CANTeam at HOMO-MEX 2024: Hate Speech Detection Towards the Mexican Spanish Speaking LGBT+ Population with Large Language Model

Le Minh Quan[1,2,*], Bui Hong Son[1,2] and Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper outlines our system for the three sub-tasks in the HOMO-MEX (Hate speech detection towards the Mexican Spanish speaking LGBT+ population) shared task at IberLEF 2024. To tackle this challenge, we developed a different approach based on fine-tuning Large Language Models with the LoRA technique for Task 1 (Multi-class Hate speech detection), Task 2 (Multi-label Fine-grained hate speech detection) and Task 3 (Binary classification Homophobic lyrics detection). LoRA (Low-Rank Adaptation) is a technique for parameter efficiently fine-tuning large language models. It significantly reduces training time and memory usage by using smaller, trainable matrices instead of modifying the entire model. This enables us to run Llama-2 on less powerful hardware. For all three tasks, we propose a fine-tuning system of the Llama-2 model by Meta AI. Our work ranked 2nd on Task 1, 1st on Task 2 and 8th on Task 3. Achieving 0.8775, 0.9730 and 0.4875 with F1 scores, respectively, for each task.

Keywords

Llama 2, Large Language Model, LoRA, Fine-tuning LLM, Prompting Engineering, IberLEF 2024, HOMO-MEX 2024

1. Introduction

The rise of hate speech online targeting the LGBT+ community is still a critical issue. This prejudice, known as LGBT+ Phobia, refers to all kinds of discrimination against the LGBT+ population on the basis of their sexual preferences and/or gender identities. With the rapid growth of social networks, LGBT+ Phobia becomes a larger problem as hateful content proliferates, normalizing discrimination and indoctrinating people with harmful ideologies. To address this growing issue of hateful content targeting LGBT+ communities online, the HOMO-MEX (Hate speech detection towards the Mexican Spanish-speaking LGBT+ population) shared task was established. This initiative aims to develop and improve automatic detection systems designed for the classification of hate speech directed at the Mexican LGBT+ community.

The HOMO-MEX 2024 shared task as part of IberLEF 2024 [1] targets researchers in natural language processing, hate speech detection, LGBTQ+ advocacy, music analysis, and content moderation to find solutions for the detection of LGBT+ phobic messages in social content [2, 3]. The shared task has proposed three different classification tasks as below:

- **Task 1** is a multi-class classification task. The objective of this task is to predict the label of each individual tweet as Phobic, Not Phobic or Not Related to LGBT+ Phobic.
- **Task 2** is a multi-label classification task, The objective of this task is to predict type(s) of LGBT+ Phobia in each tweet that contains LGBT+phobic hate speech.
- **Task 3** is a binary classification task. The objective of this task is to predict if a phrase of a song's lyrics contains LGBT+phobic hate speech.

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
20520709@gm.uit.edu.vn (L. M. Quan); 22521246@gm.uit.edu.vn (B. H. Son); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0000-0001-8340-1405 (D. V. Thin)

Large Language Models (LLMs) are more and more excelling at complex reasoning tasks across diverse fields, including in specialized domains such as creative writing and programming. Generative LLMs like ChatGPT and Gemini, with their intuitive chat interfaces, have driven widespread public adoption, mostly for educational and work purposes. For that reason, we decided to use a Generative approach using Llama 2 instead of traditional LLMs. Llama 2 has demonstrated its competitiveness with other existing open-source chat models, as well as a competency that is equivalent to some proprietary models [4, 5].

The rest of the paper is organized into 4 sections. In section 2 we will introduce the proposed pipeline for fine-tuning Llama 2. Section 3 details the experimental setup, including dataset, evaluation metric and the system setting. The results obtained in the evaluation phase are shown in section 4. Finally, our conclusion based on the result will be in section 5.

2. Related Work

Hate speech detection is the task of identifying if a textual content contains hatred and encourages violence towards a person or group of people, typically based on prejudice against sexual orientation, gender and ethnicity. Hate speech nowadays usually happens on social media platforms, including Facebook and Twitter.

Classic methods for hate speech detection involve using a Dictionary, Bag-of- word, feature extraction or embedding techniques to represent text data. These representations are then used to train classification algorithms such as SVM, Naive Bayes, and Logistic Regression algorithms [6, 7, 8].

A common approach to hate speech detection tasks are utilizing Deep Neural Network methods. [9] proposed a GRU-CNN model that combines Convolution Neural Network (CNN) with Gated Recurrent Units (GRU) to detect hate speech on Twitter. [10] utilized Bi-GRU-LSTM-CNN architecture for hate speech on Vietnamese text.

The introduction of BERT [11] marked the start of the rise of transformer-based language models. Following BERT's architecture, many pre-trained transformer models have appeared to extend its capability. BERTweet[12], a pre-trained language model specifically in tweets data, serves as a strong baseline for future research on Tweet analysis and classification tasks, especially those involving prevalent hate speech. HateBERT[13] is a domain-specific model focused on hate speech. This model is pre-trained on a large-scale dataset of social media posts in English, focusing on posts from communities banned for being offensive, abusive, or hateful.

Recently, with the rise of generative methods, many works have used LLMs for various tasks, such as hate speech detection. [14] use GPT models to understand bias and generate underlying explanations on hate speech. [15] demonstrates a Chain-of-thought prompting technique, and [16] introduces a new zero-shot prompting method to elevate generative performance. More powerful LLMs, namely GPT-4o and Llama-3, are still in development and are expected to offer improved accuracy and efficiency in various tasks, including hate speech detection.

3. Approach

3.1. Overview

Llama, which stands for Large Language Model Meta AI, is a Large Language Model developed by Meta AI. Meta AI's Llama 2 model is the successor to Llama, released only 4 months prior. The model features improvements in training technique and fine-tuning methods, a 40 % larger pre-training corpus size and doubled context length. These improvements have led Llama 2 to surpass its predecessor in various tasks, including Reasoning, Coding and Knowledge. Llama 2 offering varying parameter size (from 7 billion to 70 billion), 7B Llama 2 is the fastest model but worst performance, 70B Llama 2 offer the best performance but at the cost of slower processing speeds. Furthermore, Llama 2 comes with

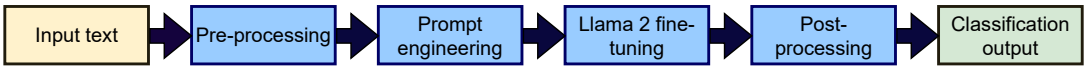

Figure 1: Overall pipeline using Llama 2 for HOMO-MEX shared task.

a "Chat" version, which has been fine-tuned specially for dialogue use cases, enabling it to provide natural conversational responses.

Figure 1 shows our step-by-step approach using Llama 2 for all tasks. First, the input text, which can be tweets or lyrics, is pre-processed before combining it with the instruction prompt. Next, we fine-tune the pre-trained Llama 2 model to generate labels for the task. Generated labels, which are in natural language, are then converted to numeric labels that match the official submission format. Below is the detailed description for each stage of the pipeline:

- **Pre-processing:** In order to improve data readability and model efficiency, we perform data pre-processing for each task:
  - **Task 1:** For the tweet dataset used by Task 1, we convert hashtags to separate words. For example, the hashtag "#BracketBusted" will be converted to "Bracket Busted".
  - **Task 2:** The process of pre-processing in Task 2 is similar to Task 1. In addition, we convert the multi-label into a natural language label. For example, the original label [0,1,0,1,0,0] will be converted to "GAY, TRAN".
  - **Task 3:** Due to the large size of each lyric in the Lyrics dataset, we truncate each data sample to the first 1,000 words to improve system efficiency.
- **Prompting:** Our observation shows that providing more information, such as the label's description to the prompt, leads to better model performance. Additionally, having distinct separators between each part of the prompt helps to clarify the instructions and makes them easier for the model to understand. A special token [INST] is utilized to separate the input prompt and answer segments. Prompts used for each task are shown in Table 1, 2, 3.
- **Fine tuning:** We fine-tune the pre-trained Llama 2 with the LoRA method. The version we use is Llama 2 chat 7B, the fastest model in the Llama 2 family and has been fine-tuned specifically for dialogue. For each task, we use a different value of training hyperparameters.
- **Post-processing:** Post-processing involves converting the output label to the submission format, which is only required for Task 2. This step transforms the label from natural language to a numerical multi-label format.
- For comparison, we used XLM RoBERTa and Multilingual T5 model as follows:
  - **XLM RoBERTa** [17]: XLM-RoBERTa is a Large multilingual Model base on RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. XLM-RBERTa significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks.
  - **Multilingual T5** [18]: Multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages. T5 uses a basic encoder-decoder Transformer architecture as originally proposed by [19]. T5 is pre-trained on a masked language modelling "span-corruption" objective, where consecutive spans of input tokens are replaced with a mask token, and the model is trained to reconstruct the masked-out tokens.

3.2. Low-rank Adaptation

In this paper, we used the LoRA technique to fine-tune Llama 2. LoRA (Low-Rank Adaptation for Large Language Models) is a popular technique to fine-tune pre-trained Large Language models and diffusion models. LoRA allows us to train some dense layers in a neural network indirectly by optimizing rank decomposition matrices of the dense layers' change during adaptation instead, while keeping the

Table 1
Prompt engineering for Task 1.

Prompt:	[INST] Classify the sentiment of a tweet: "item" ## if the tweet directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality, output "P". ## if the tweet not include any hate speech against the LGBT+ population but do mention this community, output "NP". ## if the tweet not related in any way to the LGBT+ community, output "NR". OUTPUT only P, NP, or NR. Answer: [/INST]
Response:	NP

Table 2
Prompt engineering for Task 2.

Prompt:	[INST] Predict one or more labels of a tweet: "item" ## if the tweet contains hate speech directed at homosexual people who identify as female, output "LES" ## if the tweet contains hate speech directed at homosexual people who identify as male, output "GAY". ## if the tweet directed at people who attracted to more than one gender, output "BI". ## if the tweet against transgender, output "TRAN". ## if the tweet against other sexual and gender minorities, output "OTHER". ## if the tweet is not related in any way, output "NOT RELATED". OUTPUT only the labels, nothing else. Answer: [/INST]
Response:	OTHER

Table 3
Prompt engineering for Task 3.

Prompt:	[INST] Classify the sentiment of a following lyrics from a song: "item" ## if the lyrics directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality, output "P". ## if the lyrics not include any hate speech against the LGBT+ population but do mention this community, output "NP". Answer: [/INST]
Response:	NP

pre-trained weights frozen [20]. In short, LoRa reduces the number of trainable parameters, making the training process faster and less computational cost while maintaining strong performance on downstream tasks.

4. Experimental Setup

4.1. Dataset

We use the dataset provided by the HOMO-MEX shared task. The corpus for Task 1 and 2 is composed of tweets in Mexican Spanish. For Task 3, the corpus is composed of lyrics of Spanish songs. Task 1 is a multi-class classification problem, the tweets are annotated as LGBT+phobic (P), not LGBT+phobic (NP), or not-related (NR). Task 2 is a multi-label classification problem; each tweet can have one or more of the following labels: Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), Other LGBT+phobia (O), Not LGBT+related (NR). Task 3 is a binary classification problem, and each of the lyrics is annotated as LGBT+phobic (P) or not LGBT+phobic (NP). The overall statistics of all three tasks are shown in Table 4 and 5. The dataset exhibits a class imbalance issue, particularly in Task 3.

Table 4
Classes statistic for three sub-tasks.

Task 1			Task 2			Task 3		
Class	**Train**	**Dev**	**Class**	**Train**	**Dev**	**Class**	**Train**	**Dev**
P	1072	862	L	88	72	P	39	40
NR	2246	1778	G	894	714	NP	945	560
NP	5482	4360	B	10	10	-	-	-
-	-	-	T	94	79	-	-	-
-	-	-	O	77	64	-	-	-
-	-	-	NR	0	0	-	-	-
N.o samples	8800	7000	N.o samples	1071	939	N.o samples	984	600

Table 5
Data statistics for three sub-tasks.

	Task 1			Task 2			Task 3		
Information	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
Max length	834	834	442	831	831	289	68256	68256	68256
Min length	7	7	11	8	8	12	116	145	116
Average length	125	125	146	96	96	95	2030	1508	2030
Number of tokens	184308	184308	54318	17314	13929	4257	377335	172378	377335
Number of vocabulary	43060	43060	16040	6707	5616	2162	43105	21692	43105

4.2. Evaluation Metric

The evaluation metrics for Task 1 and Task 3 are F1-score, Precision and Recall. These scores will be computed using the macro average. For Task 2 as a multi-label problem, the results are calculated based on the Sample average F1-score, Hamming loss and Exact match ratio.

4.3. System Setting

Our system code uses PyTorch framework and HuggingFace's transformers library [21]. Below is the fine-tuning setting for each model.

- **Llama 2**
  - **Training setting:** We use a learning rate of 2e-4 and a batch size of 4. For the optimizer, we use AdamW optimizer [22]. We fine-tune the model for 2 epochs for Task 1 and 5 epochs for both Task 2 and 3.
  - **LoRA setting:** For Causal Language Modeling, we configured LoRA with an attention dimension "r" of 8, an alpha parameter "LoRA_alpha" of 16, and a dropout probability "lora_dropout" of 0.05. For target modules, all trainable modules of Llama 2 were included: gate_proj, up_proj, down_proj, q_proj, k_proj, v_proj, and o_proj. With LoRA, the required training required reduce from axprox 7 billion to axprox 20 million, significantly reduced Training Time and Resources.
  - **Processing unit:** A100 80G GPU
- **XLM RoBERTa**
  - **Training setting:** We use a learning rate of 2e-5 and a batch size of 8. Fine-tuning with HuggingFace's trainer API and AdamW optimizer. For Task 1 and Task 3, we fine-tuned the model for 10 epochs. For Task 2, we fine-tune for 15 epochs.
  - **Processing unit:** P100 16G GPU
- **Multilingual T5**
  - **Training setting:** We use a learning rate of 3e-4 and a batch size of 8. Fine-tuning with HuggingFace's trainer API and AdamW for the optimizer. We fine-tune all Task 2 for 20 epochs and both Task 1 and 3 for 15 epochs.

Table 6
Performances of three models on the Task 1 development set. This table summarizes the precision, recall and F1-score for each class in Task 1, along with their average scores.

Models:		XLM-RoBERTa base			Multilingual T5 base			Llama 2 - chat 7B		
Metrics:		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Classes	NP	0.99	0.99	0.99	0.96	0.97	0.96	0.97	0.96	0.97
	P	0.98	0.96	0.97	0.89	0.83	0.86	0.89	0.87	0.88
	NR	0.99	0.99	0.99	0.96	0.98	0.97	0.94	0.99	0.96
Average (macro)		**0.99**	**0.98**	**0.98**	**0.94**	**0.92**	**0.93**	**0.93**	**0.94**	**0.94**

Table 7
Performances of three models on the Task 2 development set. This table summarizes the precision, recall and F1-score for each class in Task 2, along with their average scores.

Models:		XLM-RoBERTa base			Multilingual T5 base			Llama 2 - chat 7B		
Metrics:		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Classes	L	0.86	0.83	0.85	0.99	0.97	0.98	1.00	0.93	0.96
	G	0.99	0.97	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	B	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
	T	0.92	0.87	0.90	0.99	0.97	0.98	1.00	1.00	1.00
	O	0.00	0.00	0.00	0.91	0.92	0.91	0.95	0.91	0.93
	NR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average (samples)		**0.95**	**0.92**	**0.93**	**1.00**	**0.99**	**0.99**	**1.00**	**0.99**	**0.99**

Table 8
Performances of three models on the Task 3 development set. This table summarizes the precision, recall and F1-score for each class in Task 3, along with their average scores.

Models:		XLM-RoBERTa base			Multilingual T5 base			Llama 2 - chat 7B		
Metrics:		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Classes	NP	0.93	1.00	0.97	0.93	1.00	0.97	0.95	0.96	0.96
	P	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.30	0.33
Average (macro)		**0.47**	**0.50**	**0.48**	**0.47**	**0.50**	**0.48**	**0.66**	**0.63**	**0.65**

- **Processing unit:** P100 16G GPU

5. Result and Discussion

In this section, we present the model results and compare the performance of each model used in this research. Table 6, Table 7 and Table 8 details our evaluation results. Table 9 showcases Llama 2 performance on the test set, including its ranking on the official scoreboard with other participants for all three tasks.

XLM-RoBERTa achieved the highest F1 score of 0.98 on Task 1 of the development set, possibly because of its specialization in multilingual tasks. However, in Task 2, the model failed to predict labels "B" and "O" and resulting in zero F1 score. This is likely due to the class unbalance problem in multi-label training data. Since labels 'B' and 'O' only appear in 10 and 77 samples, respectively (as shown in Table 4), the model struggles to learn them effectively. In Task 3, our truncating techniques might have reduced some critical information in the dataset. This, along with the class imbalance, results in the model only predicting the "NP" label.

mT5 performs very well in Task 1 and Task 2 on the development set, achieving F1 scores of 0.93 and 0.99, respectively. Similar to XLM-RoBERTa, mT5 is trained specifically for multilingual tasks but with a larger parameter size. However, while it achieved good results in Task 2, mT5 still struggles in Task 3 and tends to predict all labels as "NP", neglecting the other class.

Llama 2 achieved good results across all metrics on Task 1 of the development set with an F1 score of 0.94 and the best result on Task 2 with an F1 score of 0.99. Despite being pre-trained only on English datasets, Llama 2 still demonstrates the best overall performance among the three models. This likely benefits from the massive scale of Llama-2, allowing it to capture broader linguistic patterns that might

Table 9
Performances of Llama-2 on the test set and ranking in the official scoreboard.

User	Task 1				Task 2				Task 3			
	F1-score	Presision	Recall	Ranking	F1-score	hamming-loss	exact-match-ratio	Ranking	F1-score	Presision	Recall	Ranking
Verbanex	0.9143	0.9364	0.8962	1	0.9393	0.0298	0.8880	4	0.5683	0.5575	0.6843	2
i2chuelva	0.8764	0.9098	0.8531	3	-	-	-	-	-	-	-	-
sdamians	0.8713	0.9194	0.84052	4	0.9435	0.0342	0.8470	3	0.4864	0.4794	0.4936	9
metztli	0.8562	0.8697	0.8457	5	0.9134	0.0366	0.8507	8	0.5667	0.5597	0.5766	3
Our result	**0.8775**	**0.9290**	**0.8476**	**2**	**0.9730**	**0.0149**	**0.9291**	**1**	**0.4875**	**0.4795**	**0.4957**	**8**

generalize somewhat to Spanish. Notably, while its F1 score in Task 3 is still low (at 0.65), Llama 2 is the only model that considers the minority class.

This result shows that Llama 2 and Multilingual T5 can better counter the imbalance of data and achieve better results compared to XLM-RoBERTa. The three model's performance fell short in Task 3. This could potentially be due to limitations in the data pre-processing techniques applied to the Lyrics dataset.

6. Conclusion

This paper describes the process of fine-tuning the pre-trained Large Language Model Llama 2 for the classification tasks in the HOMO-MEX shared task at IberLEF 2024. The tasks included hate speech detection, fine-grained hate speech detection and Homophobic lyrics detection. Llama 2 achieved 2nd and 1st on Tasks 1 and 2 of the official scoreboard. The results on Task 1 and Task 2 demonstrate that Llama 2 has the capability to tackle various classification tasks with high accuracy despite imbalanced datasets, thanks to its ability to reason and understand text. In future work, we would like to emphasise data quality more and employ more advanced data pre-processing techniques to improve the performance of large language models.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[2] G. Bel-Enguix, H. G'omez-Adorno, G. Sierra, J. V'asquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed toowards the mexican spanish speaking lgbtq+ population, Natural Language Processing 71 (2023).

[3] H. G'omez-Adorno, G. Bel-Enguix, H. Calvo, J. V'asquez, S. T. Andersen, S. Ojeda-Trueba, T. Alc'antara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Natural Language Processing 73 (2024).

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv e-prints (2023) arXiv–2307.

[6] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.
[7] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 27, 2013, pp. 1621–1622.
[8] P. Burnap, M. L. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics, EPJ Data science 5 (2016) 1–15.
[9] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, Springer, 2018, pp. 745–760.
[10] C. N. Vo, K. B. Huynh, S. T. Luu, T.-H. Do, Exploiting hatred by targets for hate speech detection on vietnamese social media texts, arXiv preprint arXiv:2404.19252 (2024).
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[12] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[13] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 2021, pp. 17–25.
[14] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5477–5490.
[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.
[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020.
[18] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).
[22] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).