**Pretrained và fine-tuning GPT2 model xây dựng chatbot**

**1. Dữ liệu**

Dữ liệu: khoảng 3000 dòng text về các cuộc hội thoại từ các nhóm chat.

Cấu trúc dữ liệu:

A screenshot of a computer program

Description automatically generated

Mỗi dialog là một conversation giữa các senders.

Ta trích ra các đoạn text (khoảng 3500 dòng) thành tập dữ liệu để train.

**2. Mô hình**

Mô hình: [GPT2LMHeadModel](https://huggingface.co/docs/transformers/v4.40.1/en/model_doc/gpt2#transformers.GPT2LMHeadModel)

Đây là mô hình pretrained text generation. Ta sẽ sử dụng để train bộ dữ liệu trên.

**2.1 Xử lý dữ liệu đầu vào**

Ta sử dụng GPT2Tokenizer để tokenize dữ liệu đầu vào với tham số:

* max\_length: 40
* truncation: True
* padding: “max\_length”
* return\_tensors: “pt”

**2.2 Training**

A screen shot of a computer program

Description automatically generated

Ta huấn luyện mô hình với các tham số:

* epochs: 12
* optimize: Adam với learning\_rate=1e-3

Quá trình huấn luyện:

A black background with white lines

Description automatically generated with medium confidence

Kết quả thử nghiệm:

A screenshot of a computer screen

Description automatically generated