

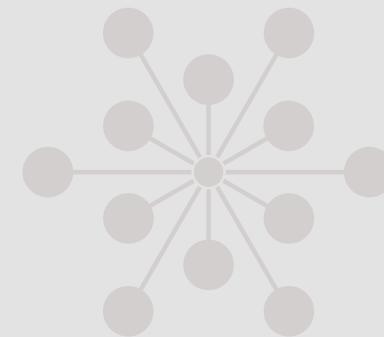
# Báo cáo nhập môn AI

Đề tài: Chatbot

Nguyễn Quốc Thái - 20225083

Phạm Hùng Phong - 20225060

# Nội dung báo cáo



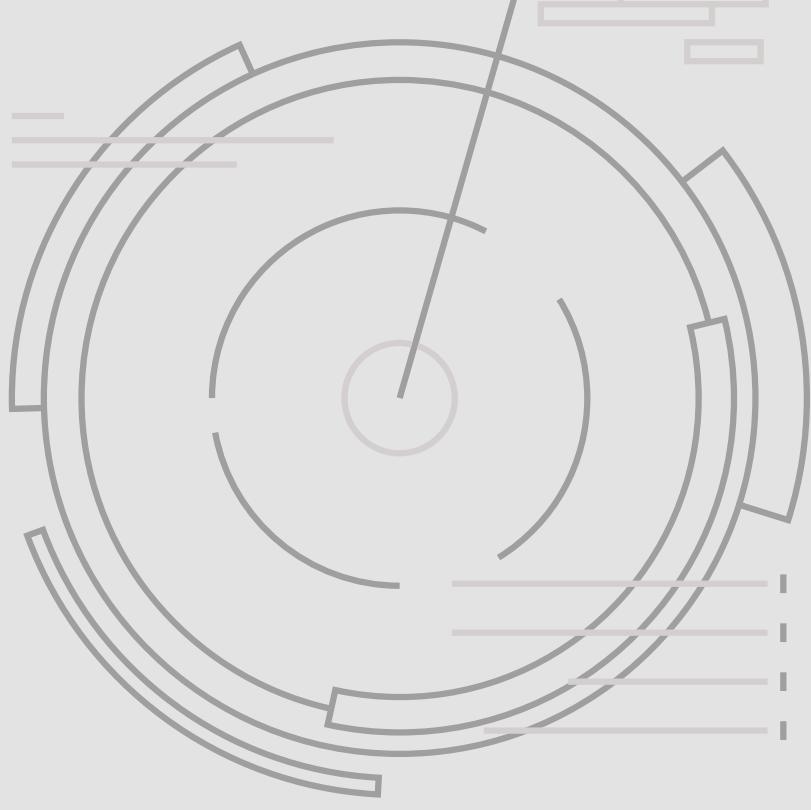
- 1 Giới thiệu bài toán
- 2 Phương pháp tiếp cận
- 3 Các bước thực hiện
- 4 Kết quả



01

# Giới thiệu bài toán

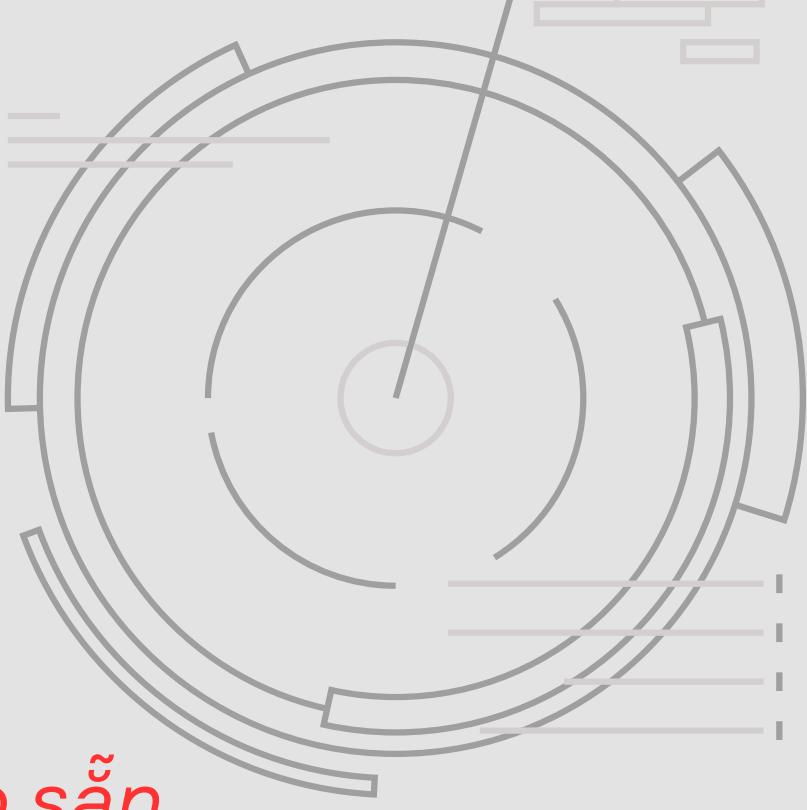




## Bài toán trả lời câu hỏi (Question Answering) là gì?

**Question Answering (QA):** một lĩnh vực trong trí tuệ nhân tạo (AI) tập trung vào việc phát triển các hệ thống máy tính có thể tự động trả lời các câu hỏi của con người một cách chính xác và đầy đủ thông tin.

## 2 dạng Question Answering chính



### **QA dựa trên truy xuất thông tin:**

- Tìm kiếm câu trả lời cho câu hỏi trong một *tập dữ liệu được cung cấp sẵn*
- Tập dữ liệu: tập câu hỏi - câu trả lời, cơ sở dữ liệu, trang web,...
- *Không có khả năng đưa ra câu trả lời nằm ngoài tập dữ liệu có sẵn*

### **QA dựa trên suy luận**

- Hệ thống có khả năng hiểu bối cảnh và ý nghĩa của câu hỏi
- *Có thể suy luận ra câu trả lời kể cả khi câu hỏi không được ghi trong tập dữ liệu*

# QA có những ứng dụng thực tế gì?



GPT (Generative Pre-trained Transformer)  
Được phát triển bởi OpenAI



Gemini  
Được phát triển bởi Google

# Ứng dụng của QA

## Dịch vụ khách hàng:

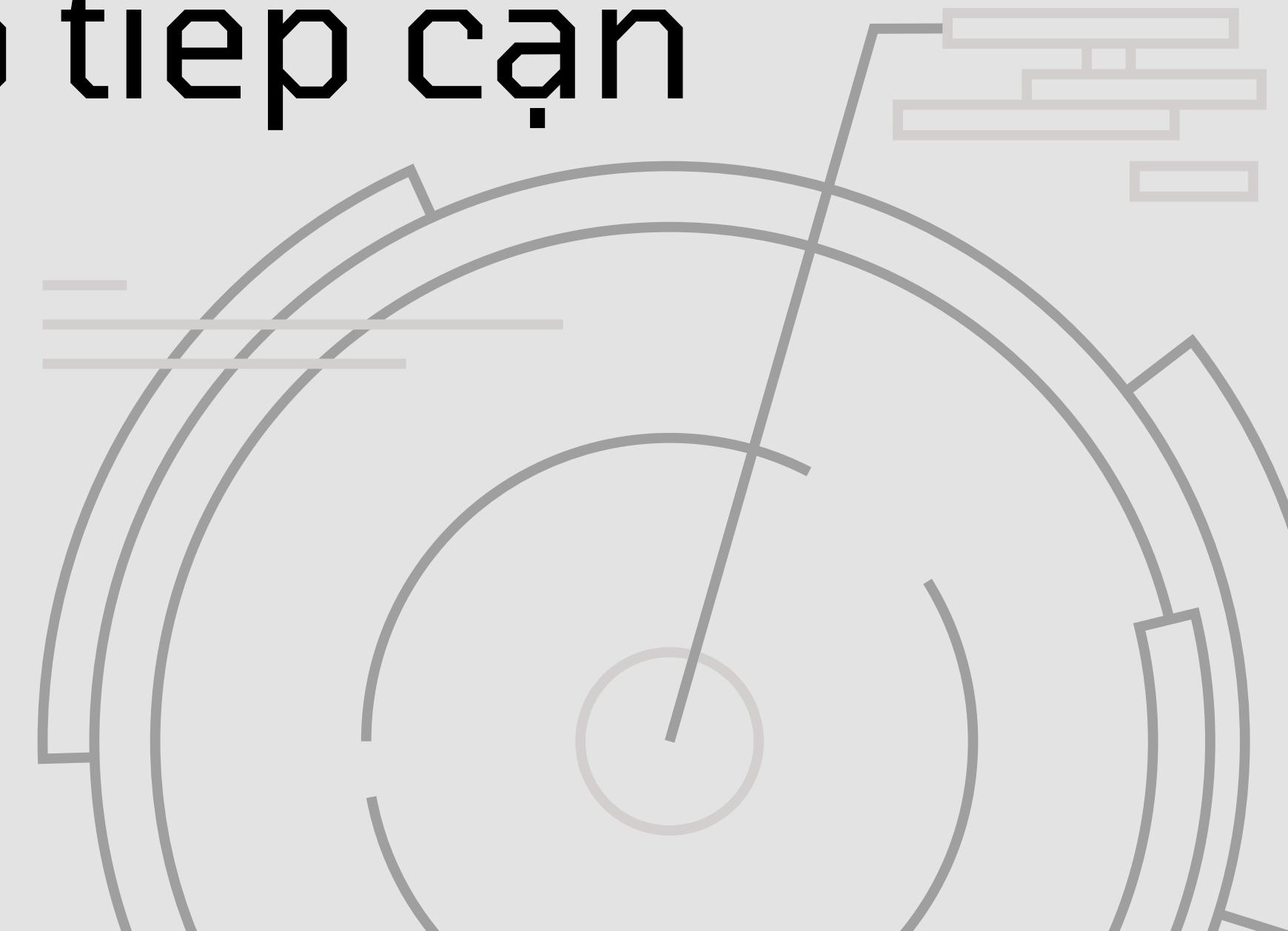
- Trả lời các câu hỏi thường gặp (FAQ) về sản phẩm hoặc dịch vụ
- Hỗ trợ khách hàng giải quyết vấn đề và cung cấp thông tin sản phẩm

## Giáo dục:

- Trả lời các câu hỏi về bài học, cung cấp giải thích bổ sung và tạo các bài tập luyện tập.

02

## Phương pháp tiếp cận



## Lựa chọn mô hình phù hợp và cài đặt mô hình

Các mô hình cho bài toán QA phổ biến hiện nay:

- GPT 3.5 (Generative Pre-training Transformer)
- BERT (Bidirectional Encoder Representations from Transformers)
- BART (Bidirectional Autoregressive Transformer)

Đều là mô hình **Transformer**

# So sánh 3 mô hình

Tính năng	GPT	BERT	BART
Kiểu mô hình	Mã hóa trước	Mã hóa hai chiều	Tái tạo tự động hai chiều
Mục tiêu huấn luyện	Tạo văn bản	Hiểu ngữ cảnh	Tạo văn bản và hiểu ngữ cảnh
Ứng dụng	Chatbot, tạo văn bản tự động, hỗ trợ sáng tạo	Phân loại văn bản, tóm tắt văn bản, trả lời câu hỏi, trích xuất thông tin	Tóm tắt văn bản, trả lời câu hỏi, tạo văn bản sáng tạo
Điểm mạnh	Tạo văn bản thực tế và mạch lạc	Hiểu mối quan hệ ngữ cảnh tốt	Cân bằng giữa tạo văn bản và hiểu ngữ cảnh
Điểm yếu	Có thể tạo ra văn bản không chính xác hoặc không phù hợp	Yêu cầu nhiều dữ liệu huấn luyện	Yêu cầu nhiều dữ liệu huấn luyện hơn GPT

# BARTPho

- Được nghiên cứu và phát triển bởi VinAI Research
- Sử dụng kiến trúc **Transformer**
- **Dành riêng cho việc xử lý Tiếng Việt**, có khả năng xử lý và tạo ra văn bản Tiếng Việt một cách hiệu quả và chính xác
- Được huấn luyện trên một tập dữ liệu khổng lồ bằng tiếng Việt (sách báo, wikipedia, mạng xã hội,...)
- Có 2 phiên bản: BARTPho-Word và BARTPho-Syllable



## BARTPho-Word và BARTPho-Syllable

Sự khác biệt giữa BARTPho-Word và BARTPho-Syllable là về cách chúng nhận dữ liệu đầu vào.

- **BARTPho Word:** Hiểu nghĩa chung của câu và đoạn văn, tạo ra văn bản trôi chảy và tự nhiên.
- **BARTPho Syllable:** Hiểu chi tiết âm tiết tiếng Việt, xử lý các hiện tượng ngữ âm Tiếng Việt phức tạp

Ví dụ:

A 5-syllable text: Tôi là một học sinh  
forms a 4-word text: Tôi<sub>I</sub> là<sub>am</sub> một<sub>a</sub> học<sub>\_</sub>sinh<sub>student</sub>.

## Thu thập training dataset

Dữ liệu để huấn luyện mô hình (training dataset) được thu thập trên trang web [soict.hust.edu.vn](http://soict.hust.edu.vn)

- Các thông tin trên trang web được tóm tắt thành một đoạn văn (context)
- Làm các bộ câu hỏi và câu trả lời về đoạn văn đó
- Dữ liệu sau đó được đưa thành file JSON với định dạng dataset chuẩn cho các bước sau

# Thu thập training dataset

Trang chủ > PGS.TS. Lê Thanh Hương

PGS.TS. Lê Thanh Hương - SoICT

soict.hust.edu.vn/pgs-ts-le-thanh-huong.html

**GIỚI THIỆU**

PGS. Lê Thanh Hương là PGS. TS. Khoa Khoa học máy tính của Trường Công nghệ Thông tin và Truyền thông, Đại học Bách khoa Hà Nội. Cô nhận bằng Tiến sĩ Khoa học Máy tính năm 2004 tại trường Đại học Middlesex, Vương quốc Anh. Các hướng nghiên cứu hiện nay của cô bao gồm trích rút thông tin, hồi đáp, tóm tắt văn bản, dịch máy, phát hiện và sửa lỗi chính tả, kiểm tra văn bản, phân tích quan điểm, hệ gợi ý... Cô hiện là phó trưởng Khoa Khoa học máy tính, trưởng nhóm nghiên cứu Xử lý ngôn ngữ tự nhiên, thuộc Trung tâm nghiên cứu quốc tế về Trí tuệ nhân tạo (BKAI).

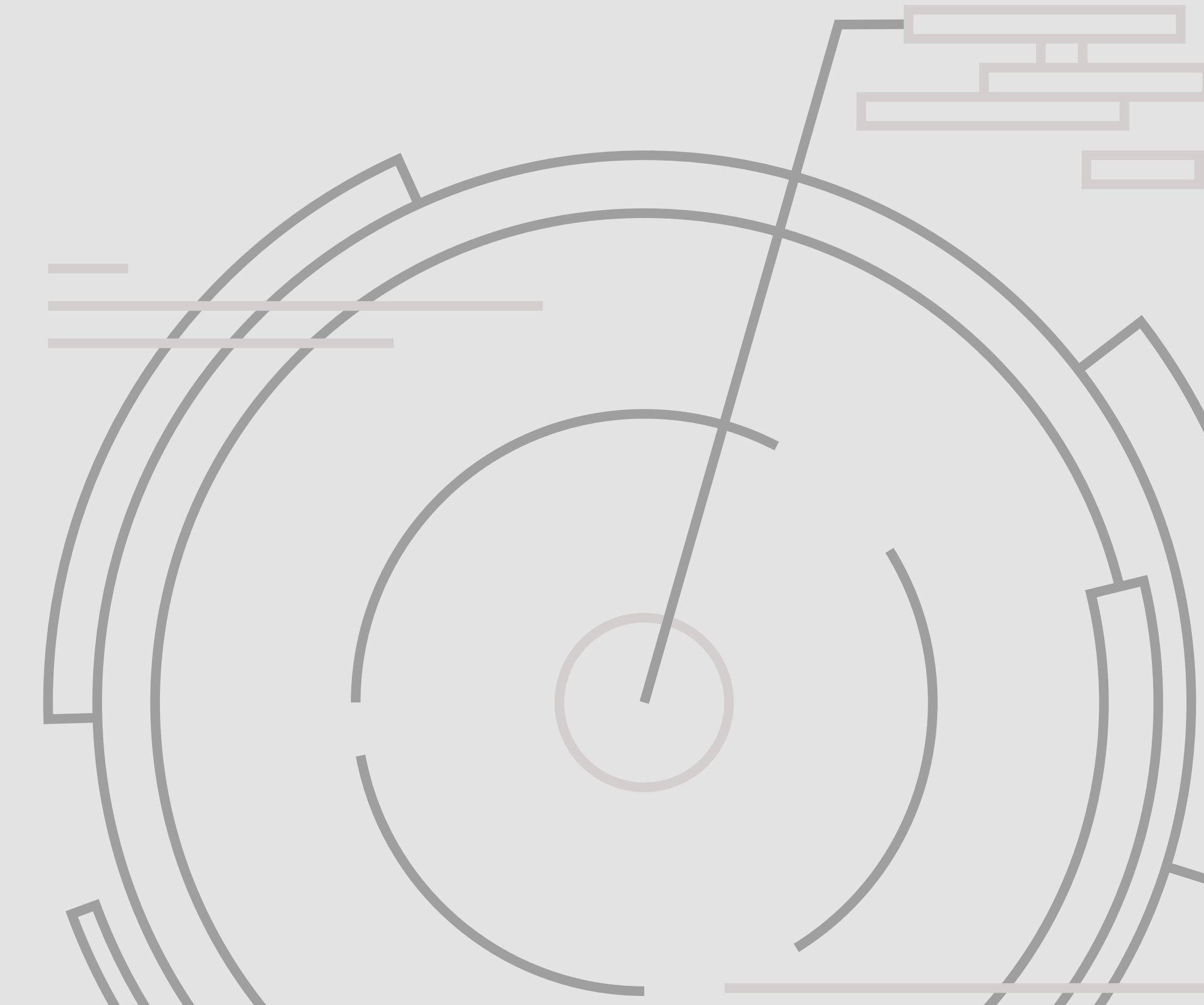
**CÁC CÔNG TRÌNH KHOA HỌC TIÊU BIỂU**

- Thai Nguyen Quoc, **Huong Le Thanh**, Hanh Pham Van (2023). Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies. *Informatica* (Slovenia). <https://doi.org/10.31449/inf.v47i3.4761>
- Tuan Luu Minh, **Huong Le Thanh**, Tan Hoang Minh (2021). A hybrid model using the pre-trained BERT and deep neural networks with rich feature for extractive text summarization. *Journal of Computer Science and Cybernetics*. Volume 37, No.2, pp. 123-143.
- Nguyen Van Son, **Le Thanh Huong**, Nguyen Chi Thanh (2021). A two-phase plagiarism detection system based on multi-layer LSTM Networks. *IAES International Journal of Artificial Intelligence*.
- Minh-Tuan Luu, **Thanh-Huong Le**, Minh-Tan Hoang (2021). An effective deep learning approach for extractive text summarization. *Indian Journal of Computer Science and Engineering*. Vol. 12, No. 2, 2021, e-ISSN:0976-5166, p-ISSN:2231-3850, pp.434-444.
- Anh Phan Duy, **Huong Le Thanh** (2023). A Question-Answering System for Vietnamese Public Administrative Services. *SOICT '23: Proceedings of the 12th International Symposium on Information and Communication Technology*. <https://doi.org/10.1145/3628797.3628965>
- Quang Nhat Nguyen and **Huong Thanh Le** (2023). Building an Efficient Retriever System with Limited

```
"title": "PGS. TS. Lê Thanh Hương",
"paragraphs": [
{
  "context": "PGS. TS. Lê Thanh Hương hiện là Phó Trưởng Khoa Khoa học máy tính tại Trường Công nghệ Thông tin và Truyền thông, Đại học Bách khoa Hà Nội. Cô nhận bằng Tiến sĩ Khoa học Máy tính từ Đại học Middlesex, Vương quốc Anh vào năm 2004, Thạc sĩ Khoa học Máy tính từ Đại học Tự do Brussels và",
  "qas": [
    {
      "question": "PGS. TS. Lê Thanh Hương hiện giữ chức vụ gì tại Đại học Bách khoa Hà Nội?",
      "id": "97-01",
      "answers": [
        {
          "text": "PGS. TS. Lê Thanh Hương hiện là Phó Trưởng Khoa Khoa học máy tính tại Trường Công nghệ Thông tin và Truyền thông, Đại học Bách khoa Hà Nội.",
          "answer_start": 0
        }
      ]
    },
    {
      "question": "PGS. TS. Lê Thanh Hương đã nhận các bằng cấp gì và từ đâu?",
      "id": "97-02",
      "answers": [
        {
          "text": "Cô nhận bằng Tiến sĩ Khoa học Máy tính từ Đại học Middlesex, Vương quốc Anh vào năm 2004, Thạc sĩ Khoa học Máy tính từ Đại học Tự do Brussels và",
          "answer_start": 140
        }
      ]
    }
  ]
}
```

03

## Các bước thực hiện



## Các bước thực hiện bài toán QA trích xuất

1. Chuẩn bị dữ liệu
2. Lựa chọn mô hình phù hợp và cài đặt mô hình
3. Huấn luyện mô hình
4. Triển khai mô hình lên API

## 1. Chuẩn bị dữ liệu

Dữ liệu có cấu trúc SQuAD (Stanford Question Answering Dataset)

- Gồm các đoạn văn bản (**context**)
- Các cặp câu hỏi, câu trả lời được trích xuất từ văn bản

Dữ liệu được thu thập trên trang [soict.hust.edu.vn](http://soict.hust.edu.vn)

# 1. Chuẩn bị dữ liệu

- **Context:** Đoạn thông tin, được lấy từ [soict.hust.edu.vn](http://soict.hust.edu.vn)
- **Question:** Câu hỏi
- **Id:** Id của câu hỏi
- **answer\_start:** vị trí bắt đầu của câu trả lời trong đoạn văn
- **is\_impossible:** xác định câu hỏi trên có thể trả lời được không

```
[{"data": [{"title": "Phạm Đăng Hải", "paragraphs": [{"context": "Ông Phạm Đăng Hải là Tiến sĩ Khoa học Máy tính, hiện đang giữ chức vụ Giảng viên chính và Trưởng khoa Khoa", "qas": [{"question": "Phạm Đăng Hải hiện giữ chức vụ gì tại trường Đại học Bách khoa Hà Nội?", "id": "uit_00001", "answers": [{"answer_start": 71, "text": "Giảng viên chính và Trưởng khoa Khoa học Máy tính"}], "is_impossible": false}, {"question": "Ông Phạm Đăng Hải hoàn thành bằng Tiến sĩ vào năm nào?", "id": "uit_00002", "answers": [{"answer_start": 287, "text": "2010"}], "is_impossible": false}]}]}]
```

# Chuẩn bị dữ liệu

- **Tập hợp dữ liệu:** liên quan đến thầy cô, bao gồm các câu hỏi thường gặp và thông tin chi tiết về giáo viên, các môn giảng dạy, giải thưởng, ...
- **Làm sạch dữ liệu:** loại bỏ nhiễu (chuẩn hóa syntax ,...)
- **Tokenization:** chia văn bản thành các **token** nhỏ hơn ( từ, cụm từ,... )
- **Vectorization:** đưa văn bản về số hoặc vector cho mô hình xử lý và phân tích dữ liệu

# Token hóa dữ liệu



'<s> Phạm Đăng Hải hiện giữ chức vụ gì tại trường Đại học Bách khoa Hà Nội?</s></s> Ông Phạm Đăng Hải là Tiến sĩ Khoa học Máy tính, hiện đang giữ chức vụ Giảng viên chính và Trưởng khoa Khoa học Máy tính tại Viện Công nghệ thông tin và truyền thông, trường Đại học Bách khoa Hà Nội. Ông hoàn thành bằng Tiến sĩ tại trường Thực hành Công nghệ cao ở Cộng hòa Pháp vào năm 2010, Thạc sĩ tại Viện tin học Pháp ngữ năm 1997 và Kỹ sư tại Đại học Bách Khoa Hà Nội năm 1995. Ông có một danh sách dài các công trình nghiên cứu khoa học tiêu biểu, trong đó có các bài báo được công bố tại các hội nghị quốc tế như HPCS 2009 và IEEE-RIVF 09, chủ yếu tập trung vào các chủ đề như tính toán hiệu năng cao, mô phỏng song song và phân tán, và hệ điều hành nhúng. Ngoài các công trình nghiên cứu, ông còn tham gia giảng dạy nhiều khóa học tại trường, bao gồm Tin học đại cương (IT1110), Hệ điều hành (IT3070), Lý thuyết Ngôn ngữ và Phương pháp dịch (IT4074), và Thực hành Xây dựng chương trình dịch (IT4182). Ông Phạm...'

<s> Phạm Đăng Hải hiện giữ chức vụ gì tại trường Đại học Bách khoa Hà Nội?</s></s> Ông Phạm Đăng Hải là Tiến sĩ Khoa học Máy tính, hiện đang giữ chức vụ Giảng viên chính và Trưởng khoa Khoa học Máy tính tại Viện Công nghệ thông tin và truyền thông, trường Đại học Bách khoa Hà Nội. Ông hoàn thành bằng Tiến sĩ tại Viện tin học Pháp ngữ năm 1997 và Kỹ sư tại Đại học Bách Khoa Hà Nội năm 1995. Ông có một danh sách dài các công trình nghiên cứu khoa học tiêu biểu, trong đó có các bài báo được công bố tại các hội nghị quốc tế như HPCS 2009 và IEEE-RIVF 09, chủ yếu tập trung vào các chủ đề như tính toán hiệu năng cao, mô phỏng song song và phân tán, và hệ điều hành nhúng. Ngoài các công trình nghiên cứu, ông còn tham gia giảng dạy nhiều khóa học tại trường, bao gồm Tin học đại cương (IT1110), Hệ điều hành (IT3070), Lý thuyết Ngôn ngữ và Phương pháp dịch (IT4074), và Thực hành Xây dựng chương trình dịch (IT4182). Ông Phạm Đăng Hải là một nhà giáo dục chuyên nghiệp, luôn nỗ lực khích lệ sinh viên và chia sẻ kiến thức và kinh nghiệm của mình với thế hệ kỹ sư tiếp theo. Ông

# Lựa chọn mô hình và cài đặt các thư viện cần thiết

1. Mô hình: BARTPho Syllable

2. Các thư viện cần thiết:

- Transformer: cung cấp các mô hình Transformers như BERT, BART, GPT
- datasets: tải và xử lý các bộ datasets
- Torch: cung cấp các công cụ mạnh mẽ và linh hoạt để xây dựng và huấn luyện các mô hình học sâu
- Tải model BARTPho từ thư viện Transformers
- Cài đặt các siêu tham số (hyperparameters) để huấn luyện

## Cài đặt cấu hình tham số

- **batch\_size**: Số lượng mẫu được xử lý cùng một lúc trong quá trình huấn luyện
- **epochs**: Số lần toàn bộ dữ liệu được sử dụng trong quá trình huấn luyện
- **learning\_rate**: Tốc độ cập nhật tham số trong quá trình huấn luyện
- **max\_length**: Độ dài tối đa của đoạn văn
- **max\_answer\_length**: Độ dài tối đa của câu trả lời dự đoán của mô hình
- **n\_best**

## Cài đặt cấu hình tham số

- batch\_size = 2
- epochs = 15
- learning\_rate = 0.001
- max\_length = 1024
- max\_answer\_length = 200
- n\_best = 20

## Mô hình sẽ đánh giá như thế nào?

- Đoạn văn đầu vào (context) sẽ được chia thành các token
- Khi có một câu hỏi được nhập vào, mô hình sẽ tính khả năng token đó là vị trí bắt đầu hoặc kết thúc của câu trả lời, được gọi là chỉ số logits
- Lấy ra một lượng n\_best các token có logits cao nhất (ở trường hợp này là 20 token)
- Tính điểm các cặp có tổng logit có thể được tạo ra từ 20 token đó
- Chọn cặp token có tổng điểm logit cao nhất

# Mô hình sẽ đánh giá như thế nào?

- Context: “Tôi là học sinh, tôi tên là Thái”, câu hỏi: “Bạn tên là gì?”
- Giả sử đoạn văn có 9 token và mô hình dự đoán các giá trị logit cho vị trí bắt đầu và kết thúc như sau:

Start logits:

- "Tôi": 0.2
- "là": 0.1
- "học": 0.1
- "sinh": 0.1
- ",": 0.1
- "tôi": 0.3
- "tên": 0.8
- "là": 0.6
- "Thái": 2.0

End logits:

- "Tôi": 0.1
- "là": 0.1
- "học": 0.1
- "sinh": 0.1
- ",": 0.1
- "tôi": 0.2
- "tên": 0.6
- "là": 0.7
- "Thái": 1.8



# Mô hình sẽ đánh giá như thế nào?

- Trong quá trình suy luận, mô hình sẽ chọn ra n\_best logits cao nhất cho cả start và end. Giả sử **n\_best = 2**:

Start logits cao nhất:

- "Thái" (logit = 2.0)
- "tên" (logit = 0.8)

End logits cao nhất:

- "Thái" (logit = 1.8)
- "là" (logit = 0.7)



## Mô hình sẽ đánh giá như thế nào?

- Mô hình sẽ tạo ra tất cả các cặp có thể từ n\_best start và end logits, sau đó tính tổng logits cho mỗi cặp:

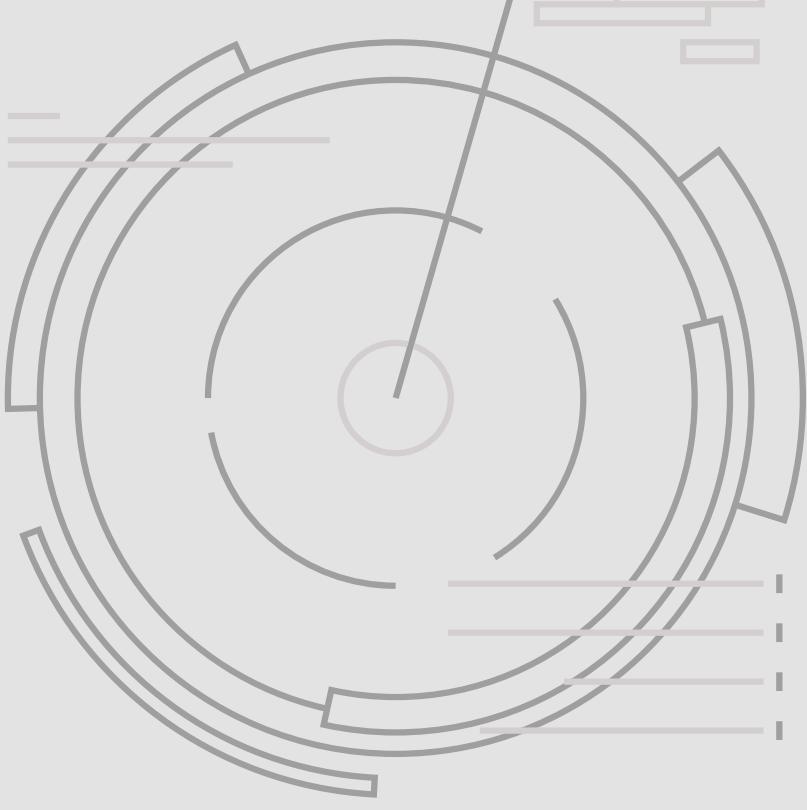
("Thái", "Thái"):  $2.0 + 1.8 = 3.8$

("Thái", "là"):  $2.0 + 0.7 = 2.7$

("tên", "Thái"):  $0.8 + 1.8 = 2.6$

("tên", "là"):  $0.8 + 0.7 = 1.5$

→ Mô hình sẽ chọn ra cặp (start\_token, end\_token) là ("Thái", "Thái")  
Câu trả lời sẽ là Thái



# Triển khai mô hình lên API sử dụng Flask

Chatbot SOICT

cô Lê Thanh Hương dạy các môn gì?

Trong giảng dạy, PGS. TS. Lê Thanh Hương phụ trách các môn học như IT3160: Introduction to Artificial Intelligence, IT4772: Natural Language Processing, và IT2000: Introduction to Information and Communication Technology. Thông

cô Lê Thanh Hương quan tâm đến các nghiên cứu gì?

Bà đặc biệt quan tâm đến các nghiên cứu về trích rút thông tin, hệ thống hỏi đáp, tóm tắt văn bản, dịch máy, phát hiện và sửa lỗi chính tả, kiểm trùng văn bản, và các ứng dụng của trí tuệ nhân tạo như hệ gợi ý và hệ chuyên

Type a message...

Send



04

## Kết quả





Chatbot SOICT

thầy Ngô Văn Linh là ai?

giảng viên tại Khoa Khoa học Máy tính, Đại học Bách khoa Hà Nội.

thầy Ngô Văn Linh nhận bằng tiến sĩ vào năm nào?

năm 2023,

Lĩnh vực nghiên cứu của thầy Ngô Văn Linh bao gồm gì?

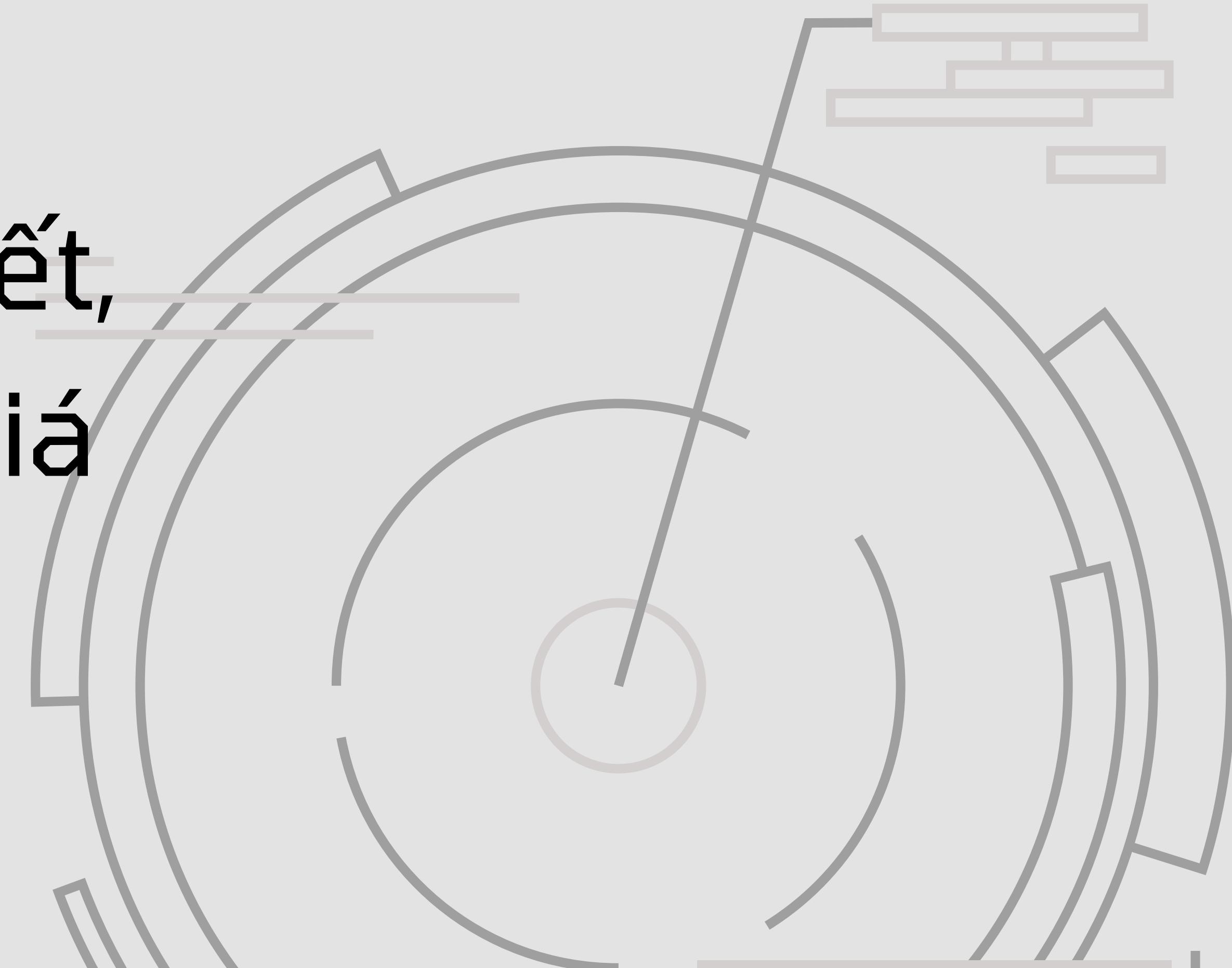
học máy, khai phá dữ liệu, hệ gợi ý, xử lý ngôn ngữ tự nhiên, mô hình chủ đề, và suy diễn biến phân.

Type a message...

Send

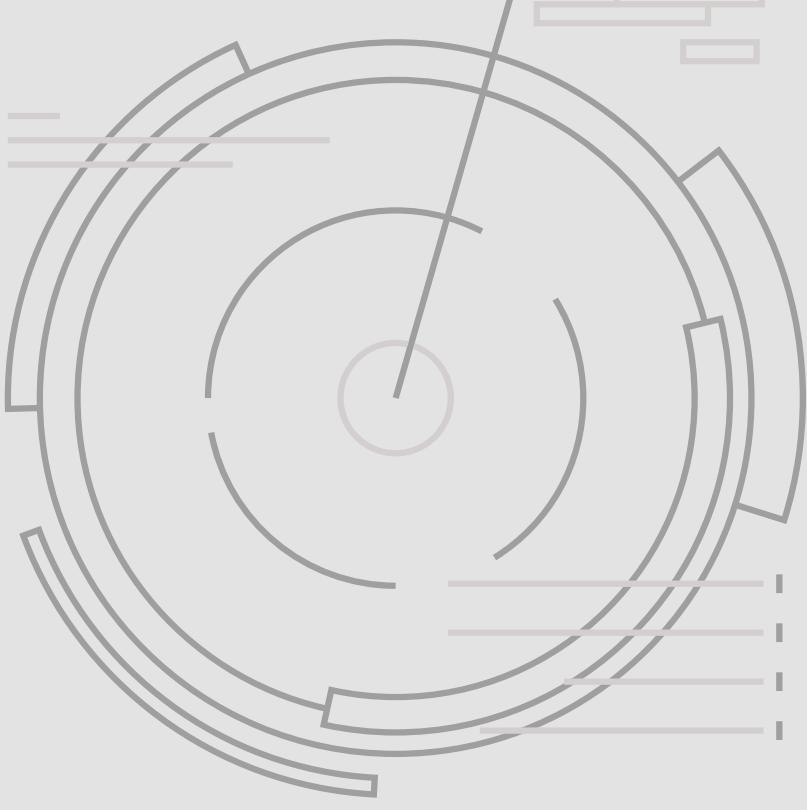
05

Tổng kết,  
đánh giá



## Những khó khăn nhóm em đã gặp phải

- Nhóm em chưa xử lý được cách để trả lời những câu hỏi nằm ngoài bộ dữ liệu có sẵn
- Bộ dữ liệu gồm các câu hỏi khá tương đồng, có thể dẫn đến hiện tượng overfitting



## Nếu có cơ hội, bọn em sẽ

- Mở rộng bộ dataset với dữ liệu về trường (lịch sử trường, các khoa, các ngành, ...)
- Triển khai lên một website công cộng để mọi người có thể truy cập được
- Xử lí bộ dữ liệu để không xảy ra vấn đề như overfitting và bias về một số thông tin.



# THANK YOU

