GreenNode-LM: Fine-Tuning Large Language Models for Advancing Vietnamese Natural Language Understanding

Hoang Quoc Viet
GreenNode.ai
viethq5@greennode.ai

Trong-Hieu Nguyen-Mau
GreenNode.ai
hieunmt@greennode.ai

Vo Tien Dat
GreenNode.ai
datvt6@greennode.ai

Duong Anh Nghi
GreenNode.ai
nghida@greennode.ai

Pham Van Ngoan *
GreenNode.ai
ngoanpv@greennode.ai

Abstract

In the VLSP 2023 challenge on Vietnamese Large Language Models, our team is participating in improving and enhancing Large Language Models for the Vietnamese language. We focus on fine-tuning two versions of language models, the 7B and 14B models, called "greennode-7b"and "greennode-14b". After refining and optimizing our model, especially a model called "greennode-14b", it stood out as a leader, achieving top positions in seven different tasks in the competition's evaluation framework: ARC-vi, HellaSwag-vi, MMLU-vi, TruthfulQA-vi, ComprehensionQA-vi, Exams-vi, GeneralKnowledgeQA-vi. This success highlights the outstanding performance of our model across various areas. These top rankings in multiple tasks demonstrate our team's dedication to making advancements and achieving excellence in processing the Vietnamese natural language. For a detailed look at the leaderboard standings and rankings, please visit the following URL.

1 Introduction

In the landscape of Natural Language Processing (NLP), the ascendancy of Large Language Models (LLMs) like ChatGPT/GPT-4 [1], BARD [2], and LLaMA (Touvron et al., 2023) has precipitated monumental advancements across diverse linguistic domains globally. This burgeoning trend has spurred an augmented interest in crafting tailored LLMs for Vietnamese, reflecting a global enthusiasm to expand NLP capabilities into non-English languages. Despite this momentum, the development of Vietnamese-specific LLMs encounters a critical impediment: the scarcity of publicly accessible evaluation data.

The VLSP2023-VLLMs (Cuong et al.) initiative emerges as a pivotal stride towards overcoming this pivotal challenge by fostering the cultivation of large language models designed explicitly for Vietnamese. This initiative seeks to construct a comprehensive evaluation dataset tailored explicitly for Vietnamese LLMs, standing apart from traditional datasets aimed at downstream NLP tasks. This unique dataset revolves around assessing primary abilities across eight distinct skills, meticulously categorized into nine diverse domains, thereby facilitating a holistic evaluation framework.

The delineated primary abilities encompass facets such as logical thinking, background knowledge, commonsense understanding, problem handling, comprehension, insightfulness, metacognition, and user alignment. Spanning domains encompassing Humanities, Language, Social Science, History, Culture, Technology, Math, Natural Science, and Health, this exhaustive framework aspires to gauge the multifaceted capabilities of Vietnamese LLMs across diverse domains of human knowledge and understanding.

2 Related Work

2.1 Large Language Models

Language modeling, a key task in natural language processing, predicts token sequence probabilities. Large Language Models employ deep neural networks with transformer architectures (Vaswani et al., 2017).

Given a sequence of input tokens, or a prompt $s = \{s_1, s_2, \ldots, s_{k-1}\}$, LLMs compute the probability of the next token as $P(s_k | s_1, \ldots, s_{k-1})$. These models are also utilized in text generation

*Corresponding author: ngoanpv@greennode.ai

[1] https://openai.com/blog/chatgpt

[2] https://bard.google.com/

Dataset:	**Clean-Instruct-3M**
Examples:	"instruction": What is the process by which organisms release energy from glucose without the involvement of oxygen? "output": Anaerobic respiration or fermentation is the process by which organisms release energy from glucose without the involvement of oxygen.
Translated examples:	"instruction": Quá trình sinh vật giải phóng năng lượng từ glucose mà không cần oxy là gì? "output": Hô hấp kỵ khí hoặc lên men là quá trình sinh vật giải phóng năng lượng từ glucose mà không có sự tham gia của oxy.

Dataset:	**chat-v2-anthropic-helpfulness**
Examples:	"user": What kind of socks do you think would keep my feet most warm "assistant": For a man, I'd recommend wool socks.
Translated examples:	"user": Anh nghĩ loại vớ nào sẽ giữ ấm chân tôi nhất? "assistant": Đối với đàn ông, tôi khuyên nên đi tất len.

Table 1: Translated examples on various datasets.

through auto-regression, appending each predicted token to the sequence for subsequent token prediction.

For this, search strategies like beam search are often applied to produce the final output sequence (Brown et al., 2020). Notably, these methods can cause variations in LLM outputs for identical inputs over different runs.

Post the initial language modeling training, LLMs can be fine-tuned for diverse tasks. This involves updating a subset of the model's parameters (Radford et al., 2018). LLMs such as GPT-4 (OpenAI) have demonstrated exceptional adaptability, capable of task adaptation by incorporating a few prompt-based examples without needing parameter modifications (Radford et al., 2018).

2.2 Fine-tuning Large Language Models

Fine-tuning a Large Language Model entails first pre-training the model on an extensive corpus for language modeling purposes. It is then adapted to a specific task by training only some of the model's parameters with a dataset dedicated to that task. This paper investigates prevalent fine-tuning approaches for LLMs, focusing on Low-Rank Adaptation (LoRA) (Hu et al., 2021).

2.3 Vietnamese Large Language Models

While the domain of open-source models for the Vietnamese language is relatively nascent, there are already some notable models available. These include Vietcuna 3B[3], Vietcuna-7B-v3[4], URA-LLaMA-7B[5], and URA-LLaMA-13B[6]. Vietcuna-3B and Vietcuna-7B-v3 were developed from the foundational models BLOOMZ-3B[7] and BLOOMZ-7B1[8] (Scao et al., 2022), respectively, and were further trained using 12GB of Vietnamese news texts for causal language modeling[9]. This process included fine-tuning with 200K instructional question and answer pairs, and 400K conversational samples. The URA-LLaMA models, originating from LLaMA-2, were pre-trained on Vietnamese content from Wikipedia and online news sources, with additional fine-tuning for instruction following. Furthermore, Dat Quoc

[3] https://huggingface.co/vilm/vietcuna-3b

[4] https://huggingface.co/vilm/vietcuna-7b-v3

[5] https://huggingface.co/ura-hcmut/ura-llama-7b

[6] https://huggingface.co/ura-hcmut/ura-llama-13b

[7] https://huggingface.co/bigscience/bloomz-3b

[8] https://huggingface.co/bigscience/bloomz-7b1

[9] https://www.vilm.org/research/how-did-we-train-vietcuna

Nguyen *et al.*(Nguyen et al., 2023) have recently introduced the PhoGPT series, a new addition to the open-source generative models for Vietnamese, which includes a base 7.5B-parameter model and its instruction-following variant.

3 Methodology

3.1 Fine-tuned Datasets

We employed the Google Translation service to train our model on multiple instructional datasets in English, with a target translation into Vietnamese. The following five publicly available datasets from Hugging Face were utilized:

1. **Clean-Instruct-3M** (Crumb, 2023): This dataset comprises 3.09 million instructions samples. It combines alpaca-cleaned (Yahma, 2023) and GPT4All dataset (Crumb, 2023).

2. **chat-v2-anthropic-helpfulness** (Bcui19, 2023): This is a dataset derived from Anthropic's HH-RLHF (Bai et al., 2022) data of instructions and model-generated demonstrations. It includes 155,000 samples.

3. **Norquinal/claude_multiround_chat_30k** (Norquinal, 2023): This dataset results from 50,000 instruction/response pairs generated by Claude. For each base instruction, two additional follow-up instructions were added, totaling 30,000 instructions.

4. **TIGER-Lab/MathInstruct** (TIGER-Lab, 2023): MathInstruct is a carefully curated instructional tuning dataset. It is lightweight yet generalizable, compiled from 13 math rationale datasets, including six newly curated datasets.

Table 1 illustrates some examples of the above datasets.

3.2 LoRA Fine-tuning

LoRA, short for Low-Rank Adaptation of Large Language Models, introduces an efficient method for fine-tuning Large Language Models (LLMs) (Hu et al., 2021). This method is designed to optimize the fine-tuning process by selectively updating the most crucial parameters while keeping the remainder fixed. This approach has proven effective in reducing the computational and financial costs associated with adapting models with billions of parameters, such as GPT-3, to specific tasks or domains.

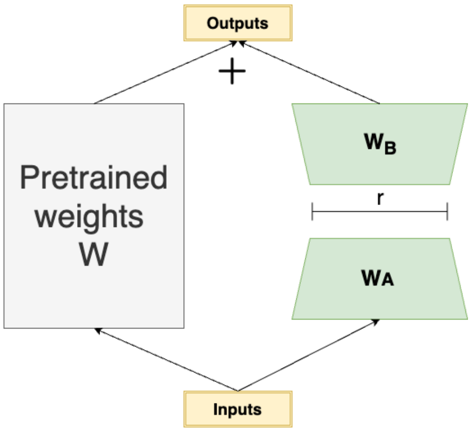

Figure 1: LoRA architecture

LoRA operates by freezing the pre-trained model weights and introducing trainable rank decomposition matrices into each layer of the Transformer architecture, substantially reducing the number of trainable parameters for downstream tasks. In Figure 1, during training, W is frozen and does not undergo gradient updates, while W_A and W_B contain trainable parameters.

3.3 Supervised Fine-tuning

QWEN (et al., 2023), developed by Alibaba Group, is a robust language model, particularly exemplified by its fine-tuned variant, QWEN-CHAT. Our choice of QWEN is driven by its extensive pretrained dataset, larger vocabulary size (compared to most models), and impressive compression rate in the Vietnamese context. This outperformance extends to models like LLaMA-7B (Touvron et al., 2023), Baichuan-7B, ChatGLM2-6B (Zeng et al., 2022), and InternLM-7B (Team, 2023).

To understand human behavior in Vietnamese, we perform Supervised Fine-tuning, refining QWEN on chat-style data with carefully selected, high-quality datasets, including translations through the Google Translation service.

3.4 Token Selection

Language models utilize the autoregressive sampling process for sequential text generation. This approach entails predicting and generating each word in a sequence while taking into account the context of preceding words. The challenge of determining the optimal top k value has led to the adoption of Nucleus Sampling, a popular decoding strategy. In this method, we choose a sufficient number of tokens to encompass a specified probability, denoted by the parameter top p through the following procedure:

1. Arrange the tokens in a descending order based on their probabilities.
2. Pick the minimum number of top tokens whose cumulative probability is equal to or exceeds the defined value of top p.
3. Perform sampling from this selected set of tokens.

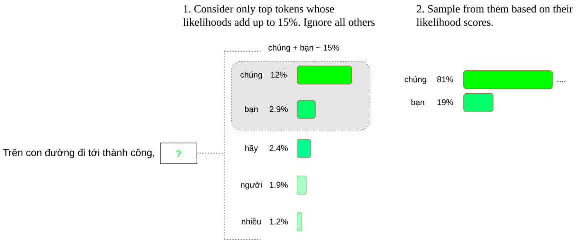
1. Consider only top tokens whose likelihoods add up to 15%. Ignore all others
2. Sample from them based on their likelihood scores.
chúng + bạn ~ 15%
chúng 12%
bạn 2.9%
hãy 2.4%
người 1.9%
nhiều 1.2%
Trên con đường đi tới thành công,
?
chúng 81%
bạn 19%
....

Figure 2: Top-p nucleus sampling.

Moreover, an additional determinant impacting the stochasticity of the model's outputs is the temperature parameter. Notably, its mode of operation diverges significantly from the two antecedent parameters. Whereas top p and top k exert direct influence on output probabilities, the temperature parameter exerts its impact on the softmax function itself:

$$\sigma(\overrightarrow{x})_i = \frac{e^{\frac{x_i}{T}}}{\sum_{j=1}^{n} e^{\frac{x_j}{T}}}$$

- If $T \rightarrow 0$, the system exhibits a propensity for extremely large exponentials. Consequently, the x_i element characterized by the highest numerical value will assert dominance, resulting in its probability nearing unity, while all other elements approach 0. This scenario is analogous to the adoption of a greedy strategy, wherein the top token consistently takes precedence.

- If $T \rightarrow \infty$, the exponentials converge to $e^0 = 1$. This transformation leads to the emergence of a uniform distribution in the output, where all probabilities become $\frac{1}{n}$, signifying equal likelihood for each token. It is evident that such a model ceases to be practical or meaningful in this context.

4 Experiment

4.1 Implementation Details

The standard AdamW (Loshchilov and Hutter, 2017) optimizer is employed with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, epsilon $\epsilon = 1e - 8$, and a learning rate set at $l = 5e - 5$. The training procedure integrates a cosine learning rate schedule, associating a specific peak learning rate with each model size. The learning rate undergoes decay, gradually decreasing to a minimum of 10% of the peak learning rate. To ensure training stability, all models are trained using BFloat16 mixed precision. Additionally, the sequence length is configured as 2048, and the training process extends over three epochs. For inference, the temperature is configured at 0.5, while the values for top p and top k are set to 0.95 and 40, respectively.

To fine-tune with LoRA, specific configurations are applied, including a LoRA rank of 16, and a lora_dropout set to 0.1. With this configuration, the 7B model learns on just 8,388,608 parameters, and the 14B model learns on 13,107,200 parameters, significantly smaller compared to the full parameter count.

All experiments for both models, greennode-7b and greennode-14b, were conducted on the same settings using an NVIDIA H100 GPU with 80GB of memory.

4.1.1 Evaluation framework

In this task from VLSP, we follow the framework for few-shot evaluation of autoregressive language models called Language Model Evaluation Harness from EleutherAI cusomized for Vietnamese dataset from VLSP (Gao et al., 2021).

4.2 Result

The evaluation of our fine-tuned models, greennode-7b and greennode-14b, involved comprehensive assessments across multiple benchmarks and tasks within the VLSP2023-VLLMs initiative, as shown in Figure 3 and Figure 4, detail in Table 2 and 3. Our models were rigorously evaluated on eight diverse tasks, each covering distinct domains of language understanding, knowledge comprehension, reasoning, and more. The evaluation outcomes highlighted the robustness and competence of our models across a spectrum of linguistic challenges. We have representative examples of greennode-14b's writing abilities in Table 4, demonstrating its proficiency across a variety of tasks.

We will deep analyze the results on each task-specific performance.

ARC-vi In the ARC-vi task, designed to evaluate AI systems' reasoning abilities, greennode-14b

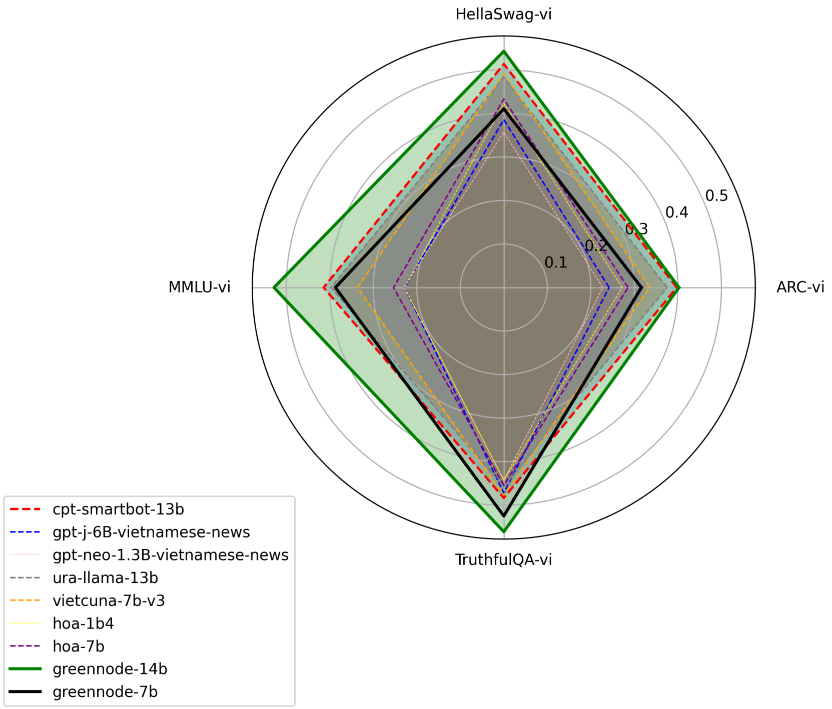

Figure 3: Performance of models on public test set including MMLU-vi, ARC-vi, TruthfulQA-vi and HellaSwag-vi.

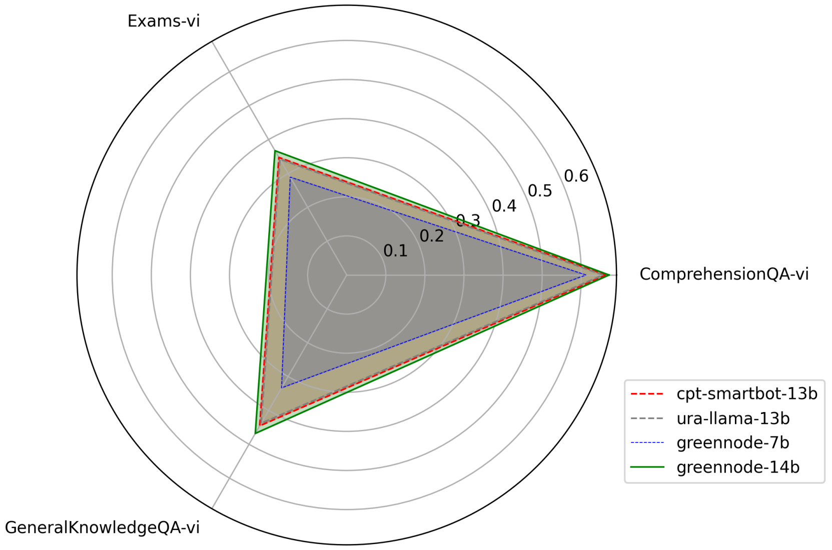

Figure 4: Performance of models on datasets Exams-vi, GeneralKnowledgeQA-vi and ComprehensionQA-vi.

demonstrated exceptional proficiency, achieving a top-1 ranking performance with $accuracy = 0.4026$. The model showcased adept logical reasoning and comprehension skills, establishing itself as a leader in this domain. The greennode-7b model received a score of $accuracy = 0.3162$. All tests were conducted with a 25-shot setting.

HellaSwag-vi In the HellaSwag-vi task, focused on commonsense reasoning, both greennode-7b and greennode-14b exhibited outstanding capabilities. Their prowess in understanding and reasoning through commonsense scenarios led to commendable standings in this benchmark. Specifically, with the greennode-14b model, we achieved an accuracy score of 0.5430, surpassing the nearest competitor by a notable margin of 2.9%. All assessments were carried out with a 10-shot configuration.

MMLU-vi In the MMLU benchmark, evaluating models in a 5-shot setting, both greennode-14b and greennode-7b demonstrated their capabilities.

Organization	Model	ARC-vi ↑	HellaSwag-vi ↑	MMLU-vi ↑	TruthfulQA-vi ↑
CPT_Smartbot	cpt-smartbot-13b	0.3974	0.5136	0.414	0.4836
ura-hcmut	ura-llama-13b	0.3752	0.483	0.3973	0.4574
vilm	vietcuna-7b-v3	0.335	0.4914	0.336	0.4771
vlsp-2023-vllm	hoa-7b	0.2855	0.4329	0.2536	0.4542
vlsp-2023-vllm	hoa-1b4	0.2718	0.4228	0.2281	0.4423
VietAI	gpt-j-6B-vietnamese-news	0.2419	0.3856	0.2282	0.4718
VietAI	gpt-neo-1.3B-vietnamese-news	0.2274	0.3567	0.229	0.4423
GreenNode.ai	greennode-7b (our)	0.3162	0.4106	0.387	0.5249
GreenNode.ai	greennode-14b (our)	**0.4026**	**0.543**	**0.5281**	**0.5612**

Table 2: Model Performance Comparison on Public Set

Model	ComprehensionQA-vi ↑	Exams-vi ↑	LAMBADA-vi ↓	GeneralKnowledgeQA-vi ↑
cpt-smartbot-13b	0.6633	0.3473	21.9864	0.4455
ura-llama-13b	0.6556	0.342	**17.5614**	0.438
greennode-7b (our)	0.6122	0.2892	189.7782	0.3335
greennode-14b (our)	**0.6711**	**0.3672**	29.5967	**0.468**

Table 3: Model Performance Comparison on Private Set

The greennode-14b model achieved an accuracy of 0.5281, showcasing strong performance in understanding and reasoning across a diverse range of subjects. Meanwhile, the greennode-7b model achieved an accuracy of 0.387, indicating a respectable level of proficiency in the task. These results provide insights into the models' ability to leverage pretraining knowledge in a challenging few-shot learning scenario, shedding light on their performance across various subjects in the MMLU dataset.

TruthfulQA-vi In the TruthfulQA-vi task, greennode-14b demonstrated remarkable proficiency with a score of 0.5612, surpassing competitors by showcasing a strong grasp of factual information and knowledge comprehension. Its ability to accurately process and retrieve factual information was a key factor in securing the leading position.

ComprehensionQA-vi GreenNode's 14B model outperformed cpt-smartbot-13b with a score of 0.6711 compared to 0.6633, showcasing slightly superior comprehension abilities. Meanwhile, ura-llama-13b obtained a score of 0.6556, trailing marginally behind GreenNode's 14B model. In contrast, GreenNode's 7B model scored 0.6122, displaying good but relatively lower comprehension performance than the other models.

Exams-vi GreenNode's 14B model demonstrated competitive performance (0.3672) compared to cpt-smartbot-13b (0.3473) and ura-llama-13b (0.342), showcasing its competency in handling exam-oriented questions. However, GreenNode's 7B model scored lower (0.2892) in this domain compared to the other models.

LAMBADA-vi In the LAMBADA-vi task, where lower scores indicate superior performance, GreenNode's 14B model achieved a score of 29.5967. However, this score was higher than the scores of cpt-smartbot-13b (21.9864) and ura-llama-13b (17.5614), indicating that GreenNode's 14B model didn't perform as well as these models in understanding contextual information within this specific task. Additionally, GreenNode's 7B model obtained a considerably higher score of 189.7782, further indicating its comparatively lower effectiveness in comprehending contextual information compared to the other models analyzed in this task.

GeneralKnowledgeQA-vi GreenNode's 14B model showcased slightly better performance in understanding general knowledge with a score of 0.468 compared to cpt-smartbot-13b (0.4455) and ura-llama-13b (0.438). However, GreenNode's 7B model scored lower (0.3335) compared to the other models in this domain.

Task:	**Brainstorming**
Prompt:	Bạn đang phát triển một ứng dụng di động, hãy nghĩ ra một tính năng độc đáo và hấp dẫn để tăng trải nghiệm người dùng.
Answer:	Một tính năng độc đáo và hấp dẫn để tăng trải nghiệm người dùng trong ứng dụng di động có thể là một công cụ tự động hóa các hoạt động thường xuyên của người dùng. Ví dụ, nếu ứng dụng được sử dụng cho việc quản lý tài chính, nó có thể tự động hóa việc theo dõi chi tiêu hàng ngày của người dùng, tạo báo cáo về ngân sách và cảnh báo khi vượt quá giới hạn chi tiêu. Điều này giúp giảm thiểu thời gian và nỗ lực cần thiết cho việc quản lý tài chính hàng ngày của người dùng, đồng thời cải thiện trải nghiệm người dùng bằng cách cung cấp thông tin dễ hiểu và hữu ích.
Task:	**Math**
Prompt:	Nam mua 5 quyển vở và đưa cô bán hàng tờ 100000 đồng. Cô bán hàng trả lại cho Nam 35000 đồng. Hỏi mỗi quyển vở Nam mua có giá là bao nhiêu? Hãy giải bài toán từng bước một.
Answer:	Tổng tiền mua vở = 100.000 - 35.000 = 65.000 đồng Số quyển vở = 5 quyển Vậy mỗi quyển vở Nam mua có giá là: 65.000 / 5 = 13.000 đồng
Task:	**Code**
Prompt:	Viết đoạn code Python mở một tệp có tên là "vlsp2023.txt" và in ra số lượng từ có trong tệp đó.

Answer:	Đây là một đoạn code Python mở một tệp có tên là "vlsp2023.txt" và in ra số lượng từ có trong tệp đó:
	`# Mở tệp`
	`with open('vlsp2023.txt', 'r') as file:`
	`    # Đọc nội dung và tính số từ`
	`    print(len(file.read().split()))`
	Kết quả sẽ là: Số lượng từ có trong tệp.

Table 4: Representative samples of greennode-14b's multifaceted writing proficiency across diverse tasks.

5 Limitations, Ethical Considerations and Future Work

Similar to other large language models, the GreenNode-LM model, available in 7B and 14B variants, faces challenges of bias and potential harmful content generation due to its training on publicly accessible datasets. These challenges are particularly pronounced in non-English languages such as Vietnamese, due to limited dataset availability. Despite these obstacles, we are dedicated to continuous improvement and plan to release updated versions with enhancements.

Recognizing the potential for AI model misuse is essential. To mitigate this, we have proactively fine-tuned GreenNode-LM to steer clear of sensitive topics like bioterrorism and cybercrime. Our ongoing efforts include further optimization of the model and establishing responsible use guidelines to reduce the risks of generating harmful content. We remain committed to responsible development and refinement of LLMs, understanding the importance of preventing misuse and bias in content generation.

6 Conclusion

Through the VLSP2023-VLLMs initiative, our team has significantly advanced Vietnamese Large Language Models, notably with our "greennode-14b"model. This initiative has been key in developing dedicated evaluation datasets and standard benchmarks for Vietnamese LLMs. Despite the challenges in fine-tuning non-English languages and addressing biases, we're committed to continuous improvement. "greennode-14b"in particular, has excelled in various tasks, achieving top rankings in multiple evaluations. This underscores not only the model's capabilities but also the broader challenges and potential in Vietnamese natural language processing. We look forward to further refining Vietnamese LLMs and fostering responsible, impactful advancements in this evolving field.

Acknowledgments

We express our deep appreciation to GreenNode.ai, our organizational and primary supporter throughout the VLSP2023-VLLMs. The pivotal role played by GreenNode.ai in providing essential resources, particularly high-performance GPUs, significantly accelerated the fine-tuning process of our models, contributing substantially to our success in the competition.

References

J. Bai et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bcui19. 2023. Chat-v2-Anthropic-Helpfulness. https://huggingface.co/datasets/bcui19/chat-v2-anthropic-helpfulness.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Crumb. 2023. Clean-Instruct-3M. https://huggingface.co/datasets/crumb/Clean-Instruct-3M.

Crumb. 2023. Gpt-4-all-clean dataset. https://huggingface.co/datasets/crumb/gpt4all-clean. Accessed: Ngày 28 tháng 11 năm 2023.

Le Anh Cuong, Nguyen Trong Hieu, Nguyen Viet Cuong, Nguyen Ngoc Que, Le-Minh Nguyen, and Cam-Tu Nguyen. Vlsp 2023 challenge on vietnamese large language models.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Nhung Nguyen, Thien Huu Nguyen, Dinh Phung, and Hung Bui. 2023. PhoGPT: Generative Pre-training for Vietnamese. *arXiv preprint*, arXiv:2311.02945.

Norquinal. 2023. claude_multiround_chat_30k. https://huggingface.co/datasets/Norquinal/claude_multiround_chat_30k.

OpenAI. Gpt-4 technical report.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. 2022. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

TIGER-Lab. 2023. MathInstruct Dataset. https://huggingface.co/datasets/TIGER-Lab/MathInstruct.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yahma. 2023. Alpaca cleaned dataset. https://huggingface.co/datasets/yahma/alpaca-cleaned. Accessed: Ngày 28 tháng 11 năm 2023.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.