**REPORT NLP TRAINING LAB**

**Name:** Nguyễn Thạch Hà Anh

**QUESTION 1:** Build a model for doing the Generative Question Answering task. Specifically, given a question and support documents related to the given question, the model will output the answer.

* **Framework:** Pytorch v4.30.0
* **Backbone:** BART

1. **The completed process:**
2. **Load and Prepare Dataset:**

* Dataset: ELI5 (ELI5.jsonl and ELI5\_val.jsonl).
* Dataset in ELI5.jsonl include:
  + question\_id
  + question
  + answers (list)
  + ctxs (contexts) (list)
* Dataset in ELI5\_val.jsonl is similar, but it includes scores (representing the relevance and confidence of contexts).
* Using 1000 first data **(Because Data have 272634 data On Embeddings and Upsert: Problems of Running. Result fail to connect again)**

1. **Pinecone service:** Allows us to interact with the index to add and retrieve vector representations.
2. **Retriever:** Using SentenceTransformer model (lax-sentence-embeddings/all\_datasets\_v3\_mpnet-base) of MPNet
3. **Embeddings and Upsert:** All the text passages in the dataset have been embedded and they have been updated or added to the Pinecone index.
4. **Generator:** loading the BART model and tokenizer.

**In Testing:** Contexts not match with Question. (State: Until now, I have fixed it. And being research more about that errors and excited with this NLP lab while doing it).