NLP - Competition

VLSP-2023 INSTRUCTION

AI VIET NAM
Nguyen Quoc Thai

CONTENT

1 – Legal Textual Entailment Recognition

Description

- Given a set of statement (a assume S is a statement) and a set of legal passages (L_1, L_2,…, L_N)
- Goal: check the set of legal passages entails statement S


```
[
  {
    "example_id": "DS-101",
    "label": "Yes/No",
    "statement": "Cơ sở điện ảnh phát hành phim phải chịu trách nhiệm trước pháp luật về nội dung phim phát hành là sai.",
    "legal_passages": [
      {
        "type": "law",
        "law_id": "05/2022/QH15",
        "article_id": "15"
      }
    ]
  }
]
```


Training Data Format


```
[
  {
    "example_id": "DS-101",
    "label": "Yes/No",
  }
]
```


Prediction Format

Source: https://vlsp.org.vn/vlsp2023/eval/lter

AI VIET NAM
@aivietnam.edu.vn

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

Stage 1: Retriever
Find the most relevant passage to the statement


```
[
  {
    "example_id": "DS-101",
    "label": "Yes/No",
    "statement": "Cơ sở điện ảnh phát hành phim phải chịu trách nhiệm
trước pháp luật về nội dung phim phát hành là sai.",
    "legal_passages": [
      {
        "type": "law",
        "law_id": "05/2022/QH15",
        "article_id": "15"
      }
    ]
  }
]
```

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

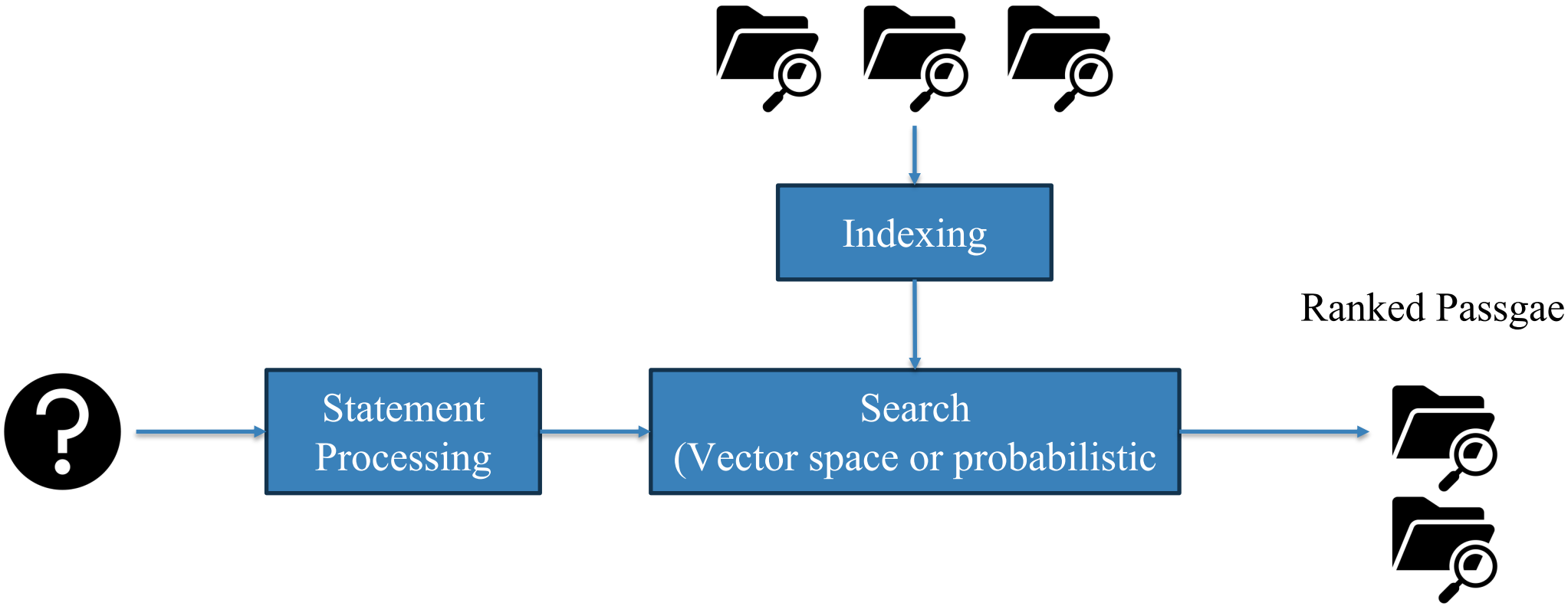

```
[
  {
    "example_id": "DS-101",
    "label": "Yes/No",
    "statement": "Cơ sở điện ảnh phát hành phim phải chịu trách nhiệm
trước pháp luật về nội dung phim phát hành là sai.",
    "legal_passages": [
      {
        "type": "law",
        "law_id": "05/2022/QH15",
        "article_id": "15"
      }
    ]
  }
]
```


Stage 2: Classifier

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 1: Retriever (Dense Passage Retrieval)



1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 1: Retriever (Dense Passage Retrieval)
- BM25

Common words less important

Repetitions of query words => good

$$\text{BM25}(D,Q) = \sum_{i=1}^{n} IDF(q_i, D) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i) + k_1 \cdot (1 - b + b \cdot (|D|/d_{avg}))}$$

More words in common with the query => good

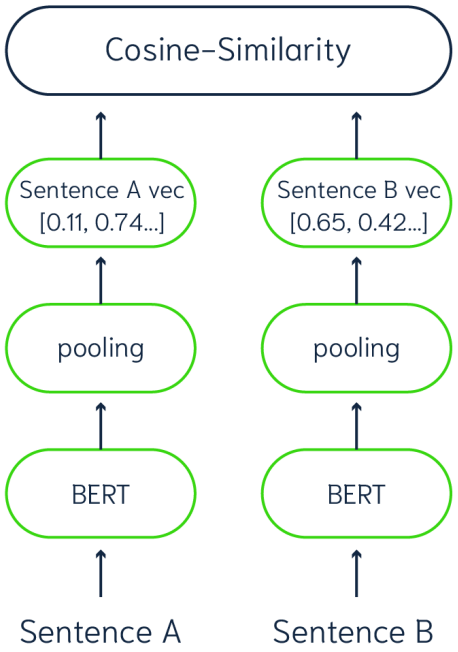
Repetitions less important than different query words

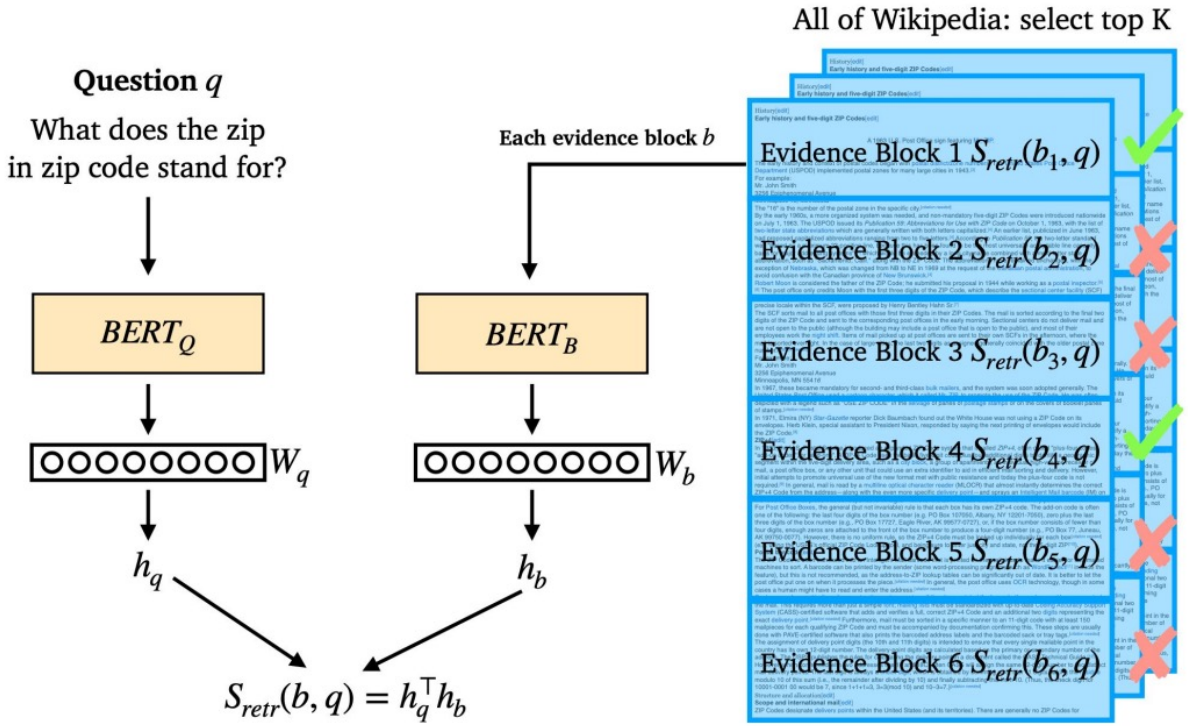
But more important if document is relatively long (average)

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 1: Retriever (Dense Passage Retrieval)
- Bi-Encoder



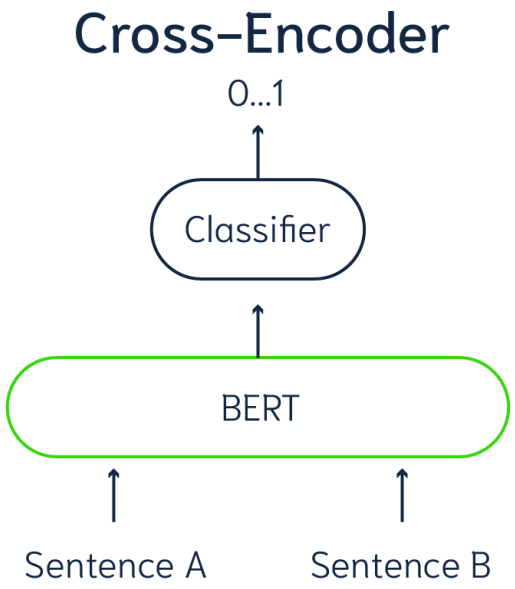


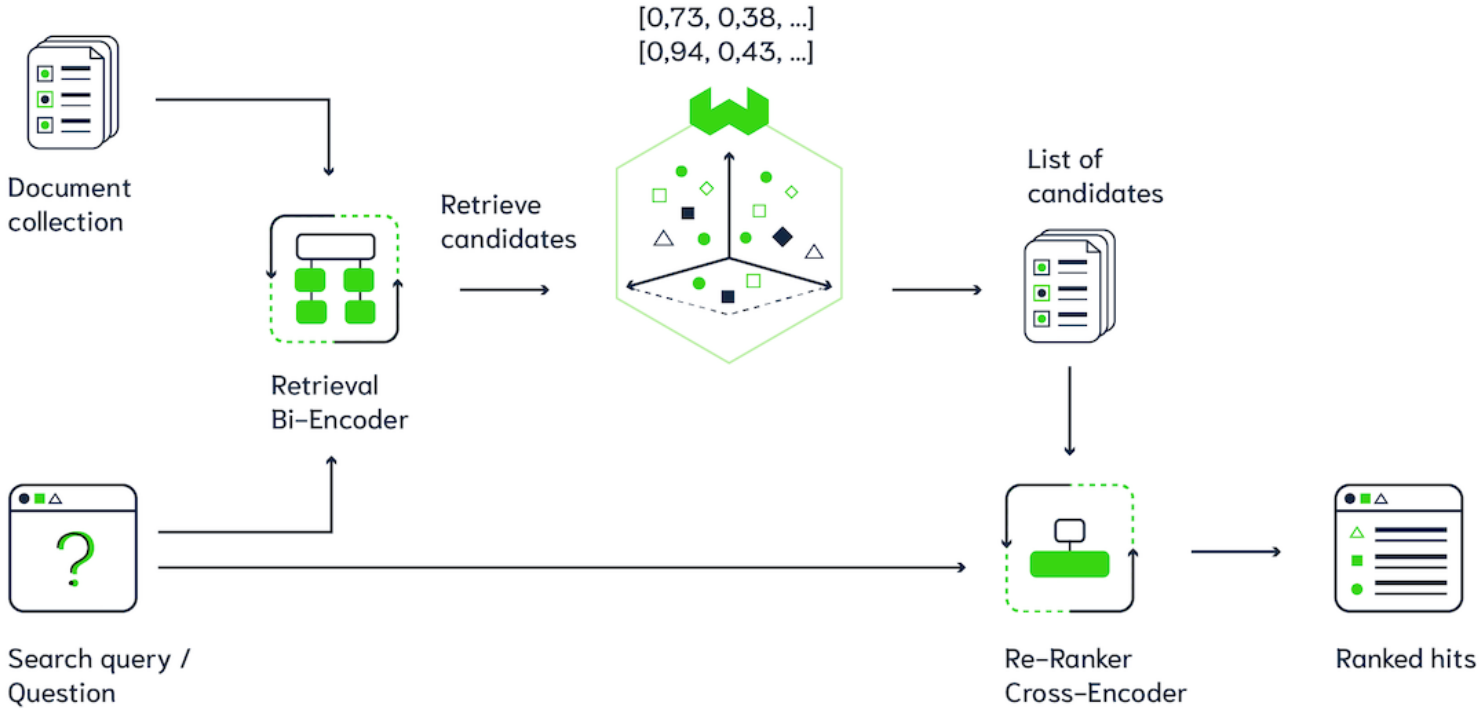
Source https://github.com/facebookresearch/DPR

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 1: Retriever (Dense Passage Retrieval)
- Cross-Encoder



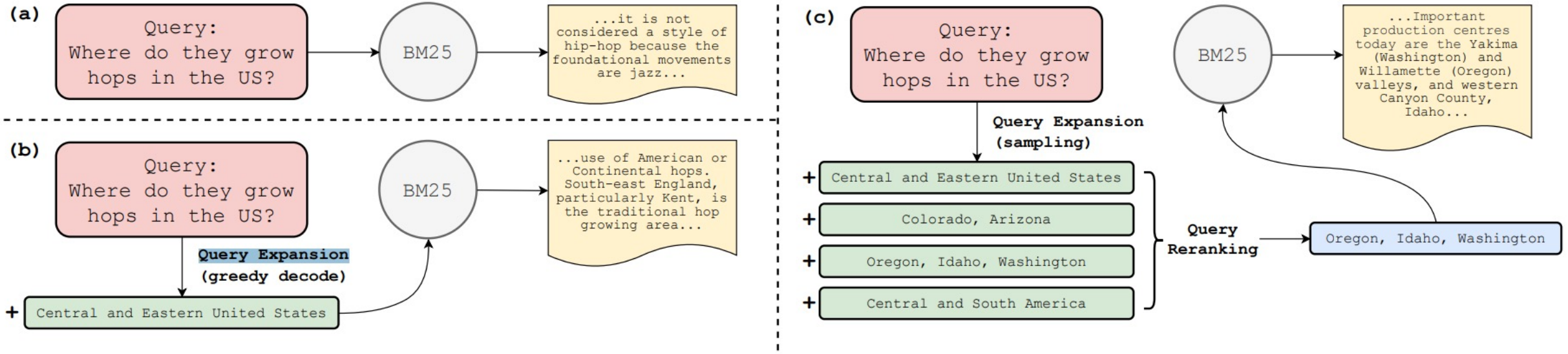


Source

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 1: Retriever (Dense Passage Retrieval)
- Expand, Rerank and Retrieve (EAR)



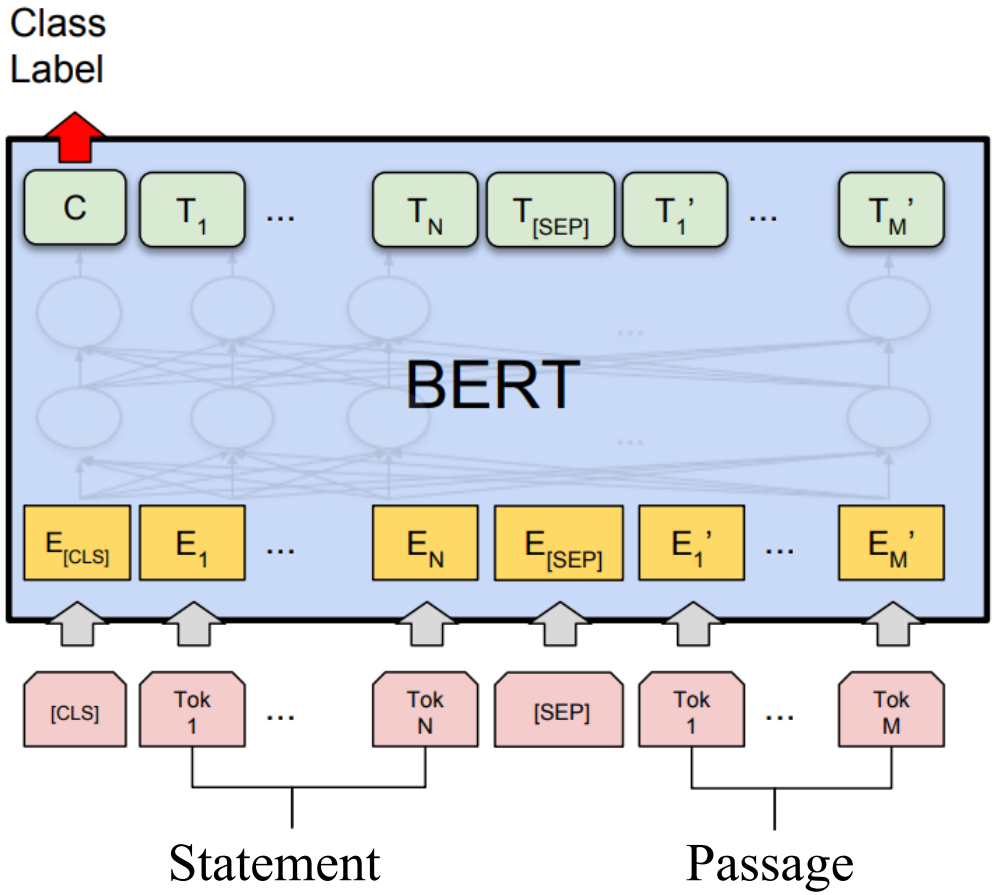
Source: https://github.com/voidism/EAR

1 – Legal Textual Entailment Recognition

Approach: Retriever - Classifier

- Stage 2: Classifier
- Metric: Accuracy

```
[
  {
    "example_id": "DS-101",
    "label": "Yes/No",
  }
]
```



2 – Comparative Opinion Mining

Description

- **Participants are required to develop models that can extract the following information, referred to as a "quintuple," from comparative sentences:**

1. Subject: The entity that is the subject of the comparison (e.g., a particular product model).
2. Object: The entity being compared to the subject (e.g., another model or a general reference).
3. Aspect: The word or phrase about the feature or attribute of the subject and object that is being compared (e.g., battery life, camera quality, performance).
4. Predicate: The comparative word or phrase expressing the comparison (e.g., "better than," "worse than," "equal to").
5. Comparison Type Label: This label indicates the type of comparison made and can be one of the following categories: ranked comparison (e.g., "better", "worse"), superlative comparison (e.g., "best", "worst"), equal comparison (e.g., "same as," "as good as"), and non-gradable comparison (e.g., "different from," "unlike").

Source: https://vlsp.org.vn/vlsp2023/eval/comon

2 – Comparative Opinion Mining

Example

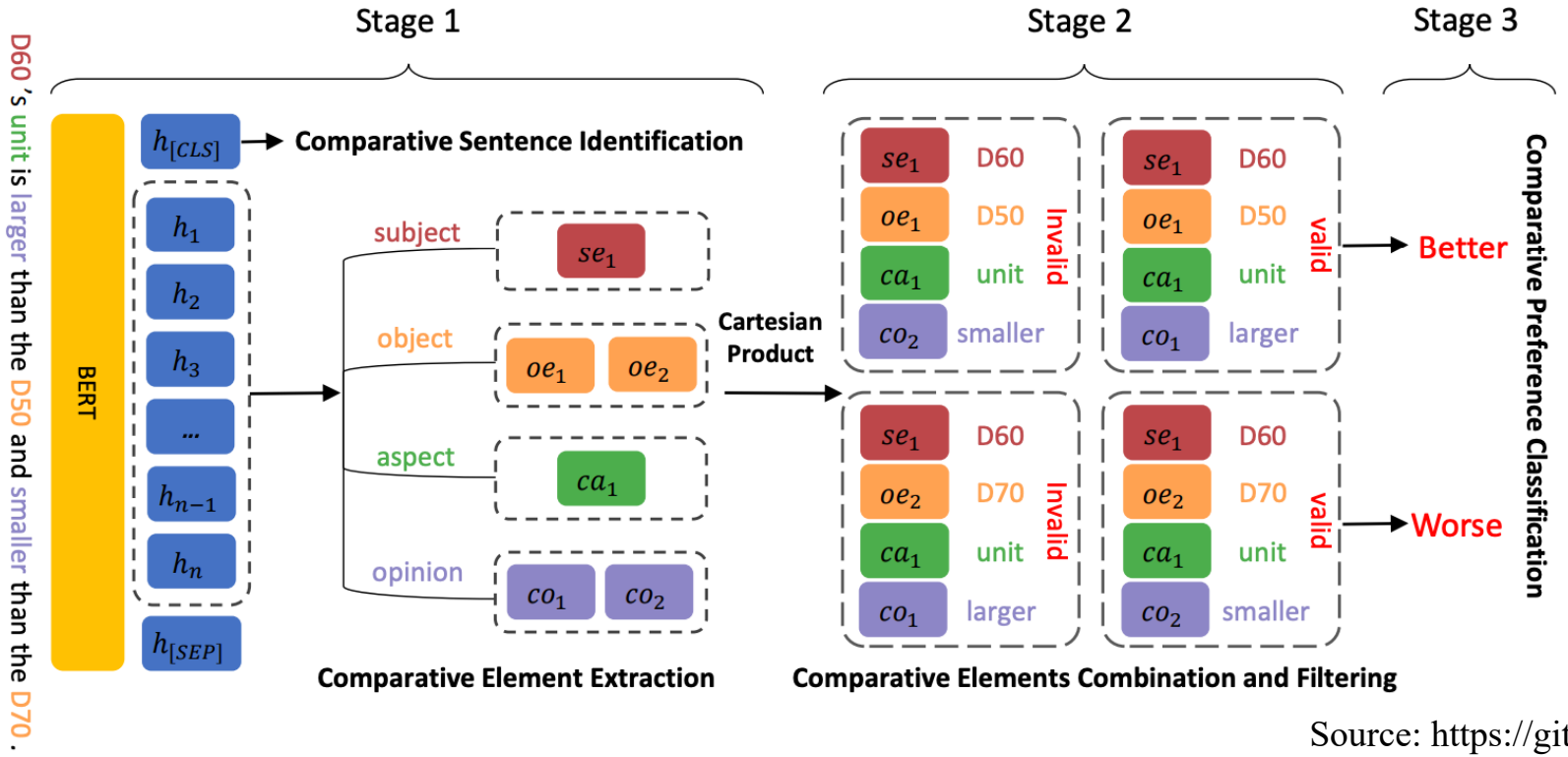
- G6 has a worse zoom than G7, but G6's battery was more reliable than G7

Elements	Subject: G6
	Object: G7
	Aspect: {zoom, battery}
	Predicate: {worse, more reliable}
	comparison type label: {ranked comparison, ranked comparison}
Quintuple	{(G6, G7, zoom, worse, ranked comparison), (G6, G7, battery, more reliable, ranked comparison)

2 – Comparative Opinion Mining

Approach

- Comparative Opinion Quintuple Extraction
- BERT-based Multi-Stage Neural Network

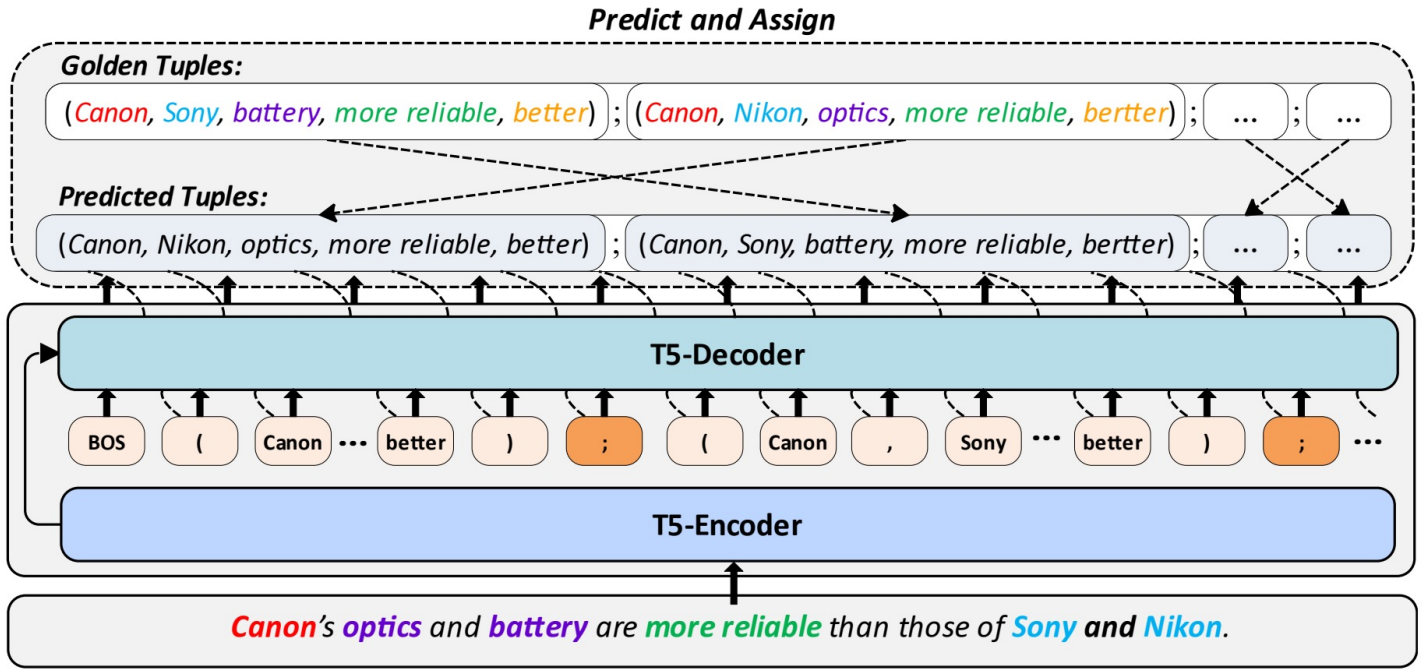


Source: https://github.com/nustm/coqe

2 – Comparative Opinion Mining

Approach

- Comparative Opinion Quintuple Extraction
- UniCOQE Model



Source: https://aclanthology.org/2023.findings-acl.775.pdf

2 – Comparative Opinion Mining

Approach

- Comparative Opinion Quintuple Extraction
- UniCOQE Model
- Data training format

Input: Canon's optics and battery are more reliable than those of Sony and Nikon.
Target:
(Canon, Sony, optics, more reliable, BETTER);
(Canon, Sony, battery, more reliable, BETTER);
(Canon, Nikon, optics, more reliable, BETTER);
(Canon, Nikon, battery, more reliable, BETTER)
Input: Canon's optics and battery are so great.
Target:
(unknown, unknown, unknown, unknown, unknown)

Source: https://aclanthology.org/2023.findings-acl.775.pdf

2 – Comparative Opinion Mining

Approach

- Comparative Opinion Quintuple Extraction
- UniCOQE Model

Models	Camera-COQE		Car-COQE		Ele-COQE	
	CSI	COQE	CSI	COQE	CSI	COQE
Multi-Stage$_{\text{CSR-CRF}}$	65.38	3.46	86.90	5.19	88.30	4.07
Joint$_{\text{CRF}}$	82.14	4.88	89.85	8.65	85.97	4.71
Multi-Stage$_{\text{LSTM}}$	87.14	9.05	92.68	10.28	96.25	14.90
Multi-Stage$_{\text{BERT}}$	93.04	13.36	97.39	29.75	98.31	30.73
UniCOQE	**95.21**	**31.95**	**98.28**	**36.55**	**98.41**	**35.46**

Source: https://aclanthology.org/2023.findings-acl.775.pdf

2 – Comparative Opinion Mining

Evaluation

- Comparative Opinion Quintuple Extraction
- Evaluation: Exact Match, Proportional Match, Binary Match

EM = 1/3

Input: *Canon's batteries are more reliable than those of Sony, but Canon's sensors are less stable than those of Sony and Nikon.*

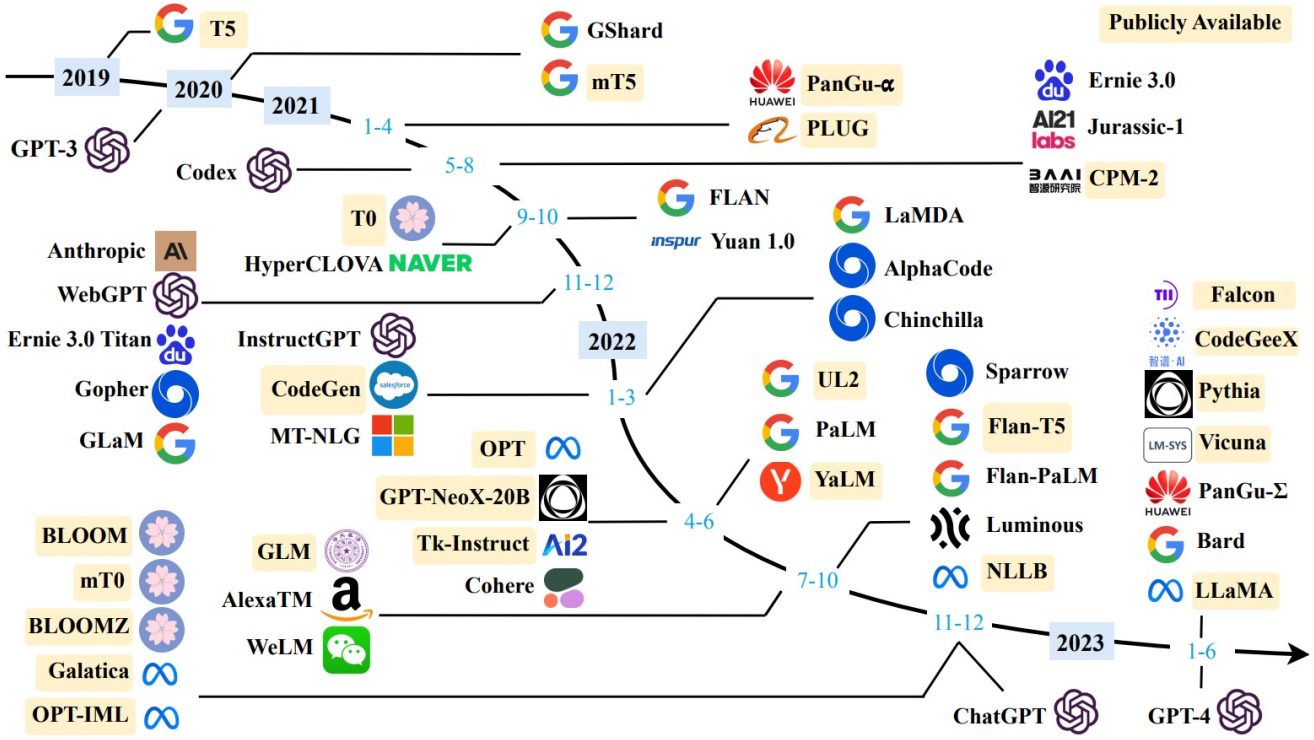
Predicted Quintuples: $Q_{pred} = p_1; p_2; p_3$	*Golden Quintuples:* $Q_{gold} = g_1; g_2; g_3$
p_1=(Canon, Nikon, sensors, less stable, WORSE)	g_1=(Canon, Sony, batteries, more reliable, BETTER)
p_2=(Canon, Sony, batteries, more reliable, BETTER)	g_2=(Canon, Sony, sensors, less stable, WORSE)
p_3=(Canon, Sony, sensors, less stable, WORSE)	g_3=(Canon, Nikon, sensors, less stable, WORSE)

Source: https://aclanthology.org/2023.findings-acl.775.pdf

3 – Vietnamese Large Language Model

Description

- The goal of VLSP2023-VLLMs is to promote the development of large language models for Vietnamese by constructing an evaluation dataset for VLLMs



3 – Vietnamese Large Language Model

Pipeline for Pre-Training LLM

- Step 1: Prepare dataset
- Step 2: Preprocessing
- Step 3: Model (Transformer, Finetuning LLMs,…)
- Step 4: Pre-training Tasks (Objective Functions)
- Step 5: Optimization Setting
- Step 6: Adaptation
- Step 7: Evaluation

3 – Vietnamese Large Language Model

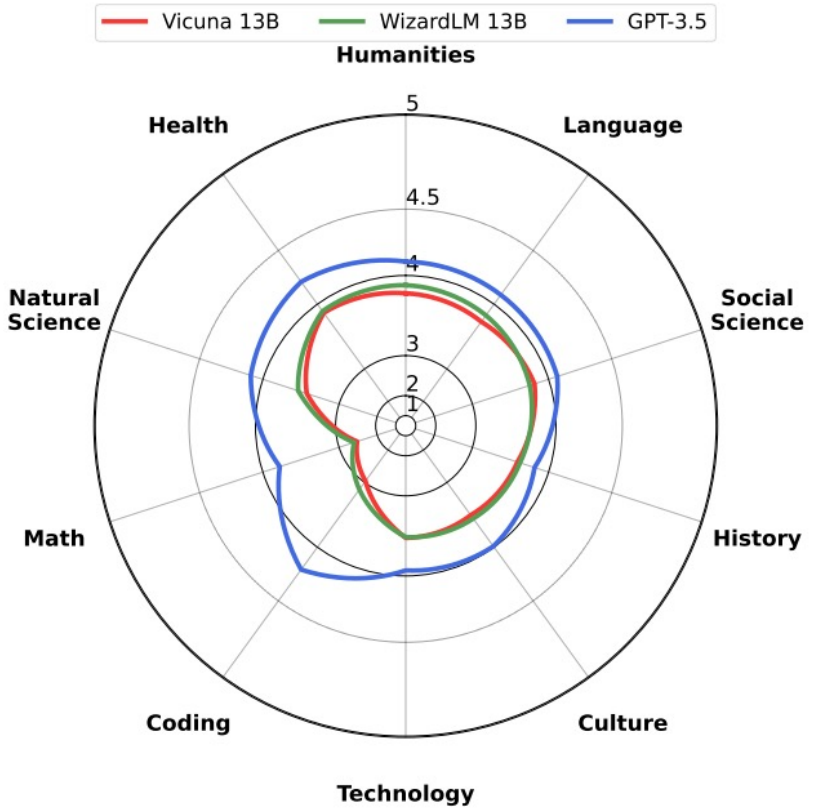
3.1. Dataset

- Monolingual Vietnamese Dataset

  CC-100: 137GB

Corpora	Size	Source	Latest Update Time
BookCorpus [134]	5GB	Books	Dec-2015
Gutenberg [135]	-	Books	Dec-2021
C4 [73]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [136]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [137]	120GB	CommonCrawl	Apr-2019
OpenWebText [138]	38GB	Reddit links	Mar-2023
Pushift.io [139]	2TB	Reddit links	Mar-2023
Wikipedia [140]	21GB	Wikipedia	Mar-2023
BigQuery [141]	-	Codes	Mar-2023
the Pile [142]	800GB	Other	Dec-2020
ROOTS [143]	1.6TB	Other	Jun-2022

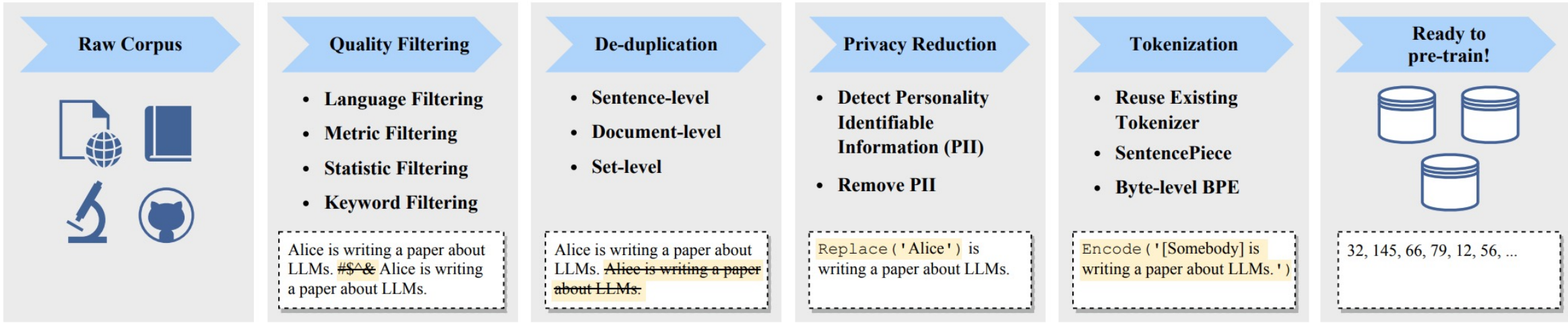
https://arxiv.org/pdf/2303.18223.pdf



https://arxiv.org/pdf/2307.10928.pdf

3 – Vietnamese Large Language Model

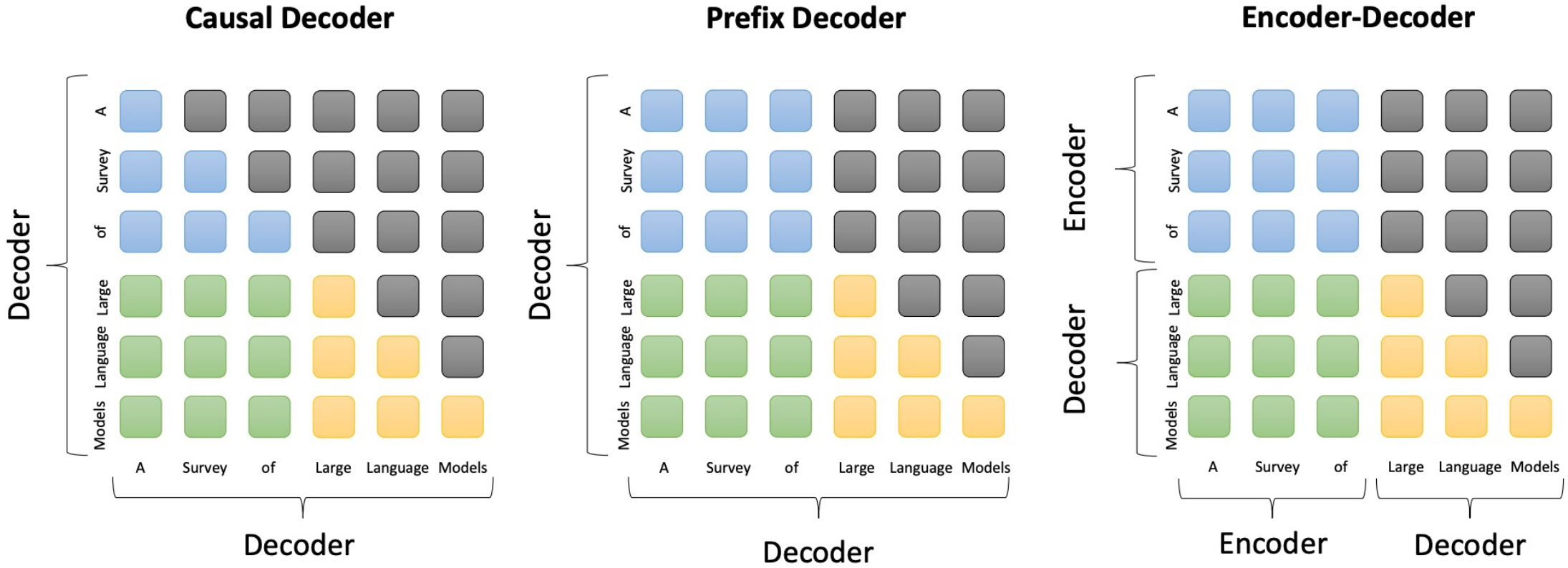
3.2. Preprocessing



https://arxiv.org/pdf/2303.18223.pdf

3 – Vietnamese Large Language Model

3.3. Architecture



https://arxiv.org/pdf/2303.18223.pdf

3 – Vietnamese Large Language Model

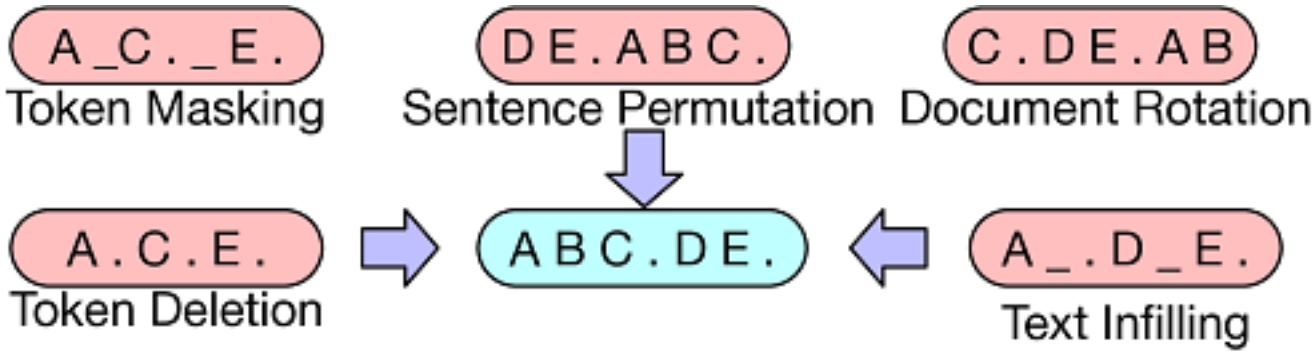
3.3. Architecture

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Causal decoder	175B	Pre LayerNorm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [75]	Causal decoder	207B	Pre LayerNorm	Learned	GeLU	✓	64	128	16384	1024
OPT [81]	Causal decoder	175B	Pre LayerNorm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Causal decoder	540B	Pre LayerNorm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [69]	Causal decoder	176B	Pre LayerNorm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [97]	Causal decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Causal decoder	280B	Pre RMSNorm	Relative	-	-	80	128	16384	2048
Chinchilla [34]	Causal decoder	70B	Pre RMSNorm	Relative	-	-	80	64	8192	-
Galactica [35]	Causal decoder	120B	Pre LayerNorm	Learned	GeLU	×	96	80	10240	2048
LaMDA [63]	Causal decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [91]	Causal decoder	178B	Pre LayerNorm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Causal decoder	65B	Pre RMSNorm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [83]	Prefix decoder	130B	Post DeepNorm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [73]	Encoder-decoder	11B	Pre RMSNorm	Relative	ReLU	×	24	128	1024	512

https://arxiv.org/pdf/2303.18223.pdf

3 – Vietnamese Large Language Model

3.4. Pre-training Tasks



BART: https://arxiv.org/abs/1910.13461

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*

T5: https://arxiv.org/abs/1910.10683

3 – Vietnamese Large Language Model

3.5. Optimization Setting

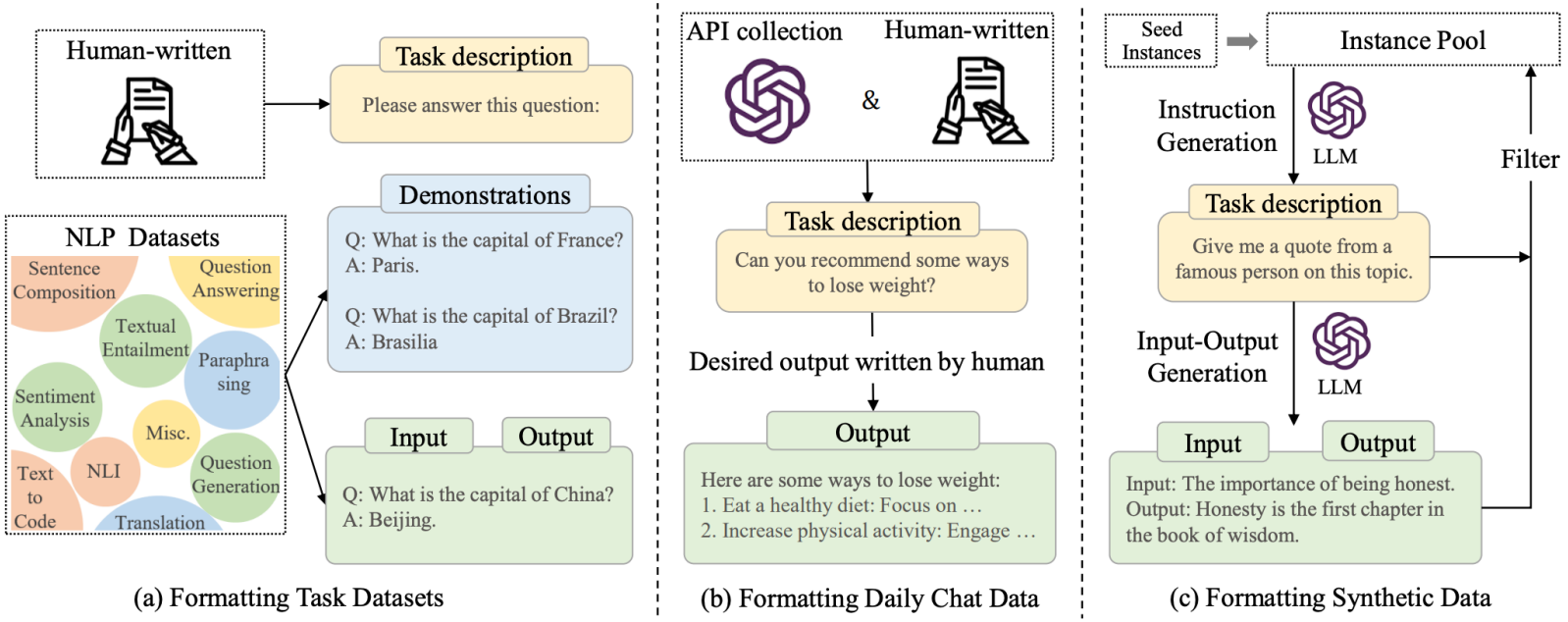
Model	Batch Size (#tokens)	Learning Rate	Warmup	Decay Method	Optimizer	Precision Type	Weight Decay	Grad Clip	Dropout
GPT3 (175B)	32K→3.2M	6×10^{-5}	yes	cosine decay to 10%	Adam	FP16	0.1	1.0	-
PanGu-α (200B)	-	2×10^{-5}	-	-	Adam	-	0.1	-	-
OPT (175B)	2M	1.2×10^{-4}	yes	manual decay	AdamW	FP16	0.1	-	0.1
PaLM (540B)	1M→4M	1×10^{-2}	no	inverse square root	Adafactor	BF16	lr^2	1.0	0.1
BLOOM (176B)	4M	6×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	0.0
MT-NLG (530B)	64 K→3.75M	5×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	-
Gopher (280B)	3M→6M	4×10^{-5}	yes	cosine decay to 10%	Adam	BF16	-	1.0	-
Chinchilla (70B)	1.5M→3M	1×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	-	-	-
Galactica (120B)	2M	7×10^{-6}	yes	linear decay to 10%	AdamW	-	0.1	1.0	0.1
LaMDA (137B)	256K	-	-	-	-	BF16	-	-	-
Jurassic-1 (178B)	32 K→3.2M	6×10^{-5}	yes	-	-	-	-	-	-
LLaMA (65B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
GLM (130B)	0.4M→8.25M	8×10^{-5}	yes	cosine decay to 10%	AdamW	FP16	0.1	1.0	0.1
T5 (11B)	64K	1×10^{-2}	no	inverse square root	AdaFactor	-	-	-	0.1
ERNIE 3.0 Titan (260B)	-	1×10^{-4}	-	-	Adam	FP16	0.1	1.0	-
PanGu-Σ (1.085T)	0.5M	2×10^{-5}	yes	-	Adam	FP16	-	-	-

https://arxiv.org/pdf/2303.18223.pdf

3 – Vietnamese Large Language Model

3.6. Adaptation of LLMs

- Instruction Tuning
- Three different methods for constructing the instruction-formatted instances



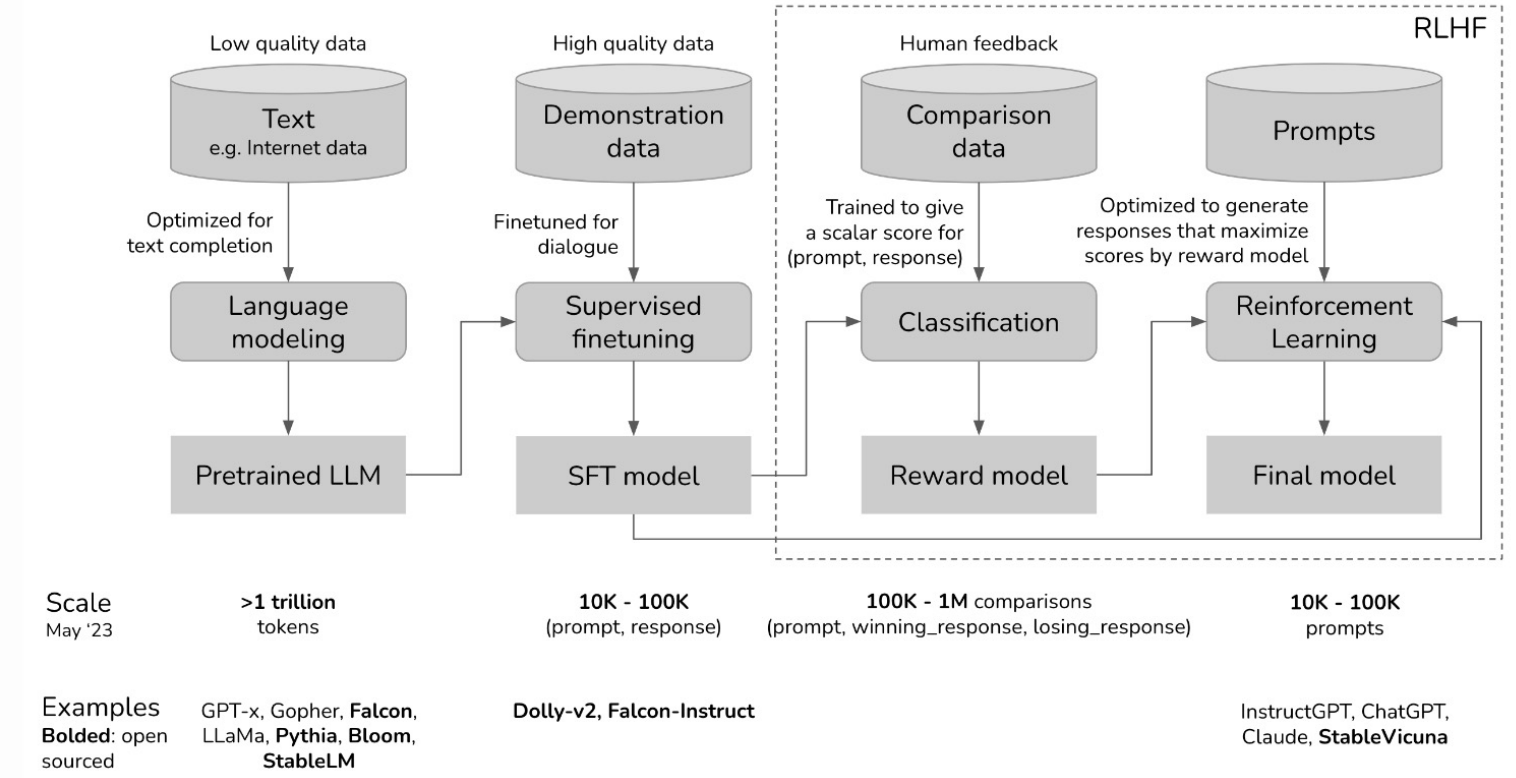
(a) Formatting Task Datasets (b) Formatting Daily Chat Data (c) Formatting Synthetic Data

https://arxiv.org/pdf/2303.18223.pdf

3 – Vietnamese Large Language Model

3.6. Adaptation of LLMs

- RLHF: Reinforcement Learning from Human Feedback



Source: https://arxiv.org/pdf/2203.02155.pdf
https://huyenchip.com/2023/05/02/rlhf.html

3 – Vietnamese Large Language Model

3.7. Library Resource

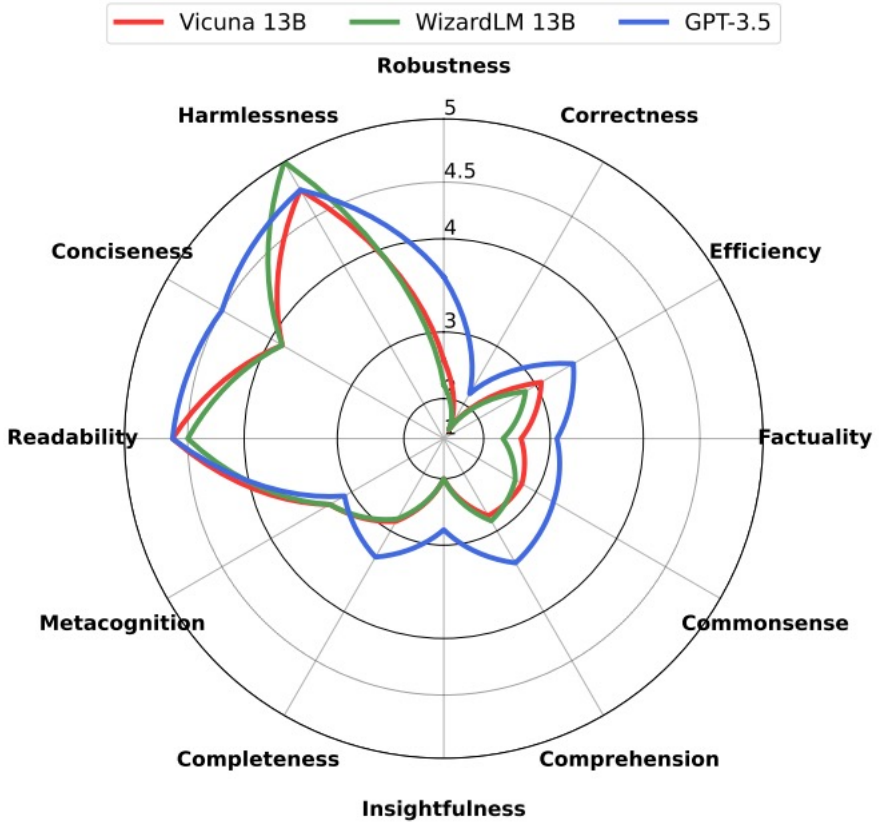
- Transformers
- DeepSpeed
- Langchain
- Megatron-LM

3 – Vietnamese Large Language Model

3.8. Evaluation

Abilities

- Logical thinking
  - Logical Correctness
- Background knowledge
  - Factuality
  - Commonsense Understanding
- Problem handling
  - Comprehension
  - Insightfulness
  - Metacognition
- User alignment
  - Harmlessness
  - Follow the Correct Instruction

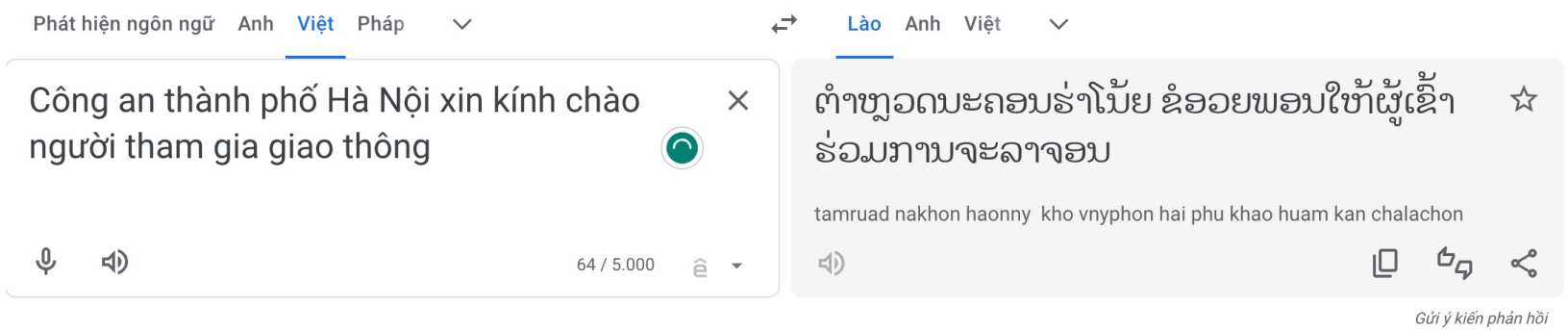


https://arxiv.org/pdf/2307.10928.pdf

4 – Machine Translation

Description

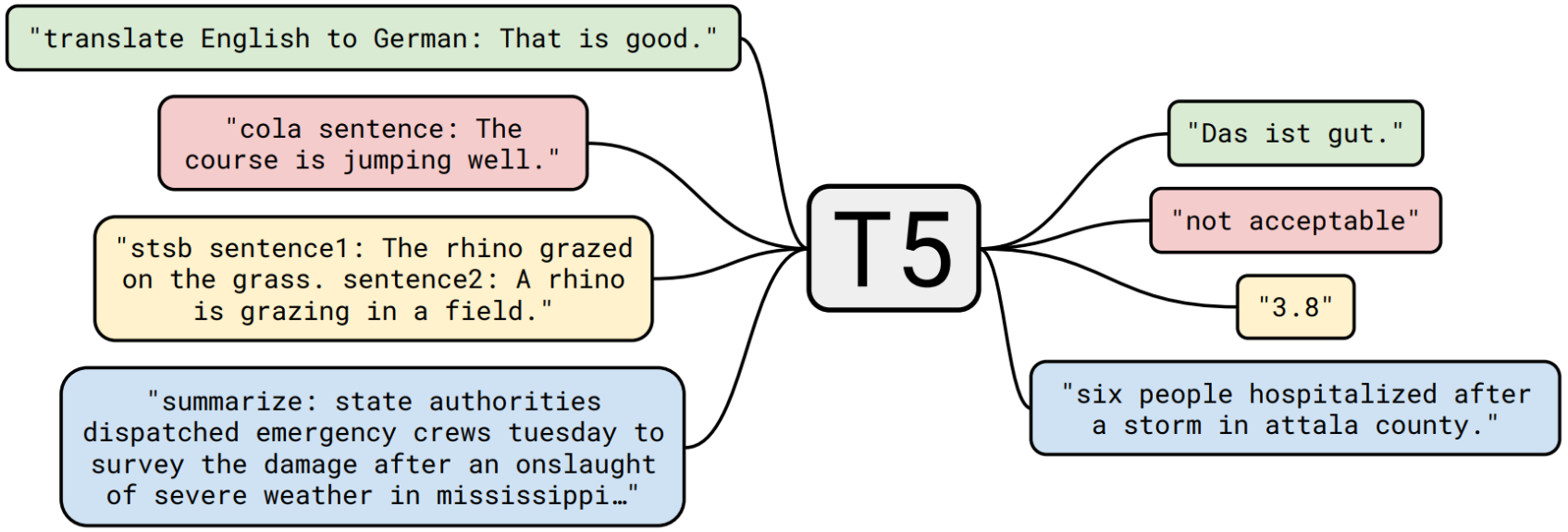
- Lao-Vietnamese Machine Translation
- Training and Test Data:
  - Parallel Corpora: Lao-Vietnamese
  - Monolingual Corpora: Lao and Vietnamese
  - Development set and test set: Lao-Vietnamese



4 – Machine Translation

Approach

- Low-resource Neural Machine Translation
- Fine Tuning: mT5 (101 languages)



Source: https://huggingface.co/google/mt5-base

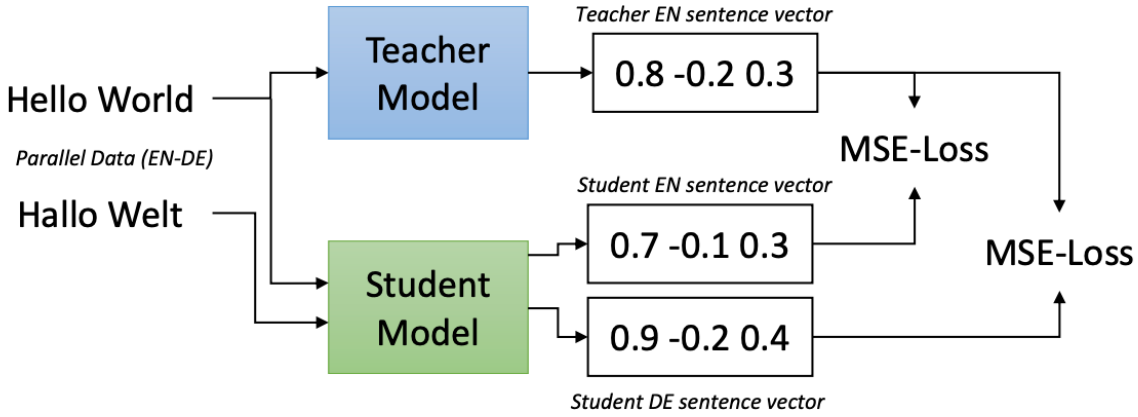
4 – Machine Translation

Improvement

- Low-resource Neural Machine Translation
- Data Filtering: Find-out low quality sentence pairs based on cosine similarity sentence-level score

Công anh thành phố Hà Nội xin kính chào người tham gia giao thông	ຕຳຫຼວດນະຄອນຮ່າໂນ້ຍ (Công an Hà Nội)

- Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

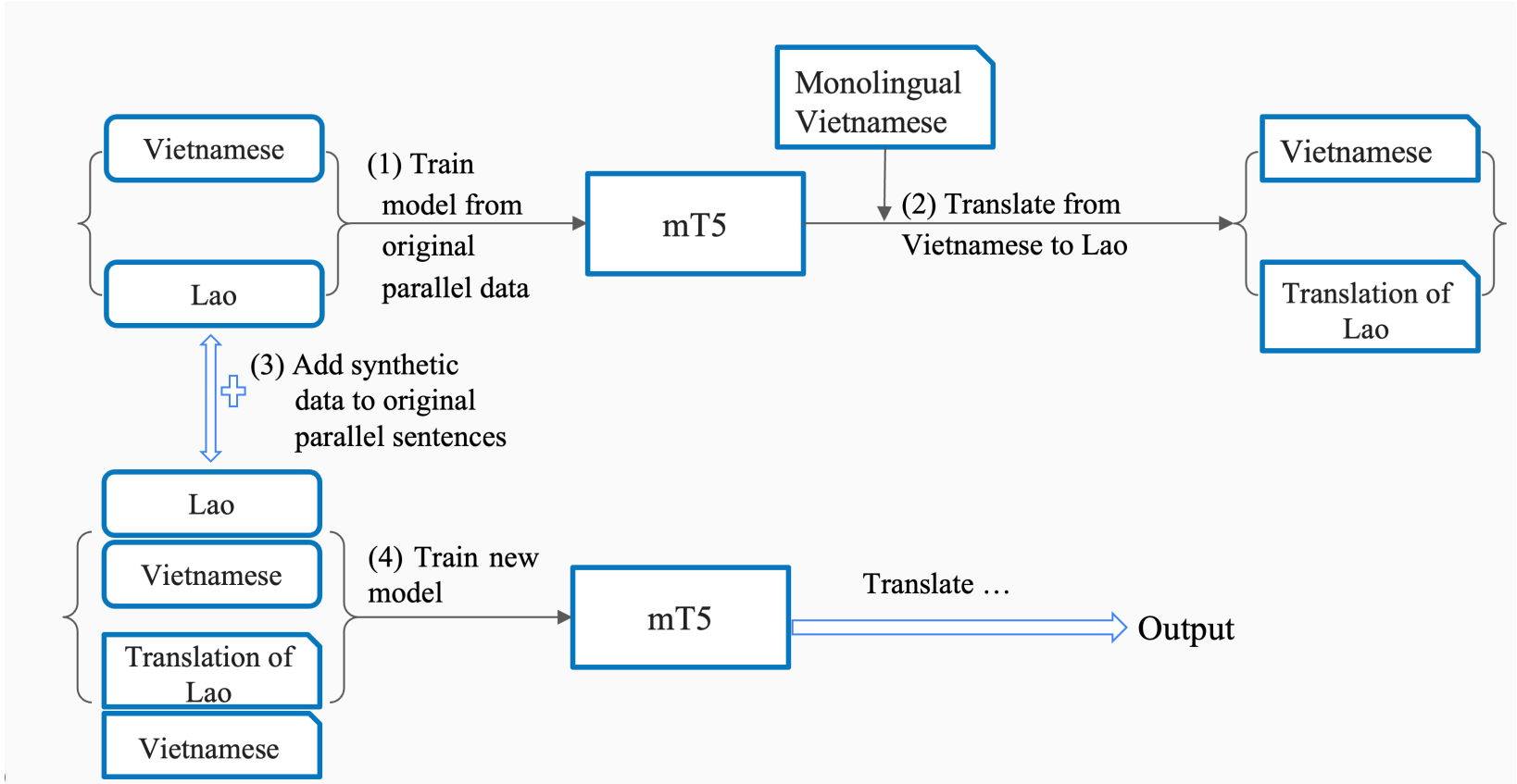


https://arxiv.org/abs/2004.09813

4 – Machine Translation

Improvement

- Back Translation



4 – Machine Translation

Evaluation

- BLEU Score

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

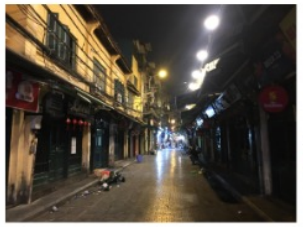
5 – Visual Reading Comprehension

Description

- The OpenViVQA dataset: **11,000+** images associated with **37,000+** question-answer pairs in Vietnam

cửa tiệm này tên gì?
Answer: cửa tiệm này tên mộc

cửa tiệm cắt tóc nam tên gì?
Answer: cửa tiệm cắt tóc bầu



quán bia saigon nằm trên con đường nào?
Answer: đường tạ hiện

dòng xe đang di chuyển về địa phận của tỉnh/thành phố nào?
Answer: dòng xe đang di chuyển về địa phận tây ninh

đây là tỉnh nào?
Answer: đây là tỉnh vĩnh long

doanh nghiệp này chọn loại hình doanh nghiệp nào?
Answer: doanh nghiệp này chọn loại hình doanh nghiệp công ty cổ phần

sản phẩm này tên gì?
Answer: sản phẩm này tên bánh cosy quế các vị

bạn nữ đang đứng ở con phố nào?
Answer: phố tạ hiện

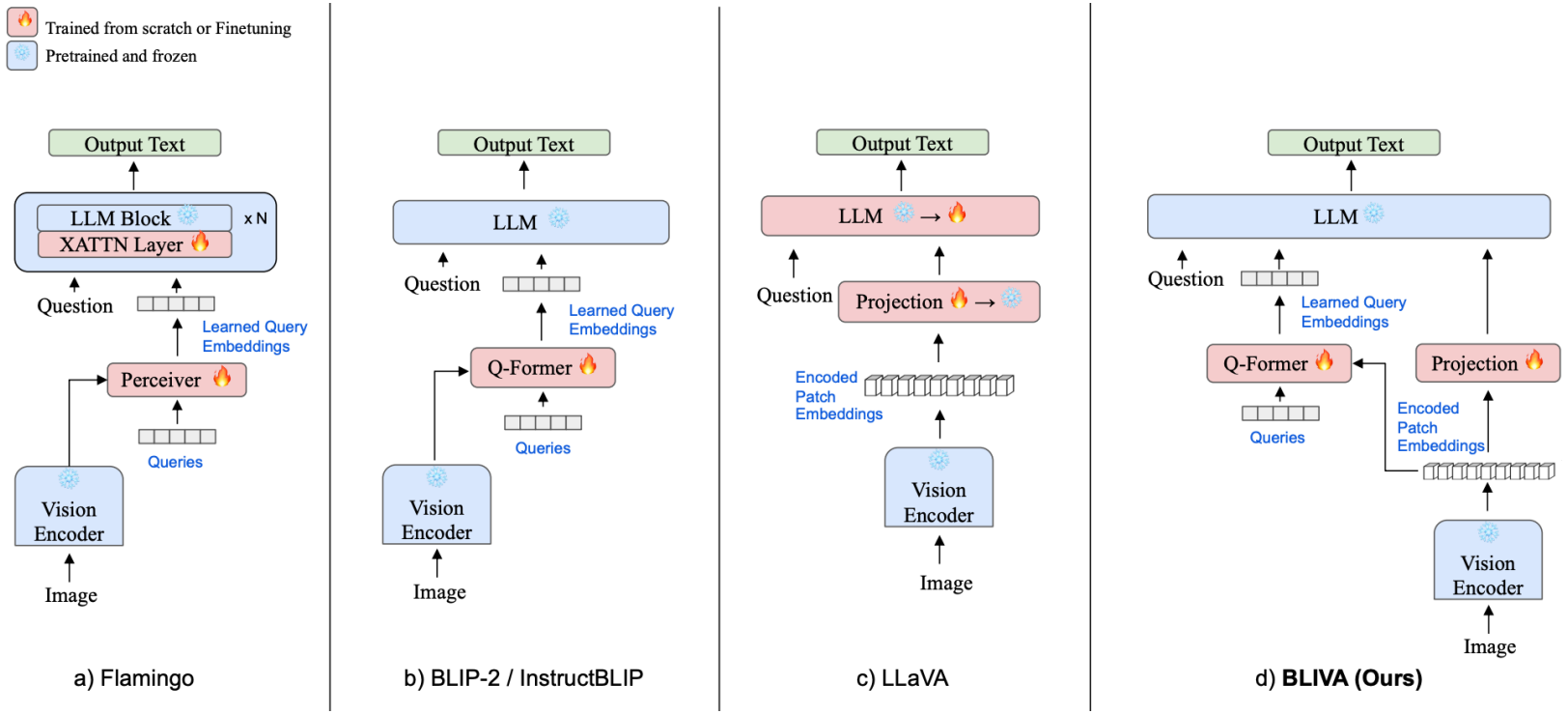
điểm bán hàng này của thương hiệu nào?
Answer: điểm bán hàng này của thương hiệu viettel

quán này làm bánh xèo theo kiểu gì?
Answer: quán này làm bánh xèo theo kiểu nhật bản

5 – Visual Reading Comprehension

Approach

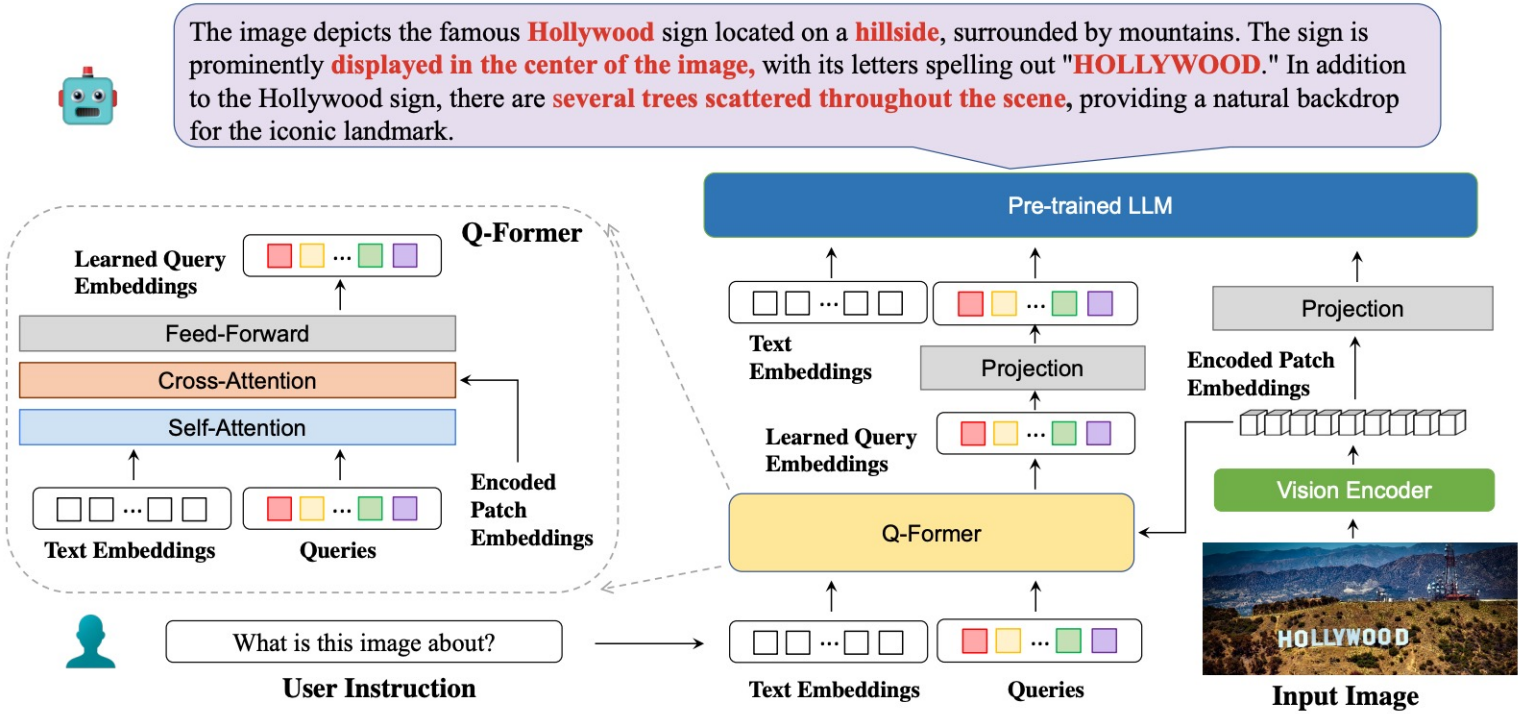
- Task: Vision Language Models (VLMs), Vision Reading Comprehension (VRC)
- Comparison of various VLM approaches



5 – Visual Reading Comprehension

Approach

- Task: Vision Language Models (VLMs), Vision Reading Comprehension (VRC)
- BLIVA Model

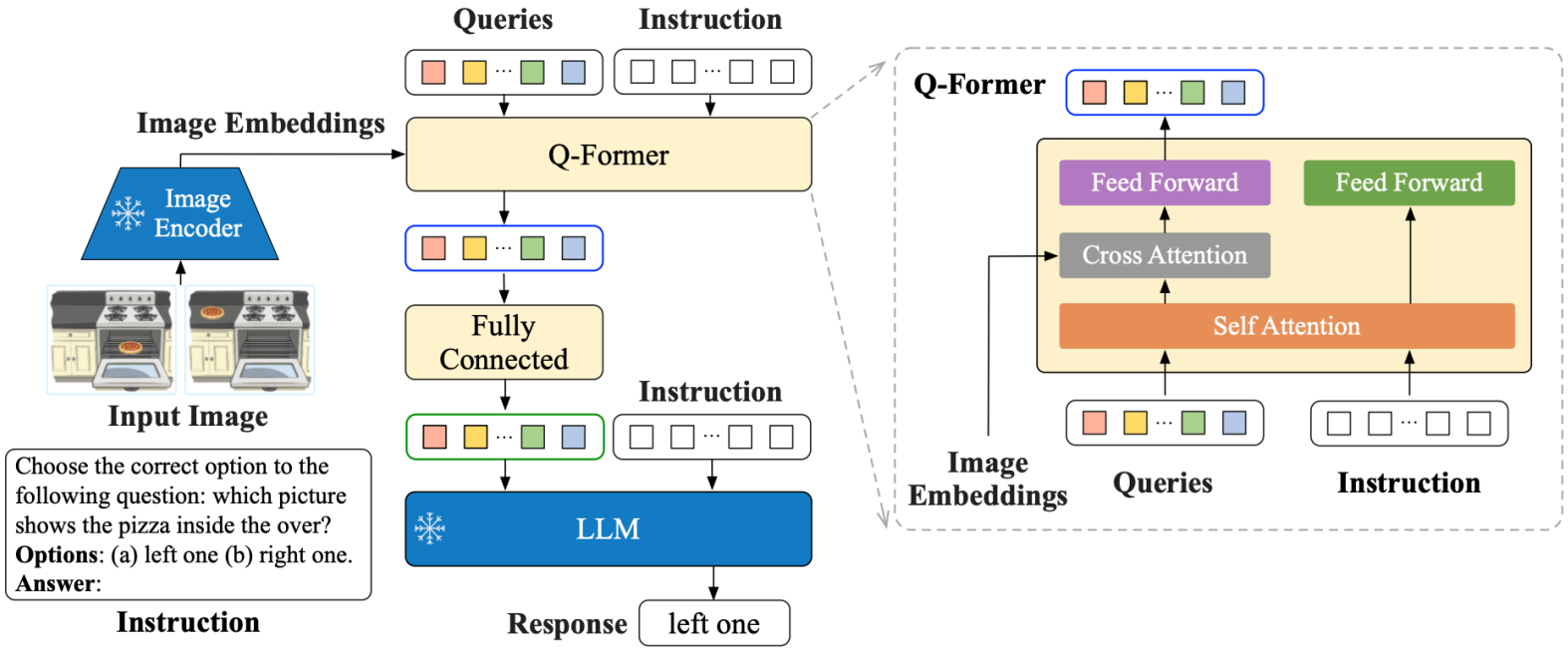


https://github.com/mlpc-ucsd/BLIVA

5 – Visual Reading Comprehension

Approach

- Task: Vision Language Models (VLMs), Vision Reading Comprehension (VRC)
- InstructBLIP

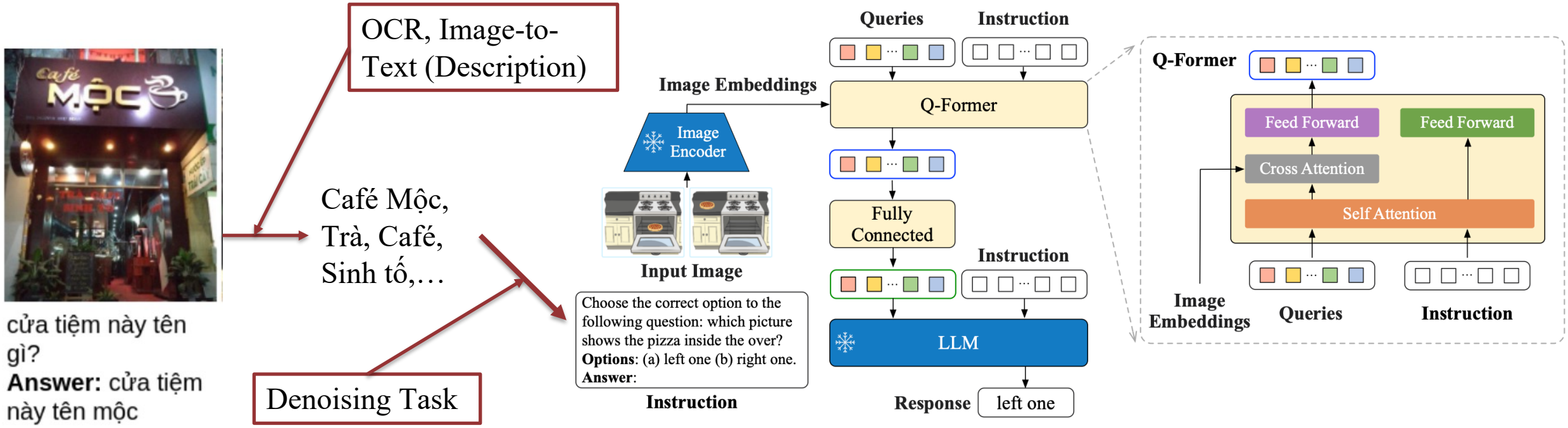


https://github.com/salesforce/LAVIS/tree/main/projects/instructblip

5 – Visual Reading Comprehension

Improvement

- Task: Vision Language Models (VLMs), Vision Reading Comprehension (VRC)
- InstructBLIP & BLIVA



5 – Visual Reading Comprehension

Evaluation

- BLEU
- CIDEr: https://arxiv.org/pdf/1411.5726.pdf

6 – Automatic Speech Recognition

Description

- Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER)

For example,

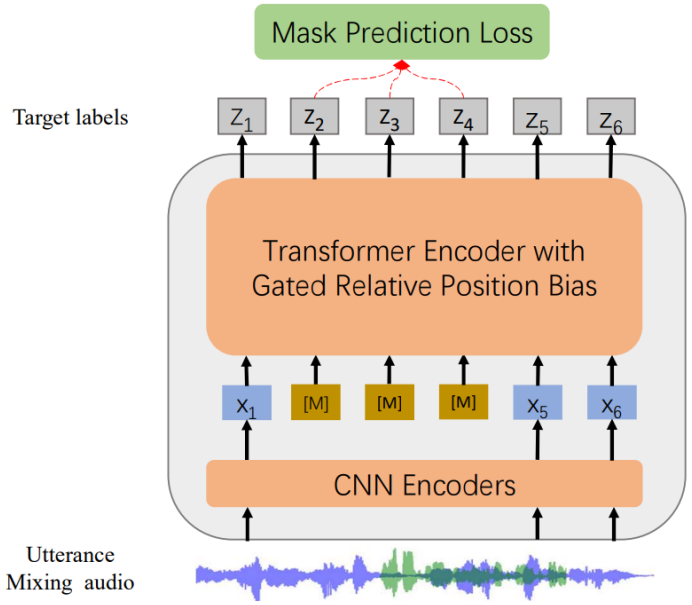
0001.wav neutral chào mừng các bạn đã tham dự cuộc thi
0002.wav negative tôi sẽ kiện ra toà

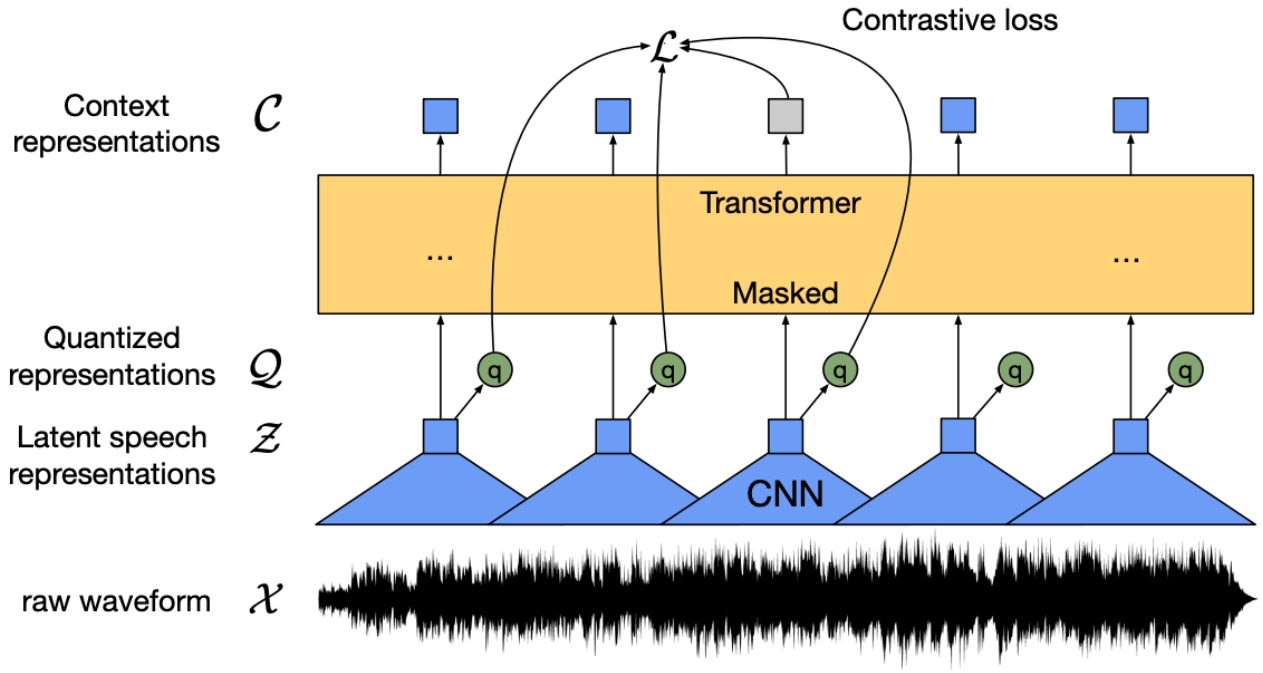
Dataset	Amount (hours)	Text Label	Emotion Label
Dataset 1	200	No	No
Dataset 2	60	Yes	No
Dataset 3	5	No	Yes
Dataset 4	40	No	Yes (Low quality)

6 – Automatic Speech Recognition

Approach

- **Automatic Speech Recognition (ASR)** and Speech Emotion Recognition (SER)
- Fine-tuning ASR model: Wav2Vec, WaLM, Whisper





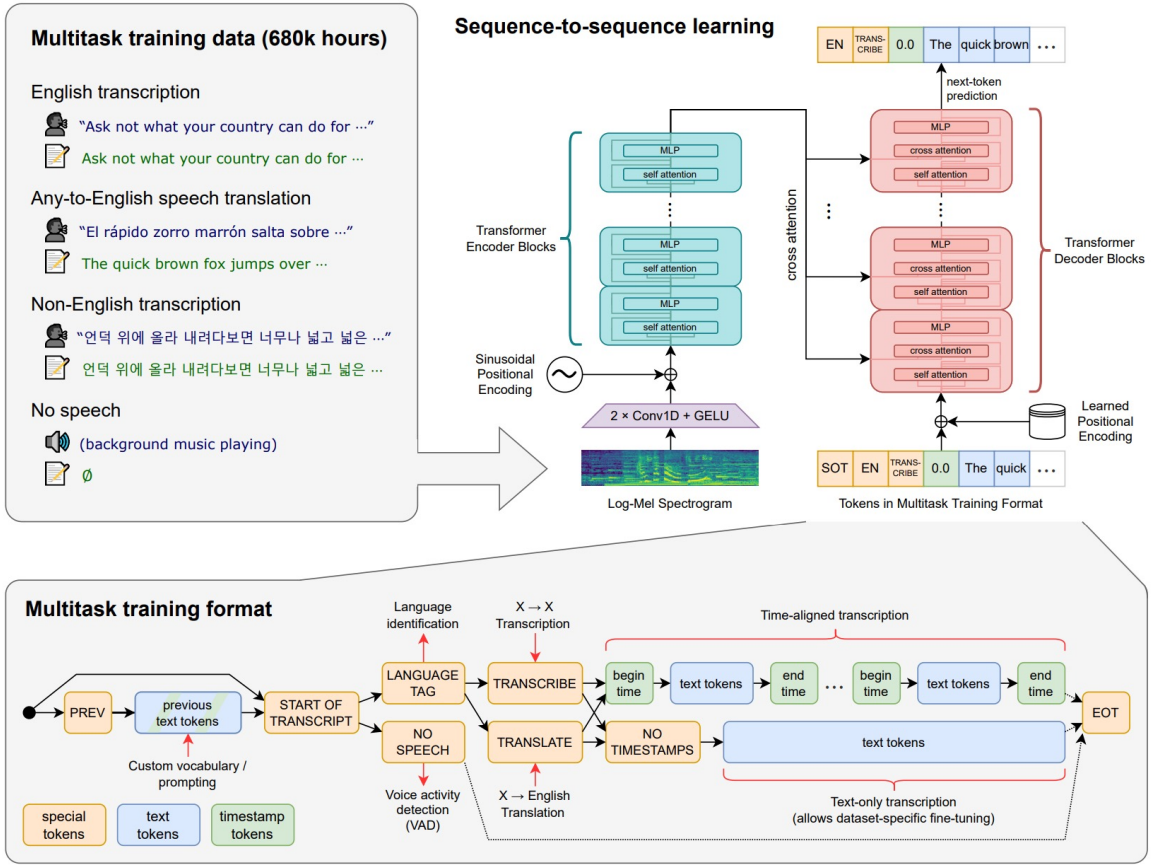
Wav2Vec: https://github.com/facebookresearch/fairseq

WavLM: https://github.com/microsoft/unilm/tree/master/wavlm

6 – Automatic Speech Recognition

Approach

- **Automatic Speech Recognition (ASR)** and Speech Emotion Recognition (SER)
- Fine-tuning ASR model: Wav2Vec, WaLM, Whisper

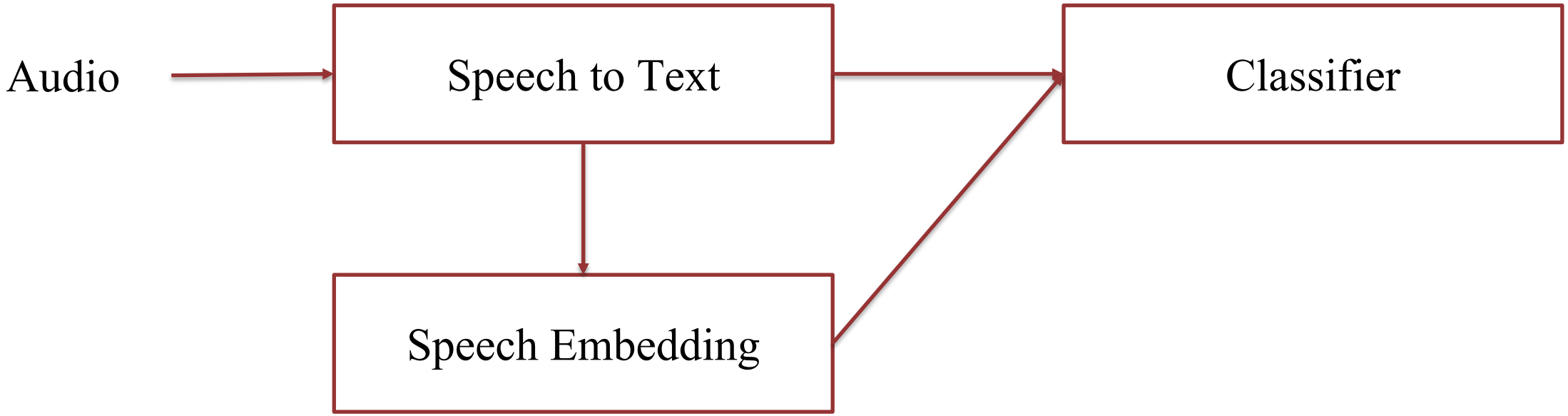


Whisper: https://github.com/openai/whisper

6 – Automatic Speech Recognition

Approach

- Automatic Speech Recognition (ASR) and **Speech Emotion Recognition (SER)**
- **Approach #1: ASR => SER**
- **Approach #2: Joint Learning**



6 – Automatic Speech Recognition

Evaluation

For each given utterance, two outputs will be submitted.

- Text sequence (ASR output)
- Emotion label (Emotion output)

The quality of the models will be evaluated by the Syllable Error Rate ($SyER_{ASR}$) and Emotion Recognition Accuracy (ACC_{SER}) metrics.

$SyER_{ASR} = (S+D+I)/N$, where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct syllables,
- N is the number of syllables in the reference (N=S+D+C)

$ACC_{SER} = (NEU_{Corr}/NEU + NEG_{Corr}/NEG)/2$, where

- NEU_{Corr} is the number of correct neutral emotion utterances
- NEU is the number of total neutral utterances
- NEG_{Corr} is the number of correct negative emotion utterances
- NEG is the number of total negative utterances.

The overall result is calculated as

$Score = 0.7*(1-SyER_{ASR}) + 0.3*ACC_{SER}$

7 – Improvement Trick

Improvement

- EDA, Noise Handling (Overfiting)
- Fine-Tuning Pre-trained LMs: Multilingual PTLMs (xlm-Roberta,…)
- Ensemble PLMs, Average Checkpoints: Ref
- Augmentation: synonym words, via pivot language, denoising, masked language modeling,…
- Contrastive Learning
- Unsupervised pre-training + fine-tuning
- Prompting with LLMs

AI VIET NAM

@aivietnam.edu.vn

Thanks!

Any questions?