An overview of foundation models for Vietnamese language processing

Dat Quoc Nguyen
Head of NLP, VinAI
https://datquocnguyen.github.io

16/12/2023

Transformer

- Transformer architecture for machine translation [1]

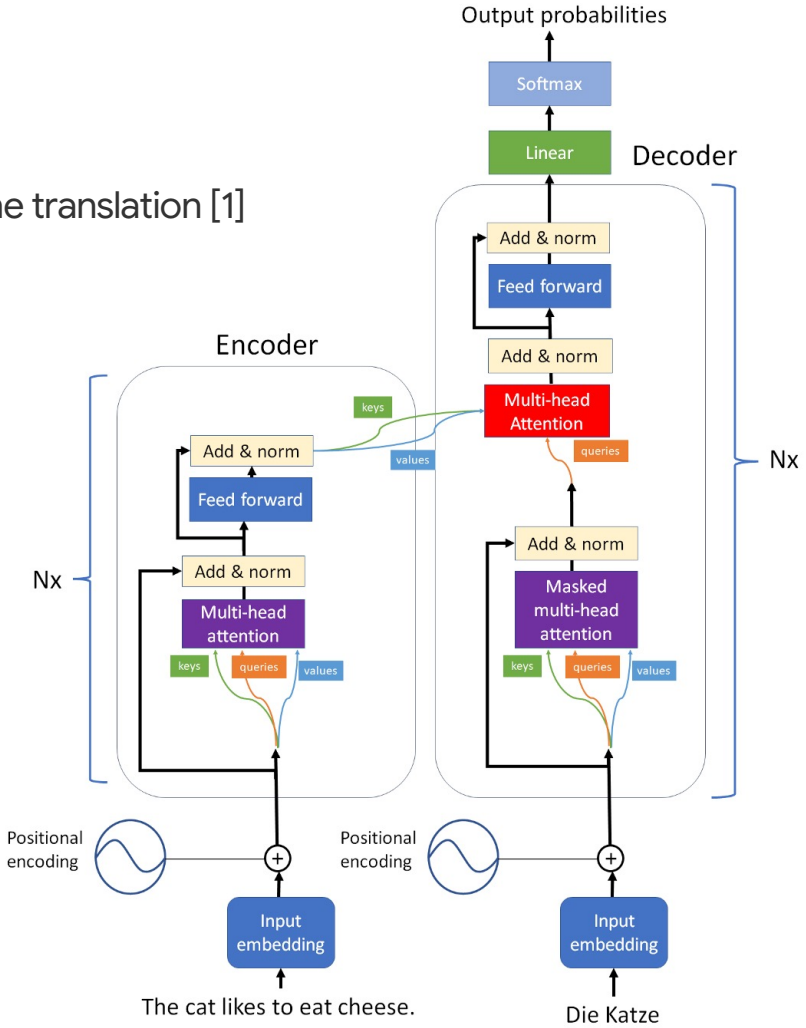


Image by Vincent Mueller: https://medium.com/towards-data-science/attention-please-85bd0abac41

Encoder-only Models

- Transformer encoder-based models
  - Pre-trained on large-scale corpora with a masked language modeling objective
  - BERT [2], RoBERTa [3], ELECTRA [4]

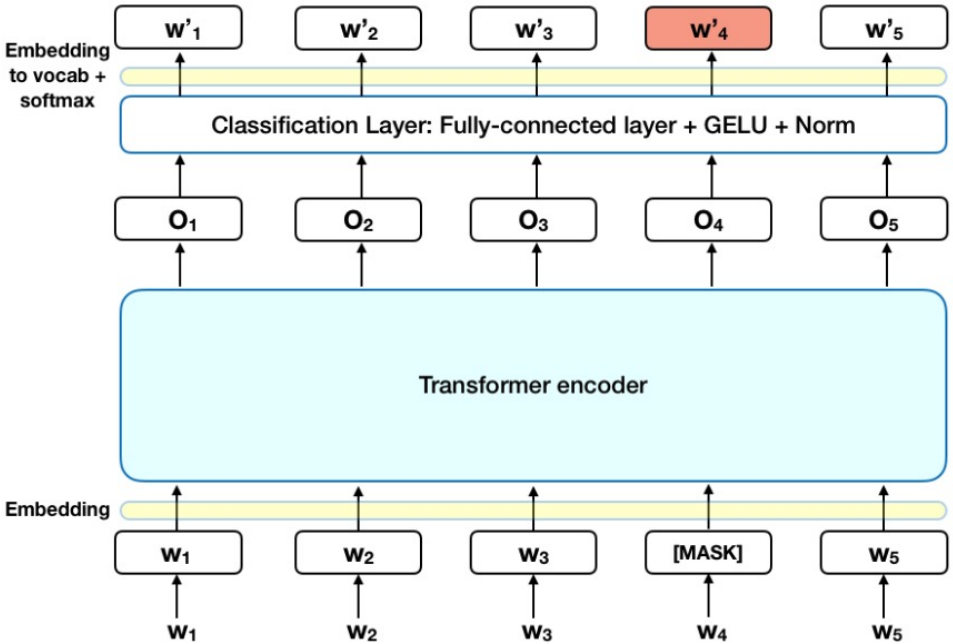


Image by Rani Horev: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Encoder-only Models

- Pre-trained BERT-type models for Vietnamese:

Model	Type	Date	Data (vi)	
mBERT [2]	Multi.	11/2018	01 GB	Syllable
XLM-R [5]	Multi.	11/2019	137 GB	Syllable
PhoBERT [6]	Mono.	03/2020	20 GB	Word
viBert [7]	Mono.	10/2020	10 GB	Syllable
vELECTRA [7]	Mono.	10/2020	60 GB	Syllable
PhoBERT (**v2**)	Mono.	04/2023	140 GB	Word
VnLawBERT [8]	Mono.	11/2020	0.32 GB	Syllable
ViHealthBERT [9]	Mono.	06/2022	19 GB*	Syllable
ViSoBERT [10]	Mono.	10/2023	01 GB	Syllable

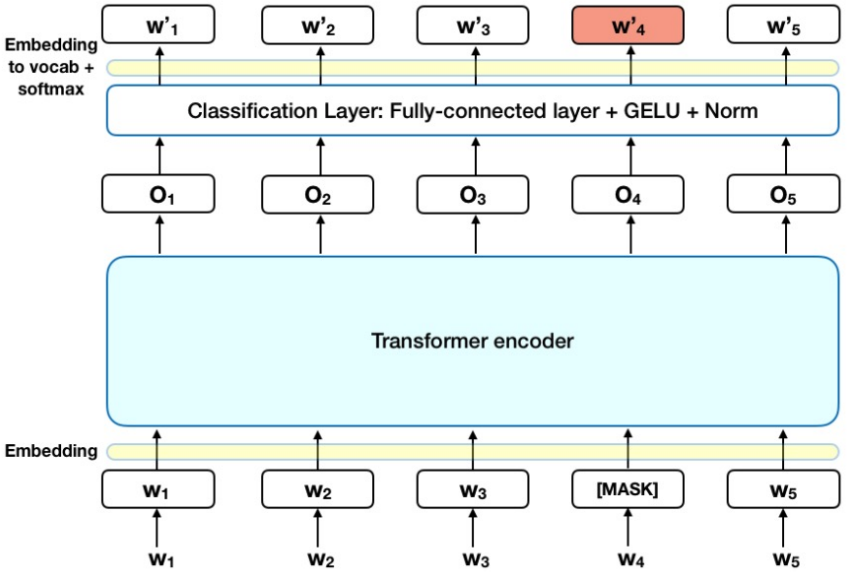


Image by Rani Horev: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

"Data (vi)" denotes the size of Vietnamese text data used for pre-training

*: 130M Sentences ~ 19 GB

Encoder-only Models

- Some comparison results

Model	Data (vi)	
mBERT [2]	01 GB	Syllable
XLM-R [5]	137 GB	Syllable
PhoBERT [6]	20 GB	Word
viBert_FPT [7]	10 GB	Syllable
vELECTRA_FPT [7]	60 GB	Syllable
viBert4news	20 GB	Word

THE EXPERIMENTAL RESULTS OF VARIOUS MONO-LINGUAL AND MULTI-LINGUAL PRE-TRAINED BERT MODELS ON VIETNAMESE ASPECT CATEGORY DETECTION TASK FOR THE RESTAURANT DOMAIN.

Types	Models	Precision	Recall	F1-score
Multi-lingual	mBERT	81.39	76.34	78.78
	mDistilBert	80.35	76.07	78.16
	XLM-R	82.98	81.40	82.18
Mono-lingual	viBert4news	79.26	77.48	78.36
	viBert_FPT	80.65	79.12	79.88
	vELECTRA_FPT	83.08	79.54	81.27
	PhoBERT	**85.60**	**87.49**	**86.53**

THE EXPERIMENTAL RESULTS OF VARIOUS MONO-LINGUAL AND MULTI-LINGUAL PRE-TRAINED BERT MODELS ON VIETNAMESE ASPECT CATEGORY DETECTION TASK FOR THE HOTEL DOMAIN.

Types	Models	Precision	Recall	F1-score
Multi-lingual	mBERT	77.93	76.26	77.09
	mDistilBert	78.59	74.97	76.73
	XLM-R	78.86	76.56	77.70
Mono-lingual	viBert4news	79.39	74.83	77.04
	viBert_FPT	81.14	74.54	77.70
	vELECTRA_FPT	79.82	76.07	77.90
	PhoBERT	**81.49**	**76.96**	**79.16**

Tables taken from [11]

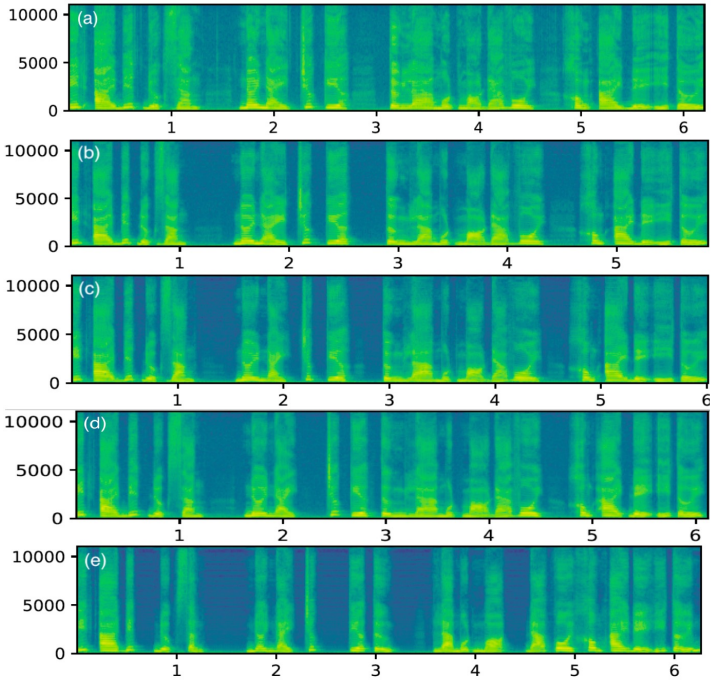
Encoder-only Models

- XPhoneBERT: Multilingual pre-training for phoneme representations for speech synthesis [12]
  - First large-scale pre-trained multilingual model for phoneme representations for speech synthesis in 94 languages and locales
  - Improve significantly the performance in the downstream speech synthesis task

Obtained results on the Vietnamese test set

	Model	MOS (↑)	MCD (↓)	RMSE_{F_0} (↓)
	Ground truth	4.26 ± 0.06	0.00	0.00
100%	Baseline VITS	3.74 ± 0.08	5.41	249
100%	VITS w/ XPB	**3.89** ± 0.08	**5.12**	**234**
5%	Baseline VITS	1.59 ± 0.05	6.20	291
5%	VITS w/ XPB	**3.35** ± 0.10	**5.39**	**248**

"XPB" abbreviates XPhoneBERT



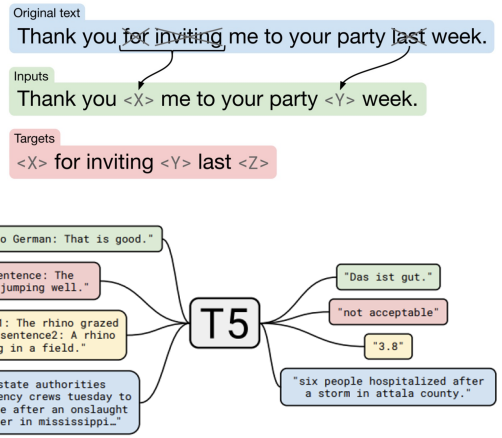
"Ít ai biết được rằng nơi này trước kia từng là một mỏ đá vôi không ai để ý tới"

(a): Ground truth; (b): VITS with XPhoneBERT (100%); (c): VITS with XPhoneBERT (5%); (d): Original VITS (100%); (e): Original VITS (5%)

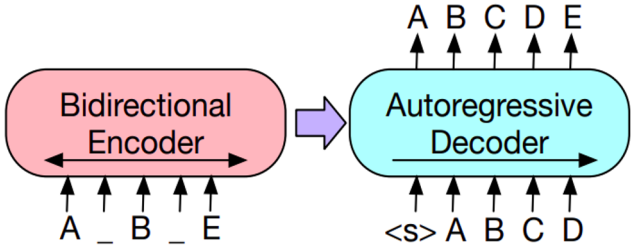
Encoder-Decoder Models

- Standard encoder-decoder Transformer-based sequence-to-sequence models, pre-trained on large-scale corpora with a denoising objective, e.g. BART [13], T5 [14], ByT5 [15]
- Pre-trained sequence-to-sequence models for Vietnamese:

Model	Type	Date	Data (vi)
mBART [16]	Multi.	01/2020	137 GB (25B syllables)
mT5 [17]	Multi.	10/2020	116 B syllables
BARTpho [18]	Mono.	09/2021	20 GB (4B syllables)
viT5 [19]	Mono.	07/2022	70 GB
enviT5 [20]	Biling.	10/2022	80 GB en & 80 GB vi
ViPubmedT5 [21]	Mono.	N/A	20 GB translated medical data



Figures taken from [16]



BART - Figure taken from [15]

Encoder-Decoder Models

- *vinai-translate*: Pre-trained BART-type Vietnamese-English translation model [22]
- *envit5-translation*: Pre-trained T5-type Vietnamese-English translation model [20]

SacreBLEU scores on the PhoMT test set

Model	English-to-Vietnamese	Vietnamese-to-English
vinai-translate	44.2	40.3
vinai-translate (v2)	**44.3**	**40.8**
envit5-translation*	44.6*	40.0*

*: Data leakage issue

	Model	Validation set		Test set	
		En2Vi	Vi2En	En2Vi	Vi2En
w/o FT	Google Translate	47.37	38.50	47.86	39.26
	ChatGPT 0-shot	34.38	29.79	34.45	30.39
	ChatGPT 1-shot	35.28	31.27	35.23	31.70
	ChatGPT 8-shot	36.09	31.87	36.02	32.57
	ChatGPT 16-shot	36.32	32.14	35.69	32.90
	ChatGPT 32-shot	34.92	32.08	36.37	32.94
	vinai-translate	44.24	33.28	44.60	33.44
	envit5-translation	42.86	31.33	43.23	32.00
FT	vinai-translate	**52.21**	**42.66**	**52.14**	**42.38**
	envit5-translation	51.14	41.47	51.27	41.17
	mBART	51.23	41.67	51.18	41.51
	envit5-base	50.10	40.66	49.94	40.36

Vietnamese-English medical translation

340K+ training; 9K validation; 9K test

Decoder-only Models

- Transformer decoder-based models, pre-trained on large-scale corpora with a standard language modeling objective, e.g. GPT-n [23-25], DialoGPT [26], LaMDA [27]

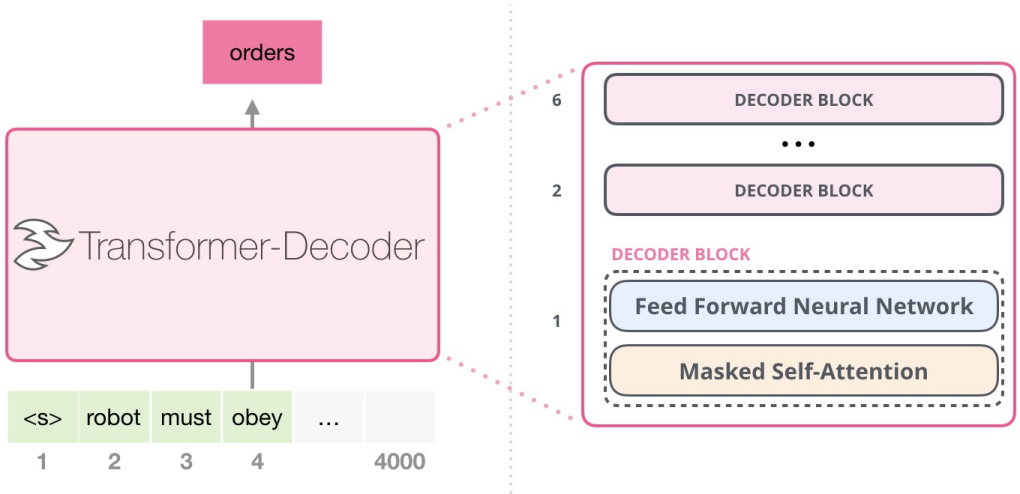


Image by Jay Alammar: https://jalammar.github.io/illustrated-gpt2/

Decoder-only Models

- “**Base**” pre-trained GPT-type models for Vietnamese

Model	Type	Date	Data (vi)
XGLM [28]	Multi.	12/2021	50 GB
BLOOM [29]	Multi.	07/2022	43 GB
gpt-j-6B-vietnamese-news	Mono.	09/2021	65 GB
Llama-2 [30]	Multi.	07/2023	~ 8 GB
Qwen [31]	Multi.	09/2023	N/A
PhoGPT-7B5 [32]	Mono.	11/2023	40 GB
SEA-LION-7B [33]	Multi.	11/2023	~300 GB
SeaLLM-7B-Hybrid [34]	Multi.	12/2023	N/A

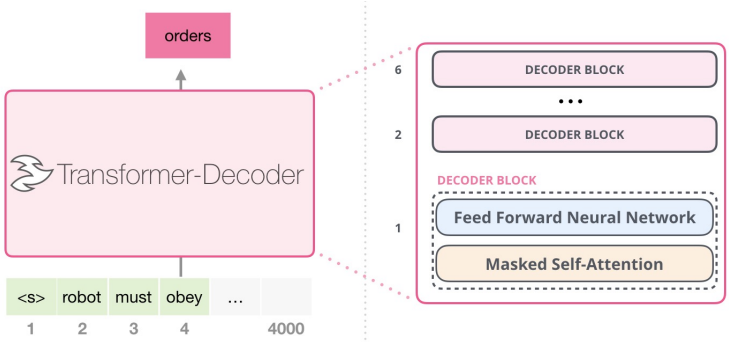


Image by Jay Alammar: https://jalammar.github.io/illustrated-gpt2/

PhoGPT

- **Motivation**
  - The success of decoder-only transformer-based generative LLMs stands out as one of the most significant achievements in recent AI research and development
    - ChatGPT/GPT-4, LLaMA/LLaMA-2, Mistral and Falcon
    - Largely limited to English

  - For Vietnamese, one might consider using the pre-trained multilingual BLOOM/BLOOMZ or the pre-trained monolingual “gpt-j-6B-vietnamesenews”

Model	Type	Date	Data (vi)
BLOOM/BLOOMZ [29]	Multi.	07/2022	43 GB
gpt-j-6B-vietnamese-news	Mono.	09/2021	65 GB

PhoGPT

- **Motivation**
  - For Vietnamese, one might consider using the pre-trained multilingual BLOOM/BLOOMZ or the pre-trained monolingual “gpt-j-6B-vietnamesenews”

Model	Type	Date	Data (vi)
BLOOM/BLOOMZ [29]	Multi.	07/2022	43 GB
gpt-j-6B-vietnamese-news	Mono.	09/2021	65 GB

  - BLOOM/BLOOMZ might not be a competitive baseline [35]
  - BLOOMZ-7B1 still performs significantly better than “gpt-j-6B-vietnamese-news” when fine-tuned for Vietnamese instruction following [36]

PhoGPT

- **Motivation**
  - BLOOMZ-7B1 performs significantly better than “gpt-j-6B-vietnamese-news” when fine-tuned for Vietnamese instruction following [36]

Model	Type	Date	Data (vi)
BLOOM/BLOOMZ [29]	Multi.	07/2022	43 GB
gpt-j-6B-vietnamese-news	Mono.	09/2021	65 GB

➔ Contradict the previously common notion that dedicated language-specific models outperform multilingual ones [6, 10, 11, 18]

➔ To reassess whether the previous common notion still holds w.r.t. LLMs:
Pre-train a similarly sized 7.5B-parameter LLM, namely PhoGPT, using closely similar Vietnamese pre-training data to BLOOM/BLOOMZ

PhoGPT

- *PhoGPT architecture*
  - A transformer decoder-based model
  - Incorporate flash attention [37] and ALiBi for context length extrapolation [38]
  - Utilize the existing multilingual BLOOM's tokenizer which supports Vietnamese well
  - Use a "max_seq_len" of 2048, "d_model" of 4096, "n_heads" of 32 and "n_layers" of 32, resulting in a model size of about 7.5B parameters

- *Pre-training Vietnamese data*: 41 GB of texts
  - 1 GB of Wikipedia texts and a 40 GB deduplicated variant of the "binhvq" news dataset

- *Pre-train PhoGPT from scratch*, employing the Mosaicml "llm-foundry" library

PhoGPT

- *Fine-tune the pre-trained PhoGPT for instruction following,* using a dataset consisting of 150K Vietnamese prompt and response pairs
  - 67K pairs from the Vietnamese subset of Bactrian-X [39]
  - 40K ShareGPT pairs without code and math, translated from English to Vietnamese by using VinAI Translate
  - 40K prompts covering hate, offense, toxicity, and safety awareness, largely including Vietnamese-translated ones
  - 1000 pairs for context-based question answering, 500 for poem writing, 500 for essay writing, 500 for spelling correction, and 500 for single-document summarization

- *Open-source PhoGPT for Vietnamese*
  - **PhoGPT-7B5**: “Base” pre-trained monolingual model
  - **PhoGPT-7B5-Instruct**: Instruction following model

Evaluation

- Conduct a human evaluation experiment to compare *PhoGPT-7B5-Instruct* with the closed-source *ChatGPT (gpt-3.5-turbo)* and other open-source instruction-following models, including *Vietcuna-3B, Vietcuna-7B-v3, URA-LLaMA-7B* and *URA-LLaMA-13B*
  - Vietcuna-3B and Vietcuna-7B-v3 are initially continually pre-trained from BLOOMZ-3B and BLOOMZ-7B1 on 12GB of Vietnamese news texts for causal language modeling, and then further fine-tuned with 200K samples of instructional question and answer pairs, alongside 400K samples of conversations
  - URA-LLaMA is continually pre-trained from LLaMA-2 on Vietnamese Wikipedia and news data, and then also fine-tuned for instruction following
  - Both the “base” pre-trained Vietcuna and URA-LLaMA, representing advanced variants of BLOOMZ and LLaMA-2 for Vietnamese, are not publicly available

- Utilize the greedy search decoding method, which is more suitable for LLM comparison [40]

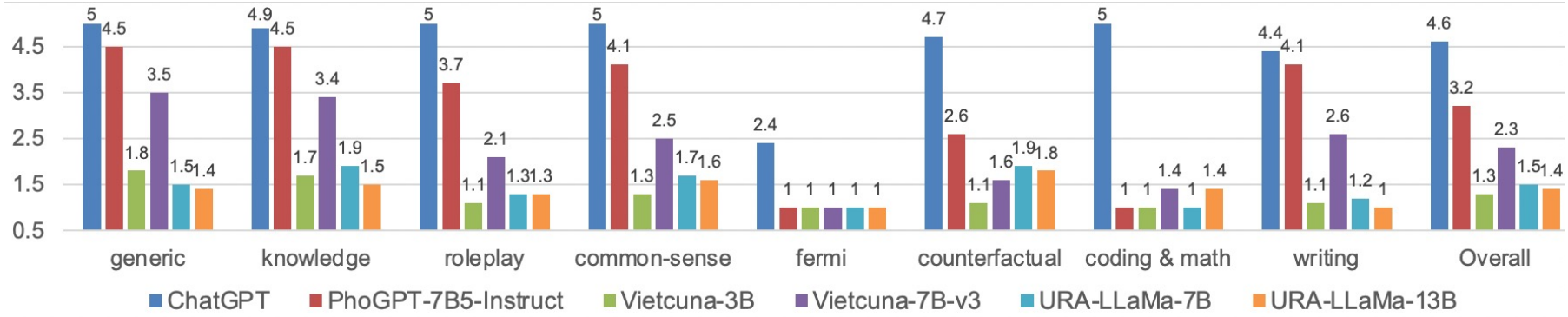
Evaluation

- The Vicuna question benchmark is manually translated into Vietnamese to create 80 evaluation questions from 8 different categories
  - Each question is fed into the 6 experimental models to generate responses, which are then anonymously shuffled

- BLOOMZ-7B1-mt, an instruction following variant of BLOOM/BLOOMZ-7B1 fine-tuned with 1.7M Vietnamese-translated examples, does not produce informative textual responses using the greedy search decoding method for the Vietnamese-translated Vicuna questions
  - BLOOMZ-7B1-mt is excluded from our evaluation

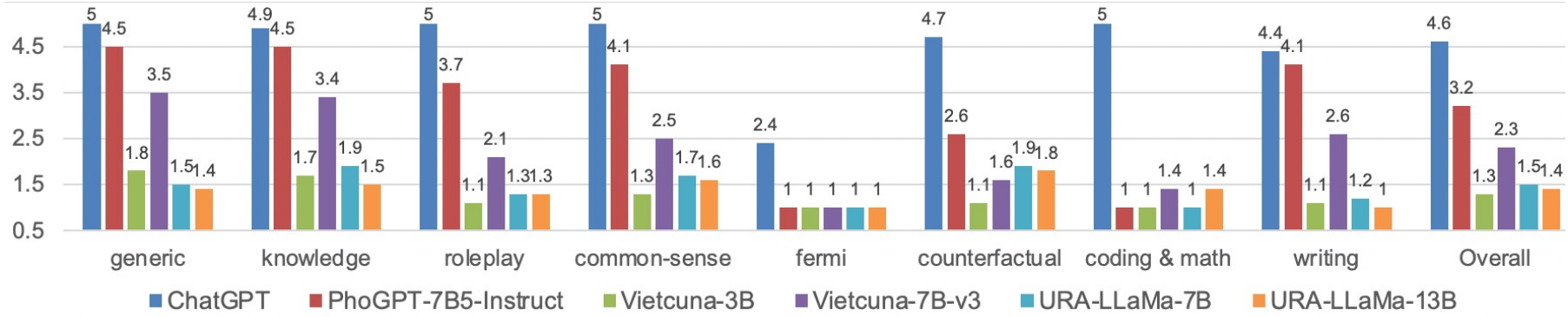
- Each generated response is then independently assessed by 3 annotators on a scale from 1 - Bad (e.g. wrong answers), 2 - Poor (e.g. partially answering the question), 3 - Fair, 4 - Good, to 5 – Excellent
  - Host a discussion session with the annotators to resolve annotation conflicts

Evaluation



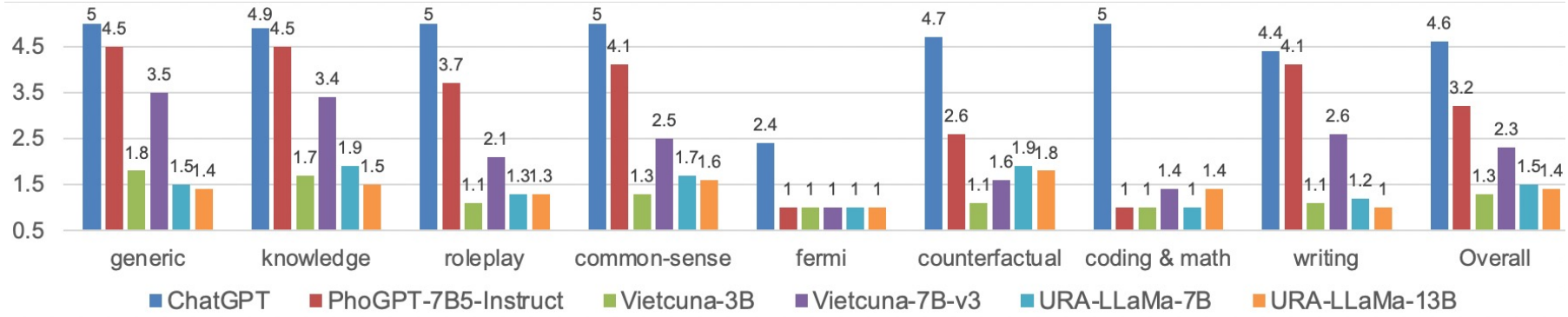
- PhoGPT-7B5-Instruct is strongly competitive compared to ChatGPT for “generic”, “knowledge”, “common-sense”, and “writing” questions
- In the “femi” category, all open-source models perform badly
- PhoGPT-7B5-Instruct substantially outperforms previous open-source instruction-following baselines for Vietnamese, except in the “coding & math” category where Vietcuna-7B-v3 and URA-LLaMA-13B perform better than PhoGPT-7B5-Instruct

Evaluation



- For “coding & math”, Vietcuna-7B-v3 and URA-LLaMA-13B perform better than PhoGPT-7B5-Instruct
  - For PhoGPT-7B5-Instruct, these results are anticipated due to the lack of “coding & math” and “femi”-type data in our pre-training Vietnamese corpus
  - Some “good” responses from Vietcuna-7B-v3 and URA-LLaMA-13B for “coding & math” questions share identical text with those from ChatGPT
    - These question-response pairs may coincidentally appear in the instruction-following datasets used for fine-tuning Vietcuna and URA-LLaMA

Conclusion



- *Open-source LLMs for Vietnamese with 7.5B parameters*
  - **PhoGPT-7B5**: “Base” pre-trained monolingual model
  - **PhoGPT-7B5-Instruct**: Instruction following model
  - A new foundation model pre-trained from scratch
  - SOTA open-source Vietnamese LLM
  - https://github.com/VinAIResearch/PhoGPT

VinAI

Thank you!

@VinAI

https://www.vinai.io/

HIRING PROCESS

BATCH 11 RECRUITMENT

VinAI

Nov 15 - Jan 5 2023 2024	Application Opening CV Screening
Jan 11 - 14	Entrance Test
Jan 22 - 28	Video Presentation
Jan 29 - Feb 4	Coding Interview
Feb 15 - 24	Research Interview

Qualitative Comparison

So sánh ảnh hưởng của việc dùng **chai nhựa sử dụng một lần** và việc dùng **chai nhựa có thể tái sử dụng** đối với môi trường và sức khỏe con người

PhoGPT-7B5-Instruct

Việc sử dụng chai nhựa một lần và chai nhựa có thể tái sử dụng đều có tác động đến môi trường và sức khỏe con người. Dưới đây là so sánh giữa hai loại chai này:

1. Tác động đến môi trường:
- Chai nhựa một lần sử dụng thường được đặt trên đất, cát hoặc đáy biển, góp phần vào việc gây ô nhiễm môi trường. Chúng giải phóng các chất độc hại vào môi trường, gây tổn hại cho các sinh vật sống trong môi trường nước và đất.
- Chai nhựa có thể tái sử dụng thường được làm từ polyethylene, một loại nhựa không gây ra tác động tiêu cực cho môi trường. Tuy nhiên, việc tái sử dụng chai nhựa có thể giảm lượng rác thải và tiết kiệm tài nguyên tự nhiên.

2. Tác động đến sức khỏe con người:
- Chai nhựa có thể chứa các chất độc hại như bisphenol A (BPA) và phthalates, có thể gây ảnh hưởng đến sức khỏe con người khi tiếp xúc lâu dài. Khi uống nước từ chai nhựa, các chất này có thể đi vào cơ thể thông qua quá trình tiêu hóa.
- Việc tái sử dụng chai nhựa có thể giảm lượng rác thải và tiết kiệm tài nguyên tự nhiên. Tuy nhiên, việc này không loại bỏ hoàn toàn các chất độc hại từ chai nhựa.

Tóm lại, việc sử dụng chai nhựa một lần và chai nhựa có thể tái sử dụng đều có tác động tiêu cực đến môi trường và sức khỏe con người. Để giảm thiểu tác động này, chúng ta cần tìm cách tái sử dụng và tái chế chai nhựa một cách hiệu quả.

Vietcuna-7B-v3

Việc sử dụng chai nhựa sử dụng một lần có tác động tiêu cực đến môi trường và sức khỏe con người. Dưới đây là một số lý do và hậu quả:

1. Tác động môi trường: Việc sử dụng chai nhựa sử dụng một lần góp phần gây ô nhiễm môi trường do chúng không thể tái sử dụng. Chúng thường được làm từ các vật liệu không thể phân hủy sinh học, chẳng hạn như polyethylene (PE) và polypropylene
(PP), có thể gây hại cho môi trường nếu không được xử lý đúng cách. Ngoài ra, việc sản xuất và sử dụng chai nhựa sử dụng một lần cũng gây ra các vấn đề về rác thải, làm tăng nguy cơ ô nhiễm và gây hại cho hệ sinh thái.

2. Tác động sức khỏe: Việc sử dụng chai nhựa sử dụng một lần cũng có thể gây hại cho sức khỏe con người. Chúng có thể chứa các hóa chất độc hại như bisphenol A
(BPA) và phthalates, có thể gây ra các vấn đề sức khỏe như ung thư, dị tật bẩm sinh và các vấn đề về hormone. Ngoài ra, việc tái sử dụng chai nhựa có thể dẫn đến việc sử dụng quá mức và lãng phí, làm tăng nguy cơ ô nhiễm và gây hại cho môi trường.

Tóm lại, việc sử dụng chai nhựa sử dụng một lần có tác động tiêu cực đến môi trường và sức khỏe con người.
Thay vào đó, sử dụng chai nhựa có thể tái sử dụng có thể góp phần bảo vệ môi trường và giảm tác động tiêu cực đến sức khỏe con người.

Qualitative Comparison

Các chính phủ có thể sử dụng các chính sách tài khóa và tiền tệ để chống lại suy thoái kinh tế như thế nào?

PhoGPT-7B5-Instruct

Có một số cách mà các chính phủ có thể sử dụng chính sách tài khóa và tiền tệ để chống lại suy thoái kinh tế. Dưới đây là một số ví dụ:

1. Tăng chi tiêu công: Chính phủ có thể tăng chi tiêu công để hỗ trợ nền kinh tế bằng cách đầu tư vào cơ sở hạ tầng, giáo dục, y tế và quốc phòng. Điều này giúp tạo ra công việc và kích thích tiêu dùng, từ đó giúp phục hồi nền kinh tế.

2. Giảm thuế: Chính phủ có thể giảm thuế cho doanh nghiệp và cá nhân để kích thích tiêu dùng và đầu tư. Điều này giúp tạo ra công việc và thu nhập, từ đó giúp phục hồi nền kinh tế.

3. Tăng chi tiêu của chính phủ cho vay: Chính phủ có thể cho vay tiền cho các doanh nghiệp và cá nhân để giúp họ phục hồi và phát triển kinh doanh. Điều này giúp tạo ra công việc và kích thích tiêu dùng, từ đó giúp phục hồi nền kinh tế.

4. Tăng thu nhập: Chính phủ có thể tăng thu nhập từ thuế hoặc các nguồn khác để đầu tư vào các dự án công cộng và cải thiện chất lượng cuộc sống của người dân. Điều này giúp tạo ra công việc và kích thích tiêu dùng, từ đó giúp phục hồi nền kinh tế.

5. Tăng dự trữ quốc gia: Chính phủ có thể tăng dự trữ quốc gia để giúp phục hồi kinh tế bằng cách đầu tư vào các công ty và doanh nghiệp. Điều này giúp tạo ra công việc và kích thích tiêu dùng, từ đó giúp phục hồi nền kinh tế.

Tuy nhiên, việc sử dụng chính sách tài khóa và tiền tệ để chống lại suy thoái kinh tế là một quyết định phức tạp và cần được xem xét kỹ lưỡng. Các biện pháp như tăng chi tiêu công, giảm thuế, tăng dự trữ quốc gia có thể giúp phục hồi nền kinh tế trong ngắn hạn, nhưng chúng cũng có thể gây ra lạm phát và mất cân bằng kinh tế.

Vietcuna-7B-v3

Các chính phủ có thể sử dụng các chính sách tài khóa và tiền tệ để chống lại suy thoái kinh tế bằng cách thực hiện các chính sách như tăng thuế, cắt giảm chi tiêu, in tiền và thực hiện các biện pháp kích thích kinh tế. Dưới đây là một số cách chính mà các chính phủ có thể thực hiện:

1. Tăng thuế: Các chính phủ có thể tăng thuế để thu thêm tiền vào ngân sách. Điều này có thể bao gồm tăng thuế doanh nghiệp, thuế thu nhập và thuế bất động sản. Tăng thuế có thể giúp tăng thu nhập và giảm chi tiêu, dẫn đến giảm chi tiêu và tăng tiết kiệm.

2. Giảm chi tiêu: Các chính phủ có thể giảm chi tiêu để giảm chi tiêu và tăng tiết kiệm. Điều này có thể bao gồm cắt giảm chi tiêu cho các dịch vụ công cộng như giáo dục, chăm sóc sức khỏe và an sinh xã hội. Giảm chi tiêu có thể giúp giảm chi tiêu và tăng tiết kiệm.

3. In tiền: Các chính phủ có thể in tiền để cung cấp tiền mới cho nền kinh tế. Điều này có thể bao gồm in tiền để hỗ trợ nền kinh tế hoặc cung cấp hỗ trợ tài chính cho các doanh nghiệp và cá nhân. In tiền có thể giúp tăng cung tiền và giảm chi phí vay, dẫn đến giảm chi tiêu và tăng tiết kiệm.

4. Thực hiện các biện pháp kích thích kinh tế: Các chính phủ có thể thực hiện các biện pháp kích thích kinh tế như giảm thuế, tăng chi tiêu và cung cấp hỗ trợ tài chính cho các doanh nghiệp và cá nhân. Các biện pháp kích thích kinh tế có thể giúp tăng chi tiêu và giảm chi tiêu, dẫn đến giảm chi tiêu và tăng tiết kiệm.

Bằng cách thực hiện các chính sách này, các chính phủ có thể giúp chống lại suy thoái kinh tế bằng cách tăng thuế, giảm chi tiêu, in tiền và thực hiện các biện pháp kích thích kinh tế.

References

1. Attention Is All You Need
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
3. RoBERTa: A Robustly Optimized BERT Pretraining Approach
4. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators
5. Unsupervised Cross-lingual Representation Learning at Scale
6. PhoBERT: Pre-trained language models for Vietnamese
7. Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models
8. VNLawBERT: A Vietnamese Legal Answer Selection Approach Using BERT Language Model
9. ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining
10. ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing
11. Investigating Monolingual and Multilingual BERTModels for Vietnamese Aspect Category Detection
12. XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech
13. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
14. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
15. ByT5: Towards a token-free future with pre-trained byte-to-byte models
16. Multilingual Denoising Pre-training for Neural Machine Translation
17. mT5: A massively multilingual pre-trained text-to-text transformer
18. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese
19. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation

References

20. MTet: Multi-domain Translation for English and Vietnamese
21. Enriching Biomedical Knowledge for Low-resource Language Through Translation
22. A Vietnamese-English Neural Machine Translation System
23. Improving Language Understanding by Generative Pre-Training
24. Language Models are Unsupervised Multitask Learners
25. Language Models are Few-Shot Learners
26. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation
27. LaMDA: Language Models for Dialog Applications
28. Few-shot Learning with Multilingual Language Models
29. BigScience Large Open-science Open-access Multilingual Language Model
30. Llama 2: Open Foundation and Fine-Tuned Chat Models
31. QWEN TECHNICAL REPORT
32. PhoGPT: Generative Pre-training for Vietnamese
33. SEA-LION (Southeast Asian Languages In One Network): A Family of Large Language Models for Southeast Asia
34. SeaLLMs - Large Language Models for Southeast Asia
35. How well do Large Language Models perform in Arithmetic tasks?
36. Efficient Finetuning Large Language Models For Vietnamese Chatbot
37. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness
38. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation
39. Bactrian-X : A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation
40. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models