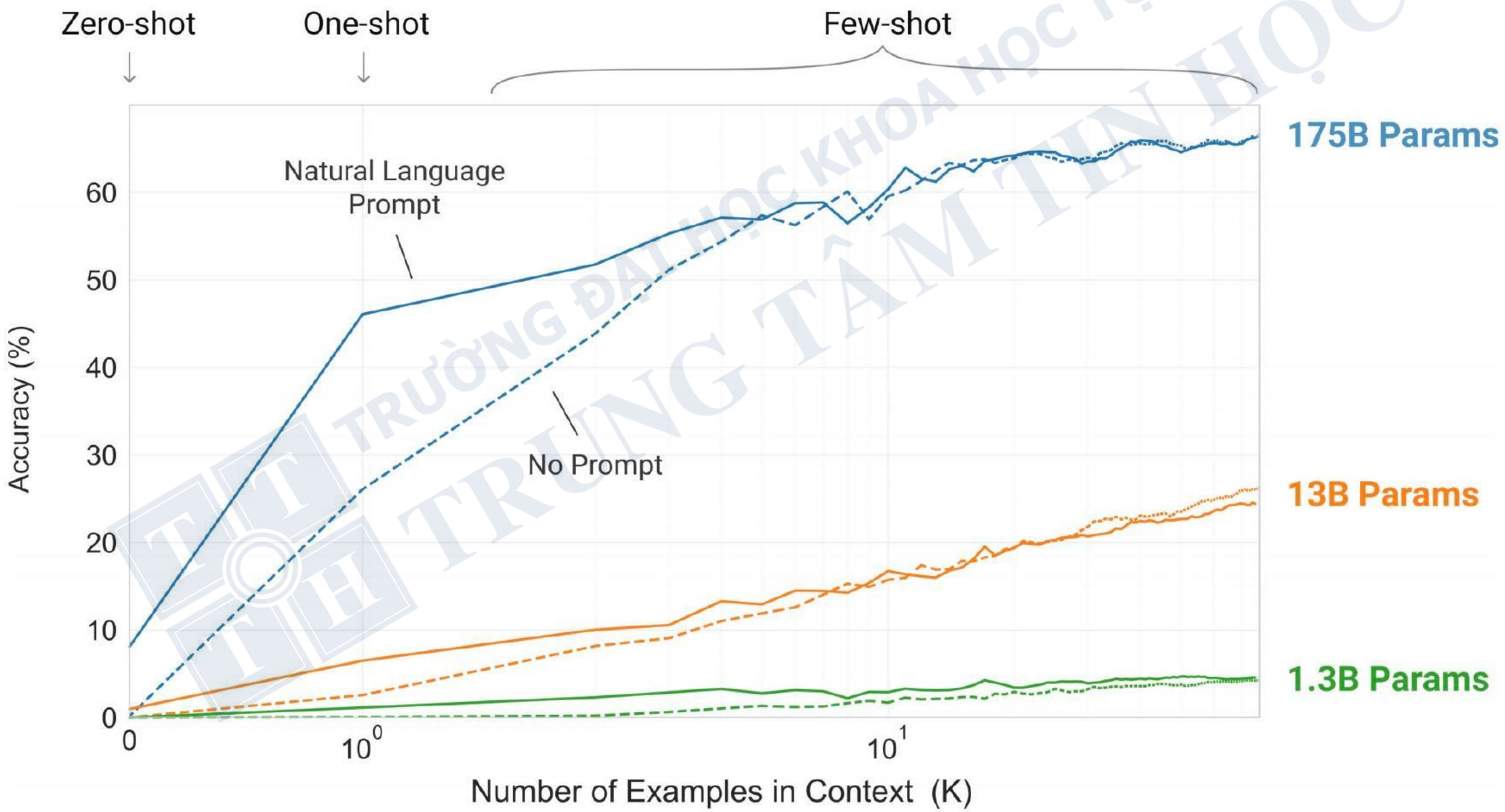
GPT-3 Model

GPT-3 Pre-training



GPT-3 Model

Zero-Shot, One-Shot & Few-Shot



GPT-3 Model

Zero-Shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description

2  cheese =>                           ← prompt
```

One-Shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

GPT-3 Model

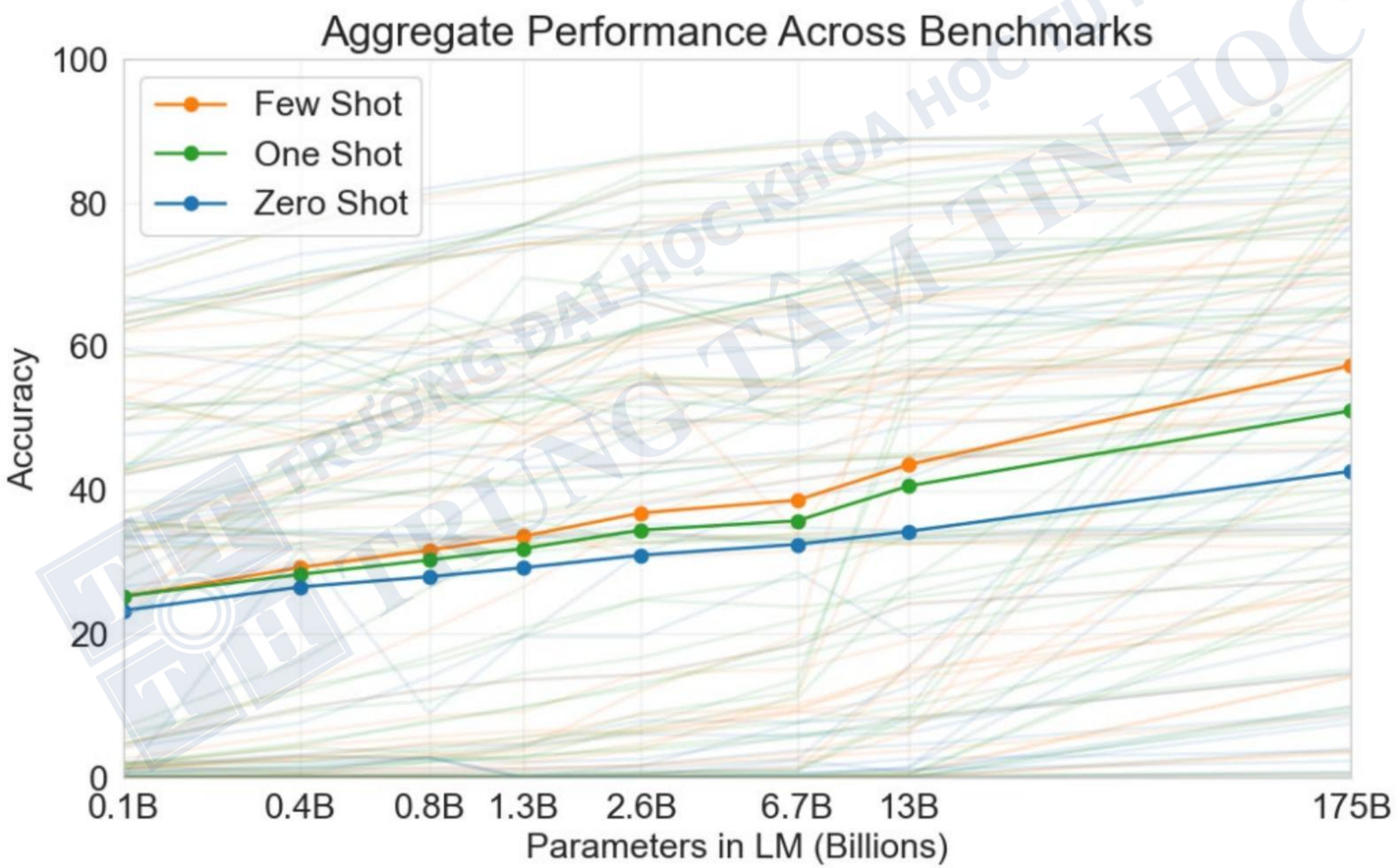
Few-Shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:          ← task description
2
3  sea otter => loutre de mer            ← examples
4
5  peppermint => menthe poivrée          ←
6
7  plush girafe => girafe peluche        ←
8
9  cheese =>                             ← prompt
```

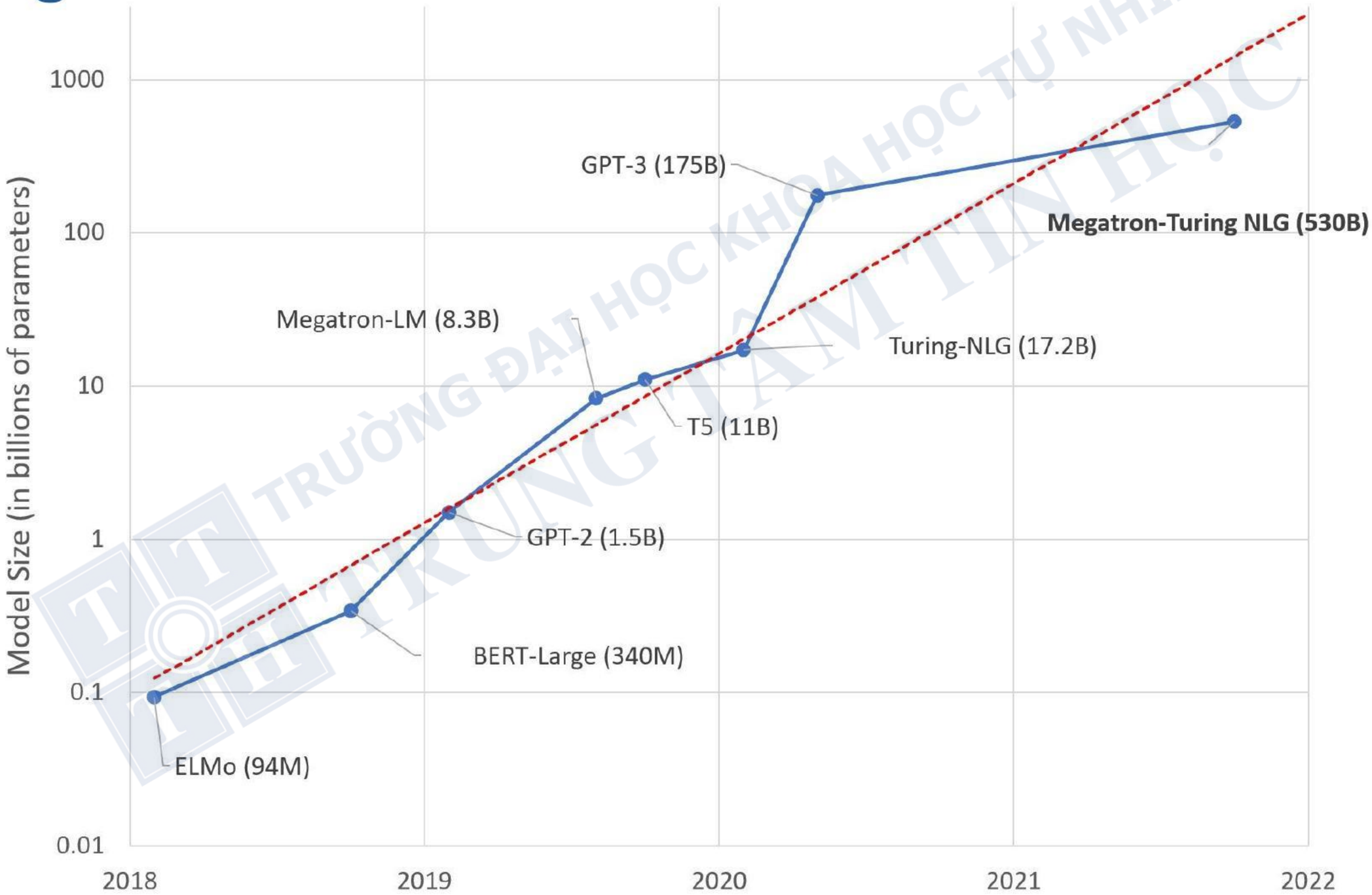
GPT-3 Model

Hiệu quả của **GPT-3**



GPT-3 Model

Large Models



GPT-3 Model

Large Models

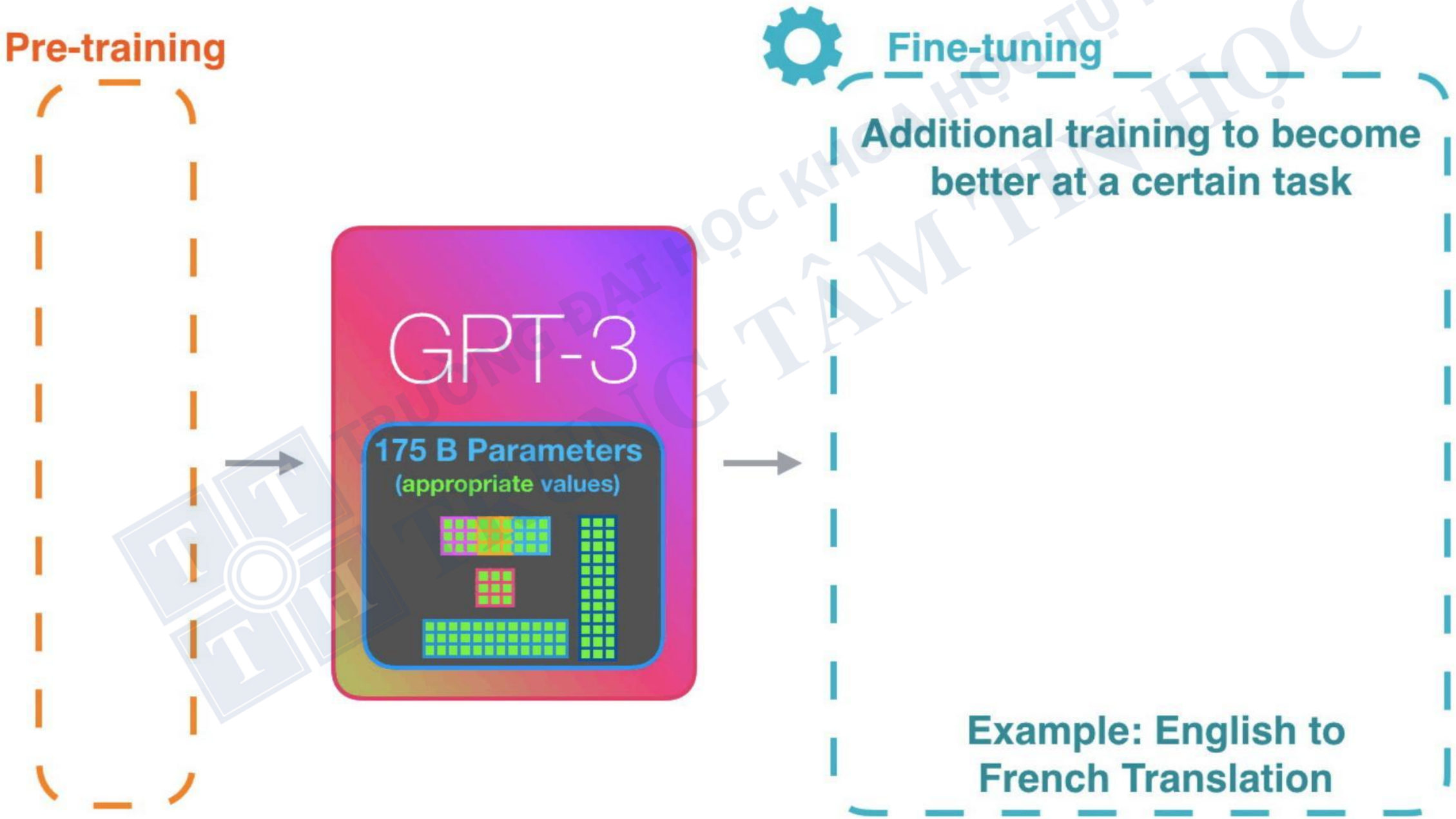
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

GPT-3 Model

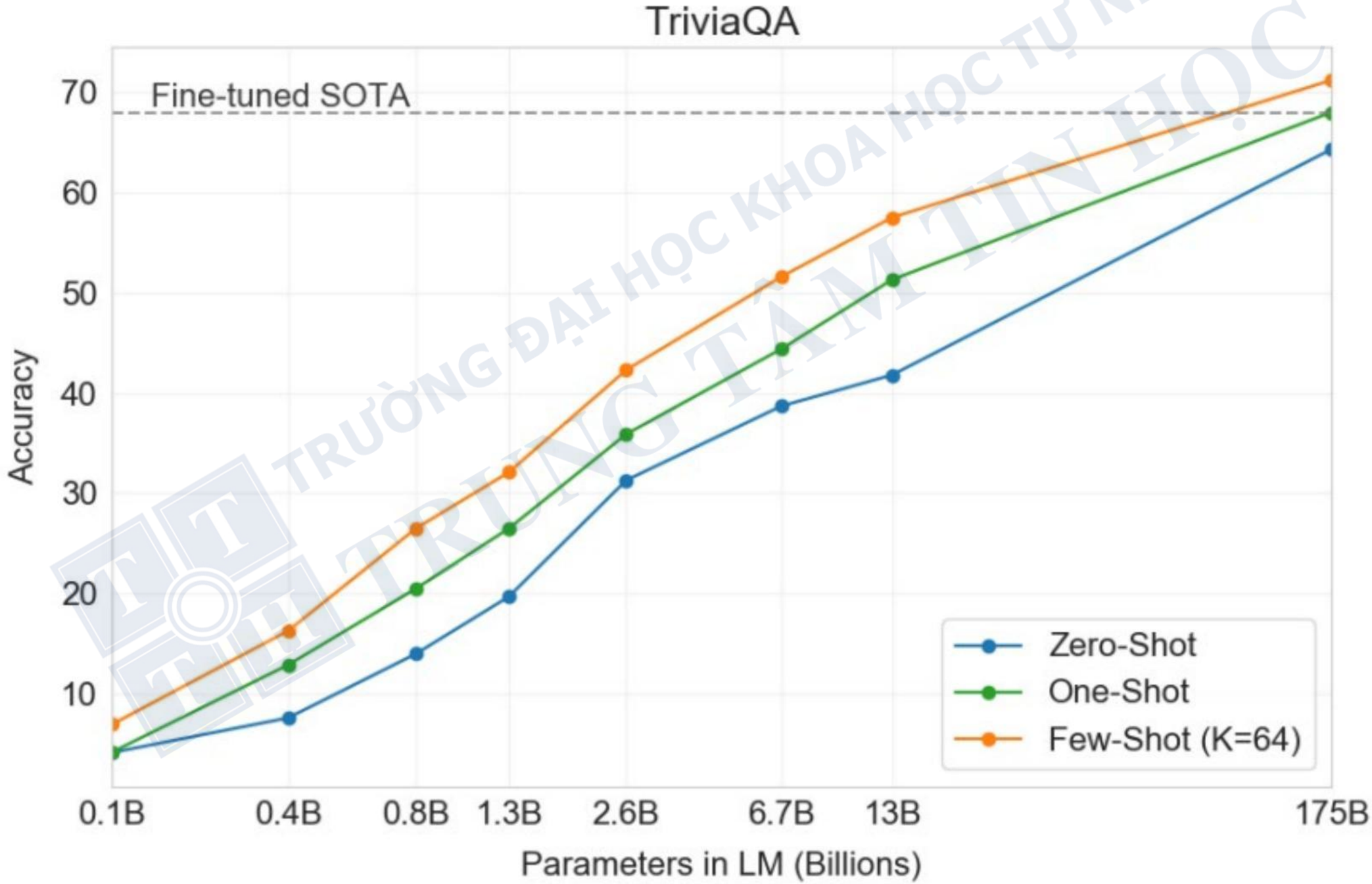
GPT-3 **Fine-tuning**

GPT-3 Fine-tuning



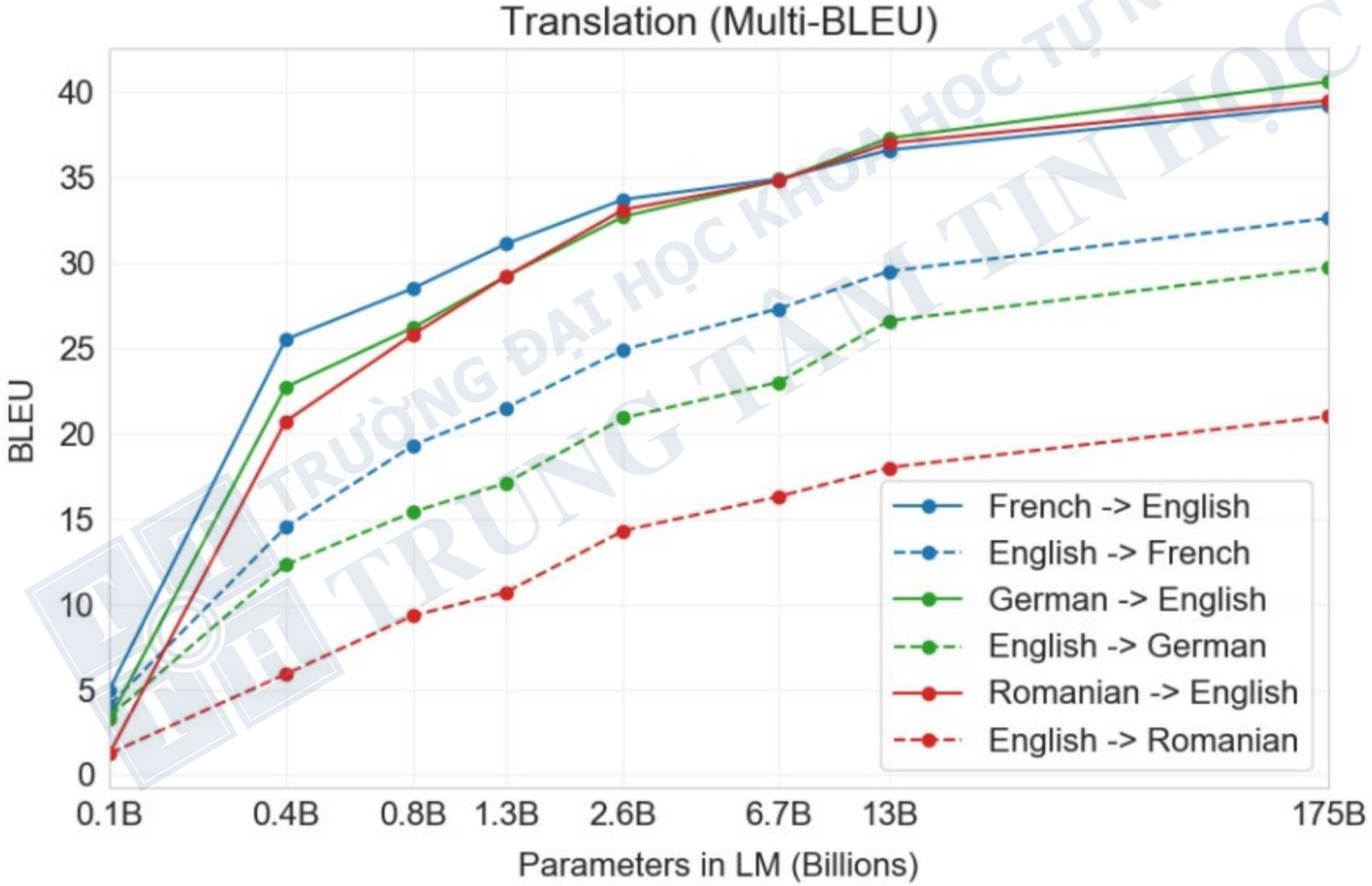
GPT-3 Model

GPT-3 Q&A



GPT-3 Model

GPT-3 Translation



GPT-3 Model

LAMBADA:

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Hạn chế của GPT-3

1. Giới hạn về **kích thước input và output.**
2. GPT-3 thiếu về **các định dạng bộ nhớ.**
3. GPT-3 **chậm** → Tốn nhiều thời gian chạy và train.
4. Khả năng suy luận hạn chế, đặc biệt là với các **prompts có độ dài lớn.**

So sánh các mô hình GPT

	Parameters	Decoder layers	Context lenght	Hidden layer size
GPT-1	117 million	12	512	768
GPT-2	1.5 billion	48	1024	1600
GPT-3	175 billion	96	2048	12288
GPT-4	1.76 trillion	120	8000*	20k*

Code Demo

DEMO

Q&A