

Intelligent QA System With NLP



GROUP:

MEMBERS:

20127258 - 20127655 - 20127625 - 20127597

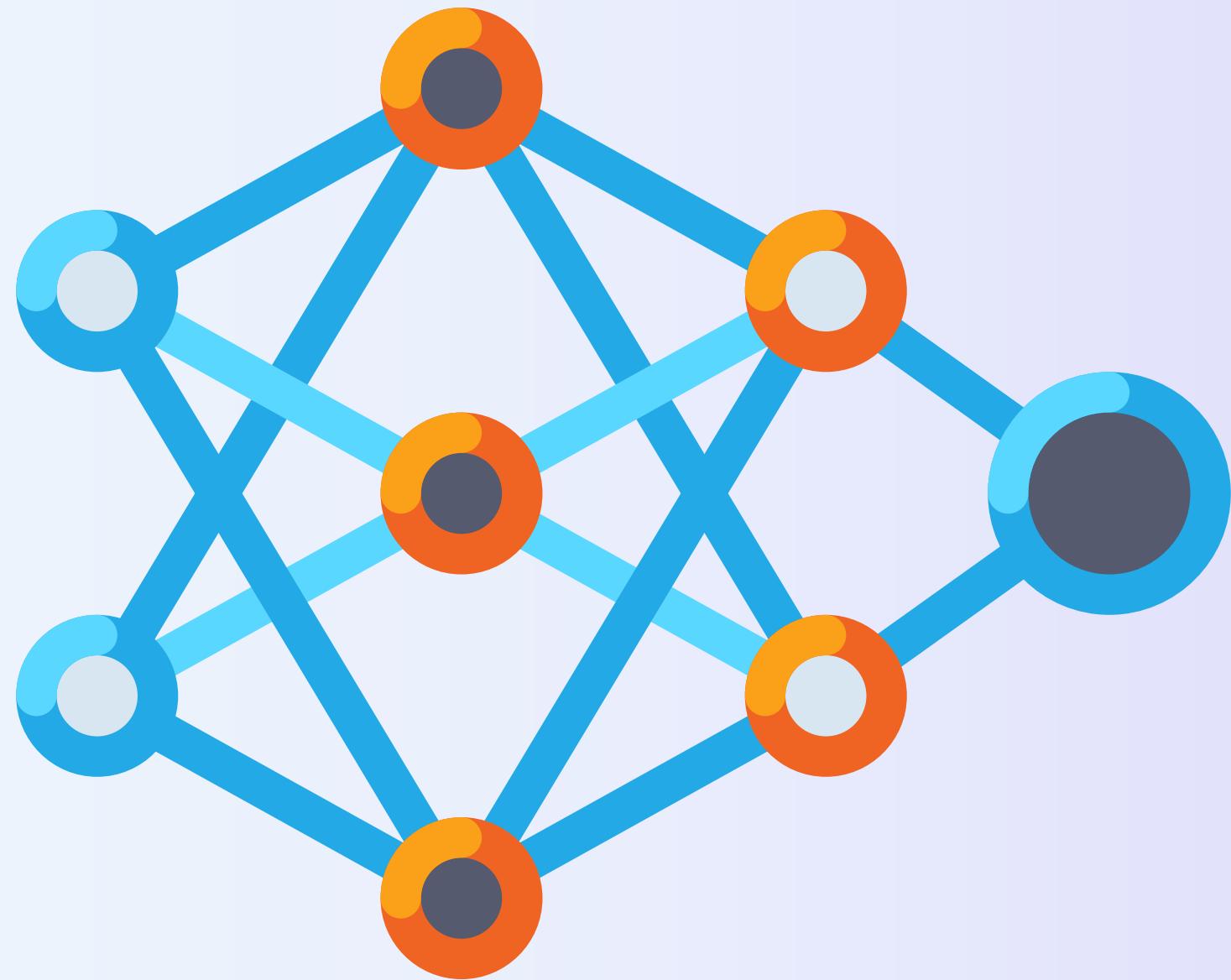
1

PHƯỚC NGUYỄN - QUỐC TRUNG - HOÀNG THÁI - TẤN PHƯƠNG

NỘI DUNG

Nội dung phủ từ lý thuyết mô hình đến kết quả
thực nghiệm của bài toán Question Answering
trong đồ án với hai mô hình PhoBERT và BARTpho.

- 1 MÔ HÌNH PHOBERT**
- 2 MÔ HÌNH BARTPHO**
- 3 THỰC NGHIỆM**



1
**MÔ HÌNH
PHOBERT**

GIỚI THIỆU PHOBERT

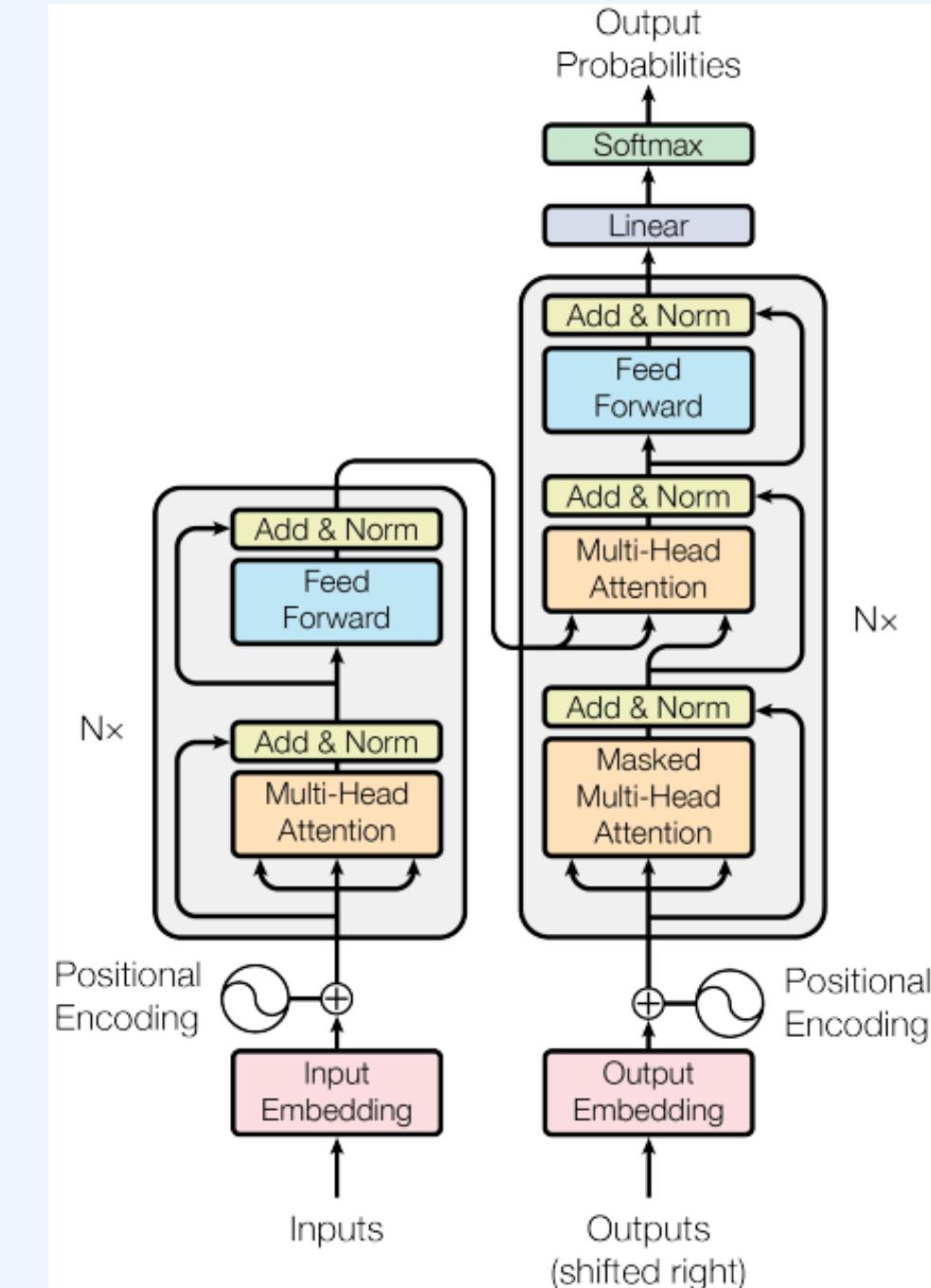
-  PhoBERT là mô hình đơn ngôn ngữ lớn đầu tiên dành cho Tiếng Việt có thể giải quyết nhiều bài toán như POS tagging, Named-entity recognition, Question Answering...
-  PhoBERT gồm hai phiên bản chính là phobert-base và phobert-large.
-  Mô hình được sử dụng là phobert-base-v2 được tinh chỉnh, kiểm thử và khảo sát với bài toán Question Answering thông qua tập dữ liệu Project2_Data.

```
{
  "context": "Về mặt kiến trúc, trường có một nhân vật Công giáo. Trên đỉnh mái vòm vàng tòa nhà chính là bức tượng vàng của Đức Trinh Nữ Ma",
  "qas": [
    {
      "id": "5733be284776f4190066117e",
      "question": "Có gì ngồi trên đầu trang của Tòa nhà Chính tại Notre Dame?",
      "answers": [
        {
          "text": "bức tượng vàng của Đức Trinh Nữ Maria",
          "answer_start": 98
        }
      ],
      "is_impossible": false
    },
    {
      "id": "5733be284776f4190066117f",
      "question": "là những gì ở phía trước của Notre Dame Tòa nhà Chính?",
      "answers": [
        {
          "text": "một bức tượng đồng của Chúa Kitô",
          "answer_start": 200
        }
      ],
      "is_impossible": false
    },
    {
      "id": "5733be284776f41900661180",
      "question": "Vương Cung Thánh Đường của trái tim Thánh tại Notre Dame là bên cạnh đê'mà cầu trúc?",
      "answers": [
        {
          "text": "Tòa nhà Chính",
          "answer_start": 304
        }
      ],
      "is_impossible": false
    }
  ]
}
```

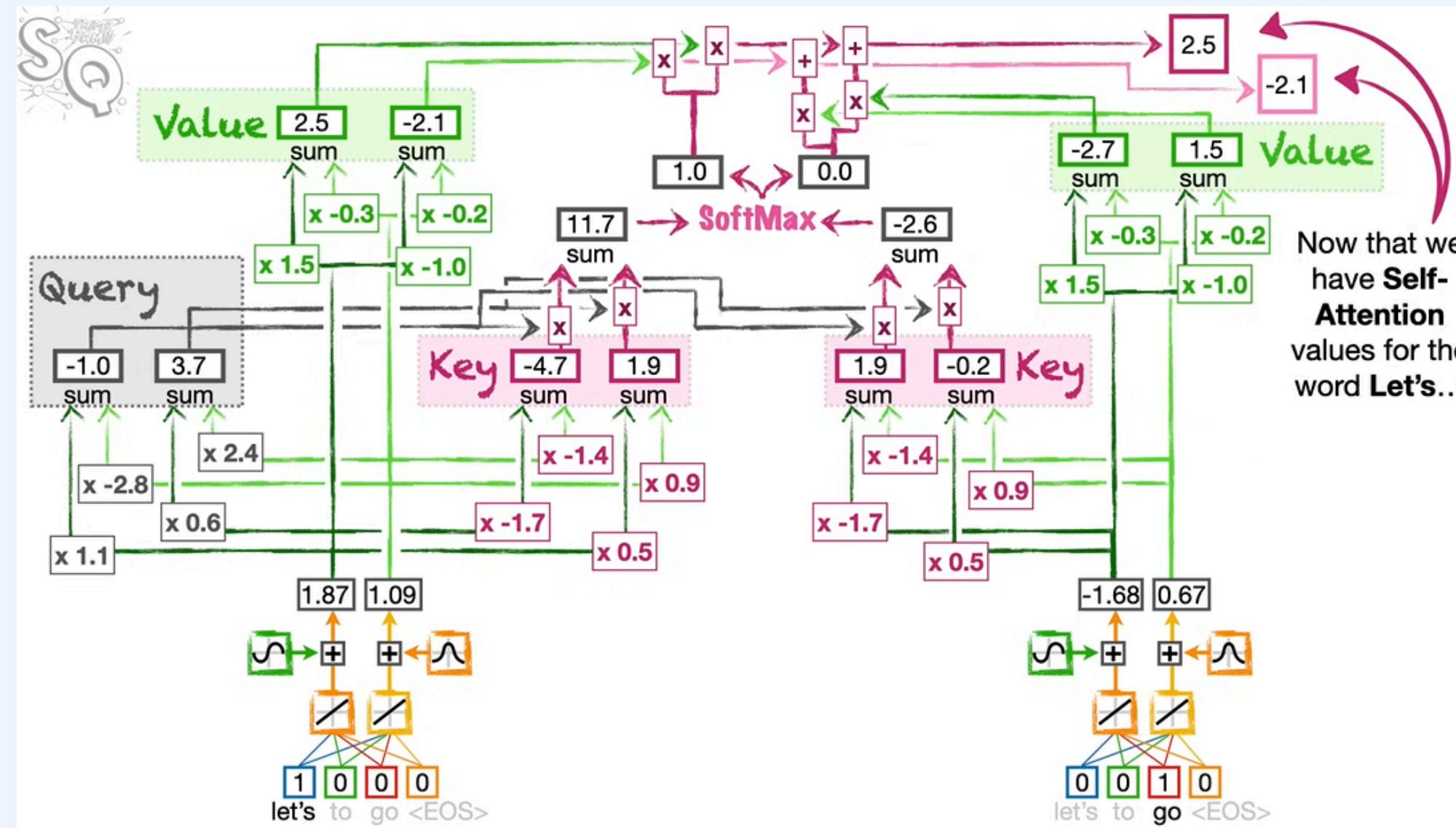
Một vài mẫu dữ liệu minh họa cho tập dữ liệu Project2_Data

CẤU TRÚC VANILLA TRANSFORMER

- Transformer được công bố lần đầu tiên vào năm 2017.
- Mô hình gồm hai thành phần chính là bộ mã hóa [encoder] và bộ giải mã [decoder] được tạo thành từ hai tầng con là tầng chú ý đa đầu [Multi-Head Attention] và mạng nơ-ron truyền thẳng [Feed-Forward Network].



Mô hình vanilla Transformer gồm bộ mã hóa [bên trái] và bộ giải mã [bên phải]



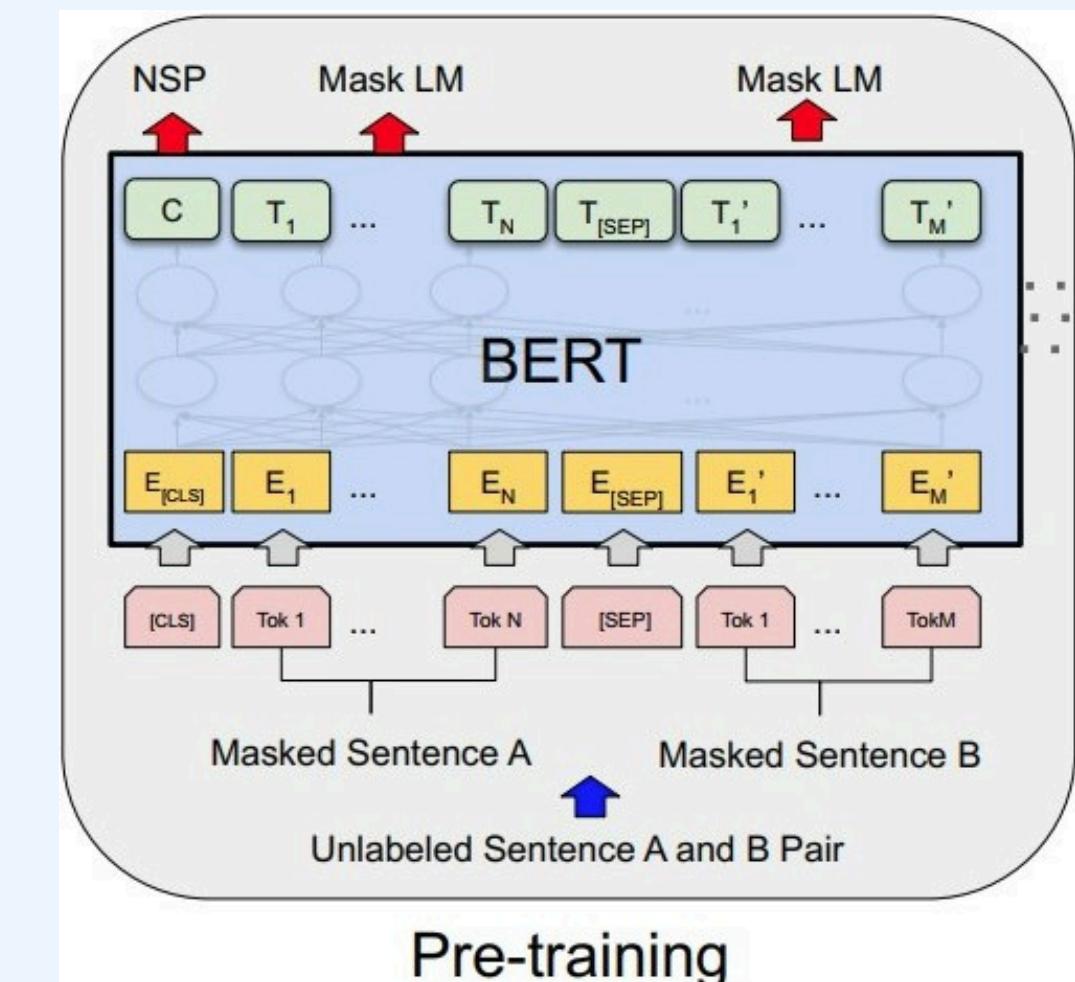
Quá trình tính toán song song của mô hình Transformer

ƯU ĐIỂM

- ✓ Tối ưu quá trình tính toán.
- ✓ Chi phí thực thi thấp.

CẤU TRÚC - BERT

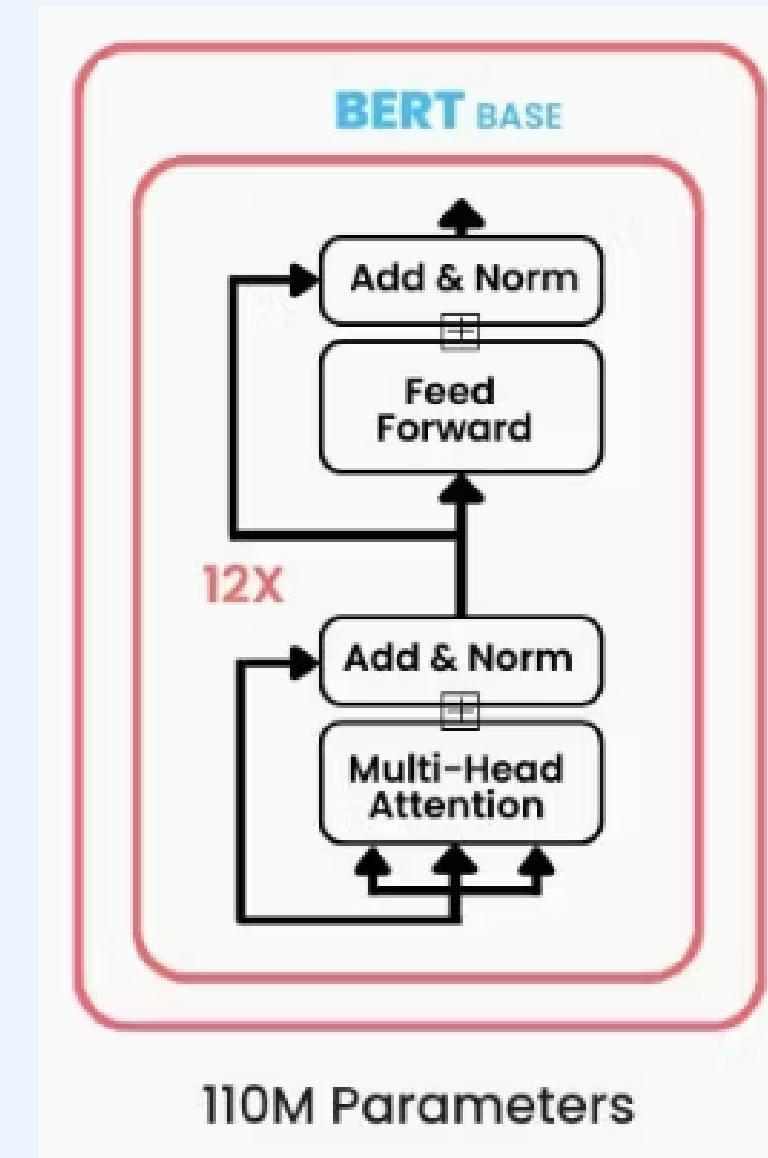
- ✓ Mô hình BERT chỉ bao gồm tầng mã hóa với tính chất hai chiều (bidirectional).
- ✓ MLM mô tả sự che khuất dữ liệu huấn luyện. Với mỗi chuỗi [sequence] đầu vào, 15% số lượng từ [token] trong chuỗi được lựa chọn ngẫu nhiên để thay thế bởi:
 - Từ [MASK] trong 80% số lần thực hiện.
 - Một từ ngẫu nhiên trong 10%.
 - Chính từ được chọn trong 10%.
- ✓ NSP mô tả quá trình lựa chọn một cặp câu [sentence pairs]. Một chuỗi bao gồm 2 câu A và B, trong đó:
 - 50% số lần thực hiện thì câu B là câu kế tiếp tương ứng với câu A (với nhãn IsNext).
 - 50% còn lại là sự lựa chọn ngẫu nhiên.



Minh họa quá trình huấn luyện của
mô hình BERT

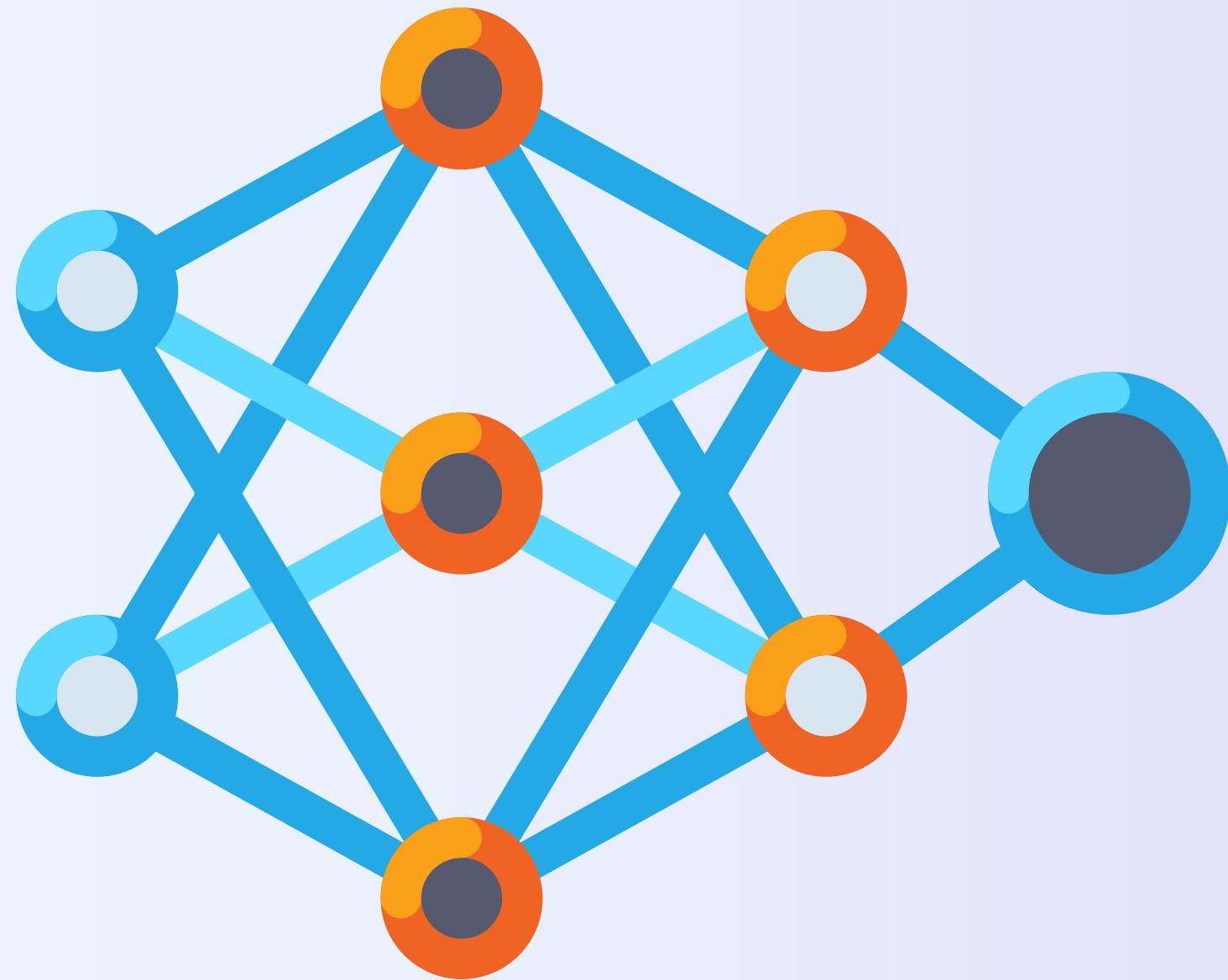
CẤU TRÚC CỤ THỂ

- ✓ Mô hình bert-base được cấu tạo từ 12 tầng mã hóa ($L=12$), với kích thước ẩn [hidden size] được sử dụng là 768 ($H=768$) và 12 tầng chú ý đa đầu ($A=12$).
- ✓ Mặc dù có cấu trúc tương tự với mô hình bert-base nhưng mô hình phobert-base-v2 có tổng số lượng tham số là 135 triệu tham số.



ĐIỂM MẠNH MÔ HÌNH

-  PhoBERT được huấn luyện với tập dữ liệu đa dạng [20GB dữ liệu từ tập dữ liệu Wikipedia Tiếng Việt 1GB và news corpus Tiếng Việt 19GB] đem lại hiệu suất trong các bài toán đặc trưng cho Tiếng Việt.
-  PhoBERT là mô hình phân biệt âm tiết (syllable) và từ (token). Ví dụ với một câu 5 âm tiết (5-syllable) “Tôi là một sinh viên” sẽ tương ứng với 4 từ “Tôi là một sinh_viên”.
-  PhoBERT kế thừa cấu trúc mô hình từ BERT và các cải tiến tối ưu từ mô hình RoBERTa.



2

MÔ HÌNH BARTPHO

GIỚI THIỆU BARTPHO

- ✓ BARTPho là mô hình seq2seq được huấn luyện trước cho tiếng Việt, dựa trên kiến trúc BART, giải quyết vấn đề xử lý đặc điểm âm tiết và từ của tiếng Việt.
- ✓ BARTPho được cài đặt sẵn trong transformers, dễ dàng sử dụng cho các ứng dụng xử lý ngôn ngữ tự nhiên tiếng Việt.
- ✓ Lý do chọn mô hình: Hỗ trợ cực kỳ tốt cho tiếng việt, được tích hợp sẵn trong transformers, fine-tune dễ dàng trên nền tảng Kaggle.

VinAIResearch/
BARTpho

BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese (INTERSPEECH 2022)



1
Contributor



0
Issues



93
Stars



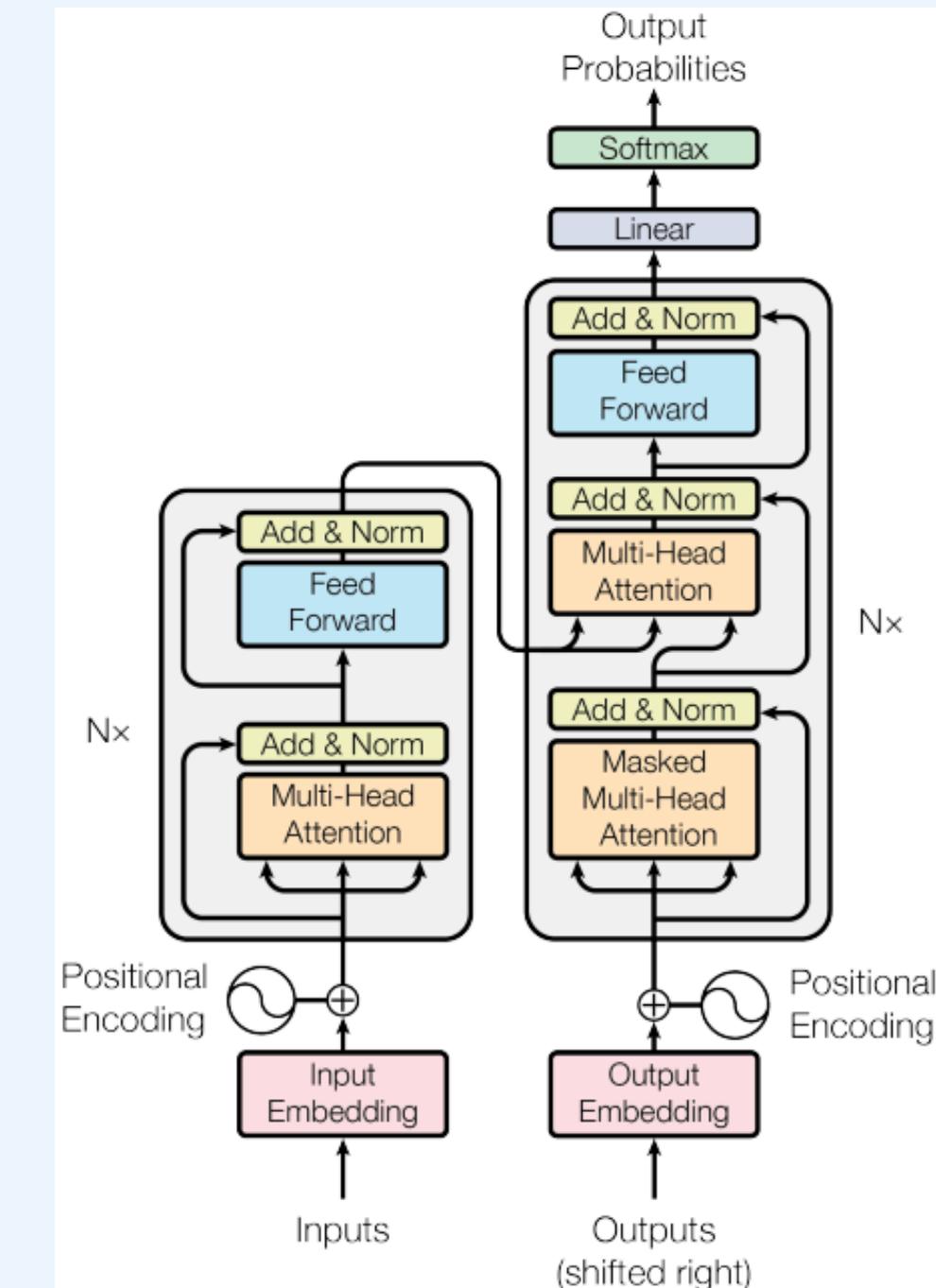
7
Forks



KIẾN TRÚC MÔ HÌNH

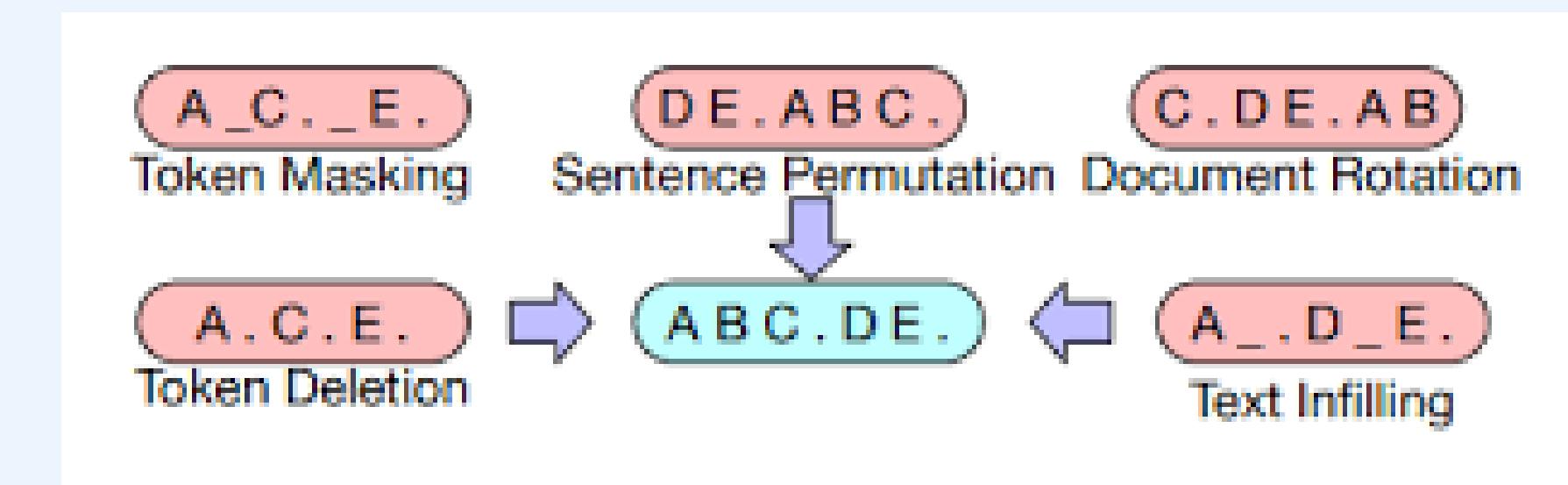
- ✓ BARTPho sử dụng kiến trúc Transformer "large" với:
 - 12 lớp encoder và decoder
 - Hàm kích hoạt GeLU thay vì ReLU
 - Lớp chuẩn hóa (layer normalization) trên cả encoder và decoder

- ✓ Hai phiên bản:
 - BARTpho-syllable: Xử lý đầu vào là âm tiết, huấn luyện trên bộ dữ liệu gồm văn bản và âm tiết tương ứng.
 - BARTpho-word: Xử lý đầu vào là từ, huấn luyện trên bộ dữ liệu văn bản.



CƠ CHẾ HUẤN LUYỆN

- ✓ BARTpho sử dụng cơ chế huấn luyện seq2seq denoising autoencoder. Cơ chế có 2 giai đoạn:
 - ✓ GĐ1: Phá hủy văn bản đầu vào.
 - Mục đích: Tạo phiên bản "nhiều" để mô hình học cách khôi phục và hiểu cấu trúc ngôn ngữ.
 - Kỹ thuật:
 - Sentence permutation [hoán vị câu]: Xáo trộn ngẫu nhiên các câu trong khối 512 token.
 - Text filling [điền vào chỗ trống]: Thay thế đoạn văn bản ngẫu nhiên bằng token đặc biệt.



CƠ CHẾ HUẤN LUYỆN

- ✓ BARTpho sử dụng cơ chế huấn luyện seq2seq denoising autoencoder. Cơ chế có 2 giai đoạn:
 - ✓ GĐ2: Học cách khôi phục văn bản gốc:
 - Mục tiêu: Mô hình học cách dự đoán token tiếp theo dựa trên các token trước đó trong văn bản nhiều.
 - Tối ưu hóa cross-entropy giữa đầu ra của decoder và phiên bản gốc.

Intuitively Understanding the Cross Entropy

$$H(P^* | P) = - \sum_i P^*(i) \log P(i)$$

TRUE CLASS DISTIRBUTION PREDICTED CLASS DISTIRBUTION



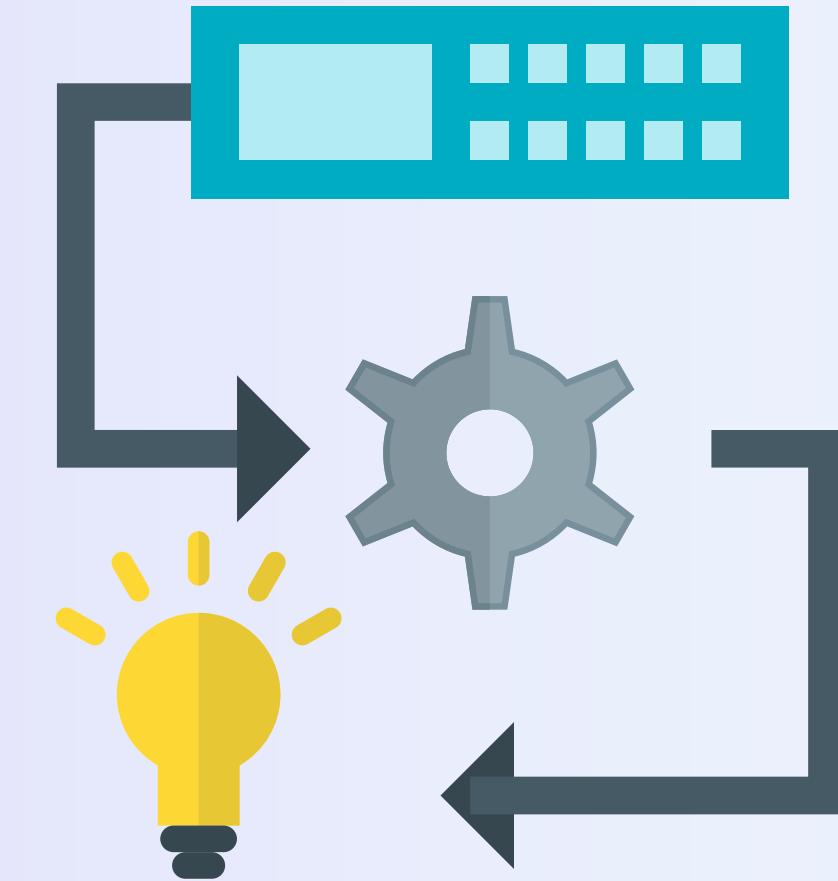
The image shows a blue advertisement for BARTpho. At the top right is the VinAI Research logo. Below it, the text "BARTpho" is displayed in large white letters, preceded by a red arrow pointing right. Underneath "BARTpho", the text "Pre-trained Sequence-to-Sequence Models for Vietnamese" is written in orange. At the bottom left, the URL "www.vinai.io/NLPworkshop2021/" is shown in green. The background features a grid of small red dots.

ƯU ĐIỂM

- ✓ Hiệu suất cao: Tóm tắt, phục hồi dấu câu, viết hoa, tạo văn bản, dịch máy, chatbot,...
- ✓ SOTA tóm tắt văn bản tiếng việt
- ✓ Hai phiên bản: BARTphosyllable và BARTphoword, đáp ứng nhu cầu và dữ liệu đầu vào khác nhau.
- ✓ Dễ triển khai: Tích hợp với thư viện phổ biến fairseq và transformers.

3 **EXPERIMENTS**

Những thực nghiệm và các kết luận của nhóm với phần Q2. Phần này tập trung mô tả các thông số và kết quả cuối cùng của mô hình.



MODELS

Có 5 models được huấn luyện, chia thành 2 loại chính

PhoBERT

1

2

BARTPho

- No fine - tune PhoBERT
- Fine - tune PhoBERT

- No fine - tune BARTPho
- Fine - tune BARTPho

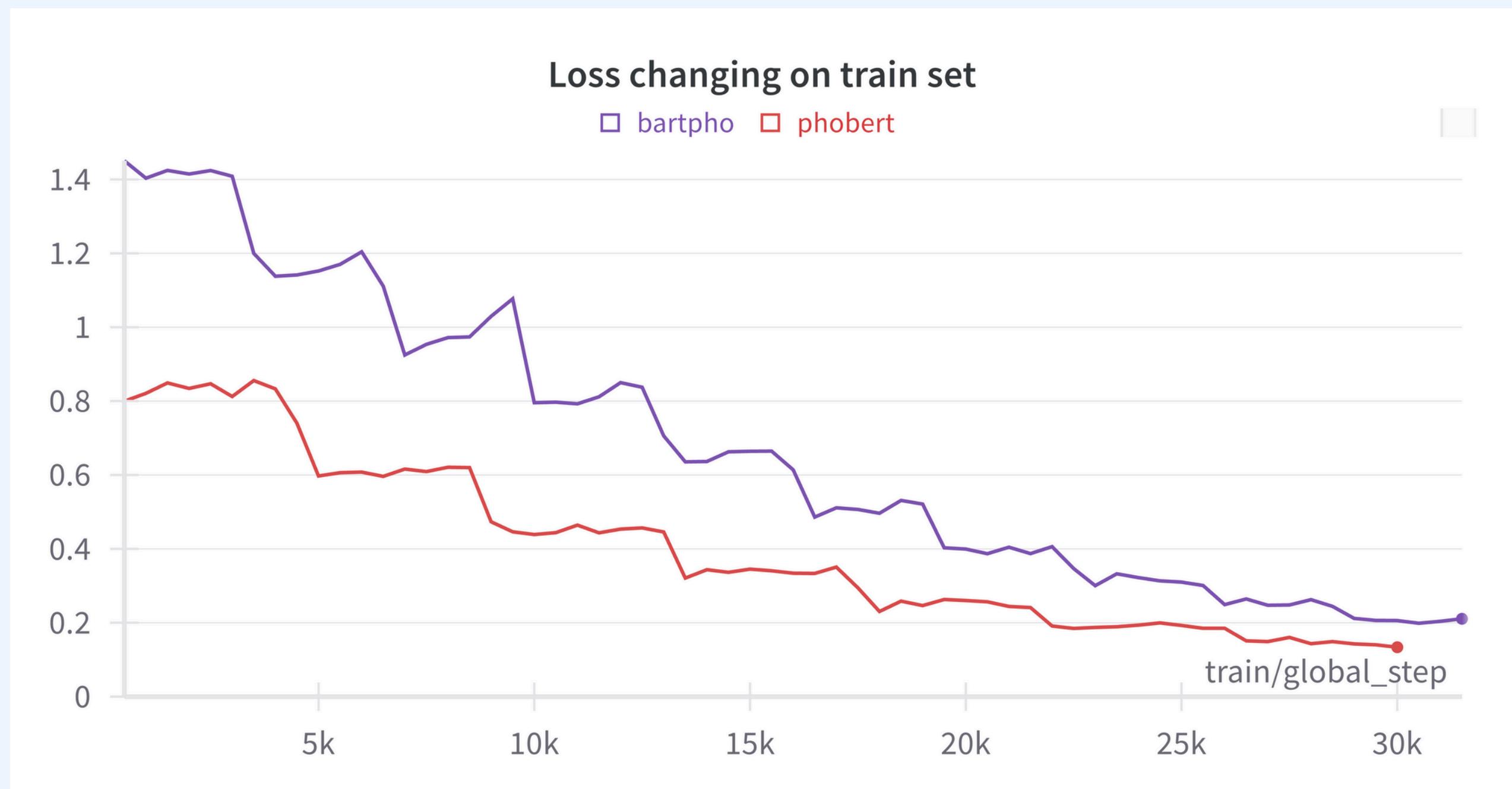
PARAMETERS

Các tham số mô hình mạng để thực hiện huấn luyện mô hình. Những tham số này được tham khảo từ bài báo Convolutional Neural Networks for Sentence Classification của Yoon Kim

MÔ TẢ THAM SỐ	GIÁTRI
Word vector đầu vào	FastText/PhoW2
Embedding size	300
Filter size	(3, 4, 5)
Số các filters	(100, 100, 100)
Hàm kích hoạt	ReLU
Pooling	1-max pooling
Dropout rate	0.5

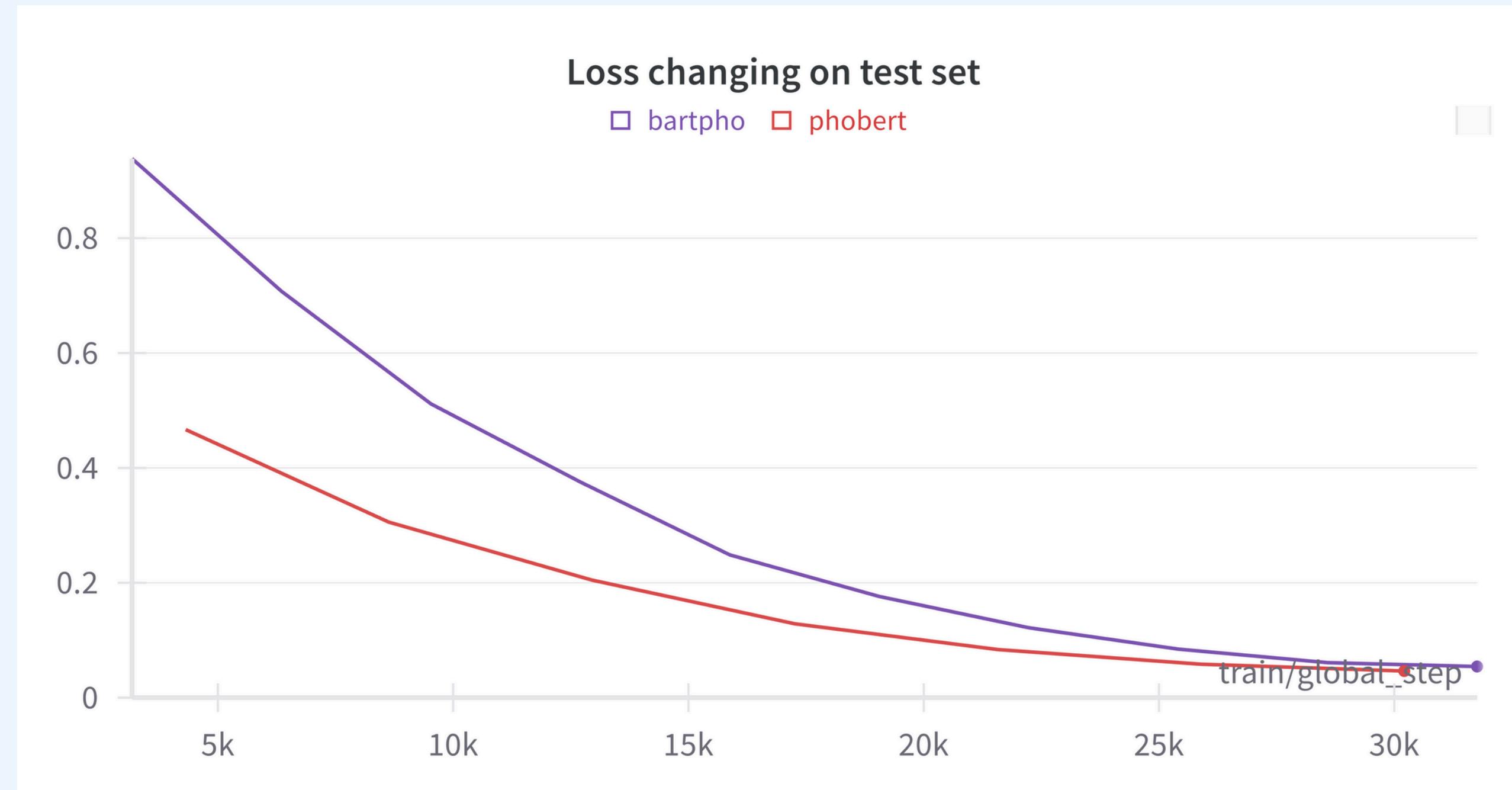
RESULT - LOSS

Train set



RESULT - LOSS

Test set



Total training time

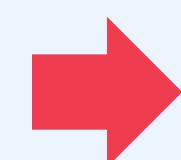


COMMENT #1

Độ lỗi của cả 2 model đều giảm trên
tập train lẫn test trong quá trình
huấn luyện

COMMENT #2

Thời gian huấn luyện của BARTPho lớn
hơn rất nhiều so với PhoBERT, điều
này là dễ hiểu khi kiến trúc của
BARTPho phức tạp hơn PhoBERT.

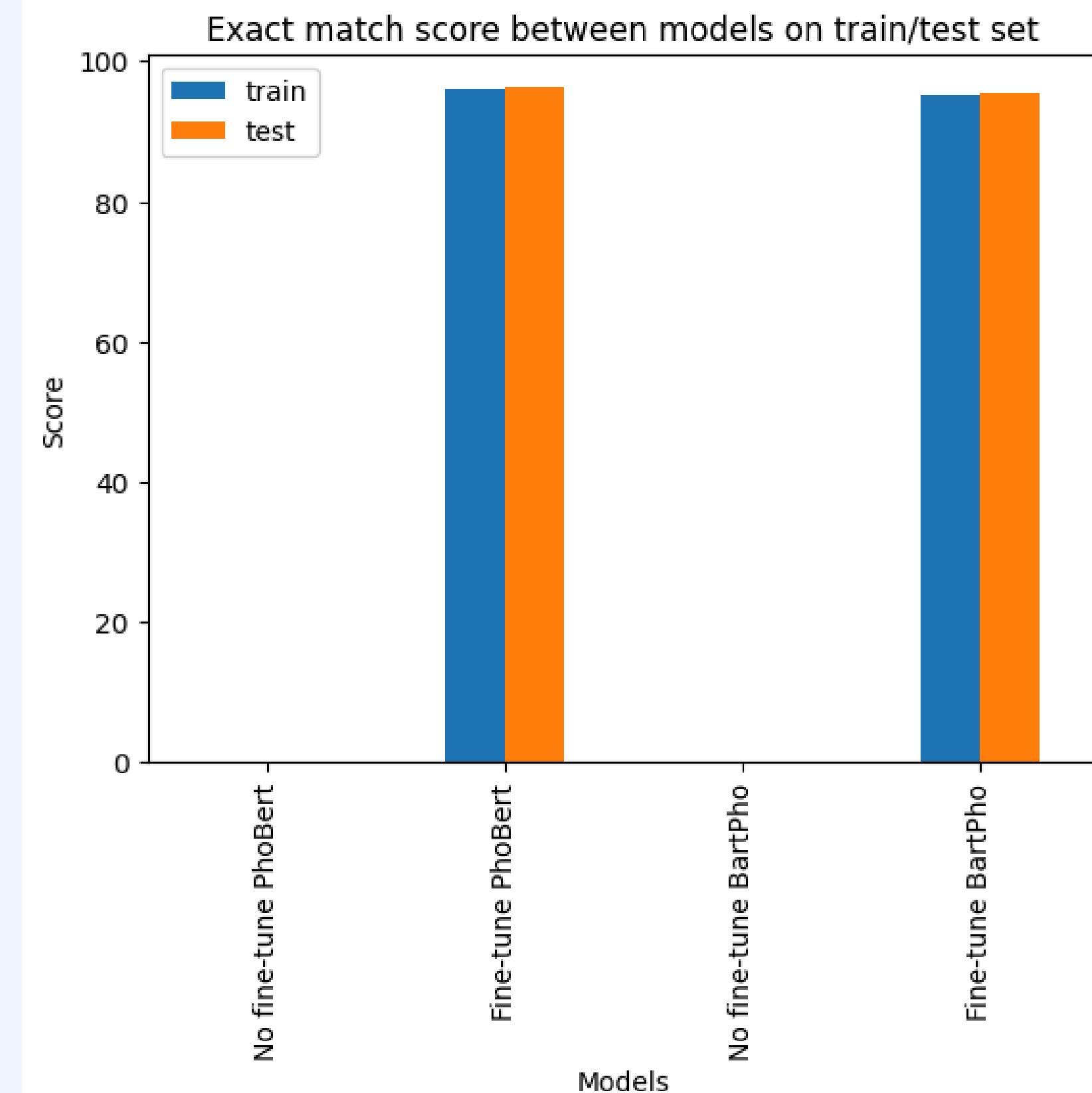


**2 model đều không bị
overfitting**

RESULT - EXACT MATCH

TÊN MÔ HÌNH THỬ NGHIỆM	Train Set	Test Set
No fine-tune PhoBERT	0.009991	0.0
Fine-tune PhoBERT	95.977620	96.154808
No fine-tune BARTPho	0.019982	0.037514
Fine-tune BARTPho	95.118393	95.360760

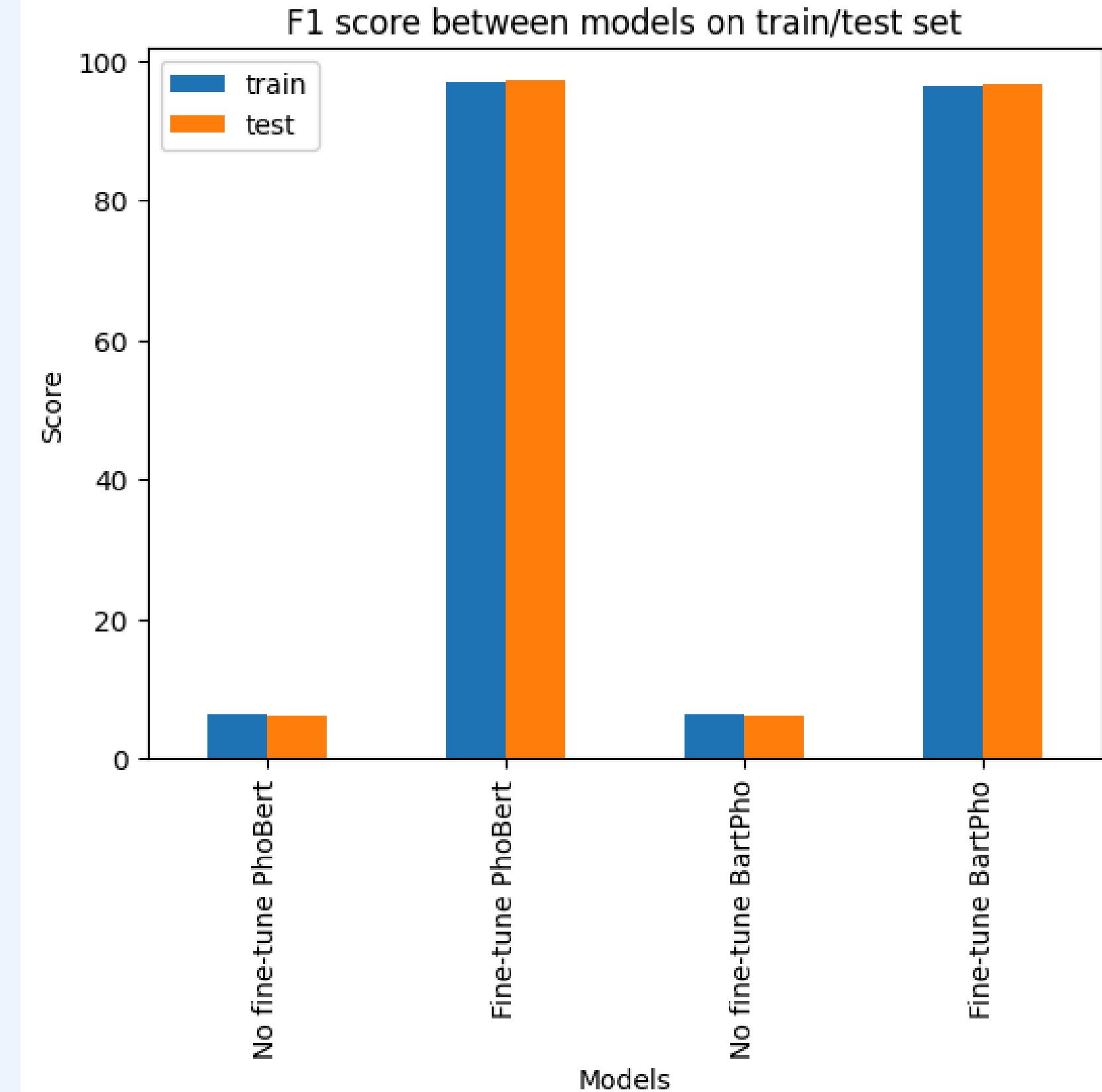
RESULT - Exact Match



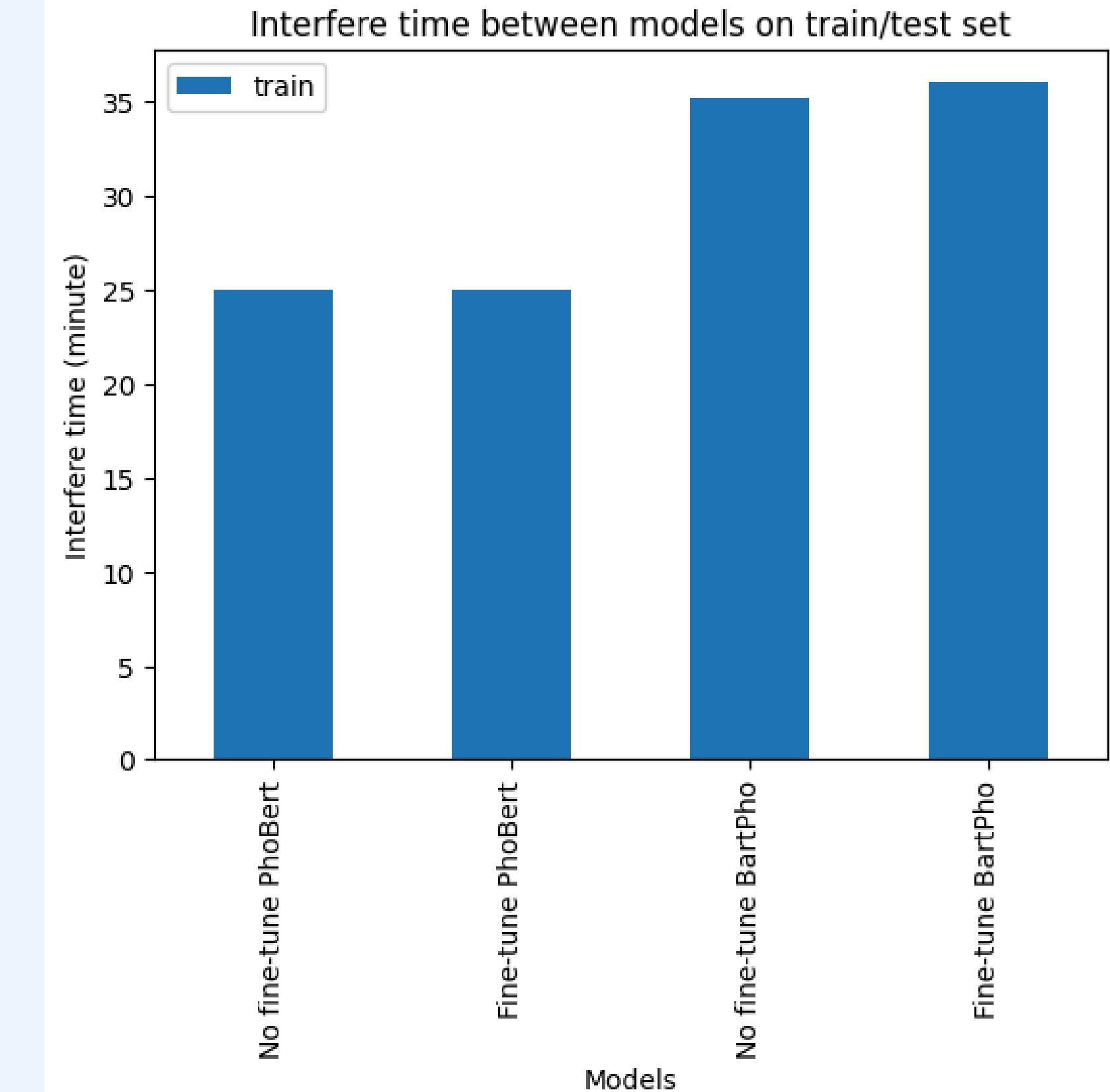
RESULT - F1 Score

TÊN MÔ HÌNH THỬ NGHIỆM	Train Set	Test Set
No fine-tune PhoBERT	6.471797	6.290414
Fine-tune PhoBERT	96.844422	97.015947
No fine-tune BARTPho	6.379219	6.305312
Fine-tune BARTPho	96.379627	96.527669

RESULT - F1 Score



RESULT - Interfere time



COMMENT

- ✓ Các model không được tinh chỉnh (fine-tune) đạt kết quả rất tệ trên cả 2 độ đo exact match hay f1-score. Điều này được lý giải bởi vì các lớp thực hiện tính toán đầu ra cho tác vụ downstream chưa được pretrained
- ✓ Các model được tinh chỉnh thì thu được kết quả rất tốt trên cả 2 độ đo [95-96.xx%] trên cả tập train và test.
- ✓ Thời gian suy luận của các model giống nhau là tương tự nhau dù có được tinh chỉnh hay không



fit@hcmus

VNUHCM - UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

COURSE: TEXT MINING
THEORY LECTURER: PhD. LE THANH TUNG
PRACTICE LECTURER: Ms. NGUYEN TRAN DUY MINH

**THANK YOU
FOR LISTENING**

GROUP:

MEMBERS:

1
20127258 - 20127655 - 20127625 - 20127597

PHƯỚC NGUYỄN - QUỐC TRUNG - HOÀNG THÁI - TẤN PHƯƠNG