GreenMind: A Next-Generation Vietnamese Large Language Model for Structured and Logical Reasoning

Luu Quy Tung[1] **Hoang Quoc Viet**[1*] **Pham Bao Loc**[1] **Vo Trong Thu**[2]
[1]GreenNode.ai [2]John Von Neumann Institute
{tunglq,viethq5,locpb}@greennode.ai, thuvt@jvn.edu.vn

Abstract

Chain-of-Thought (CoT) is a robust approach for tackling LLM tasks that require intermediate reasoning steps prior to generating a final answer. In this paper, we present **GreenMind-Medium-14B-R1** [1], the Vietnamese reasoning model inspired by the finetuning strategy based on Group Relative Policy Optimization. We also leverage a high-quality Vietnamese synthesized reasoning dataset and design two reward functions to tackle the main limitations of this technique: i) Language mixing, where we explicitly detect the presence of biased language characters during the process of sampling tokens, and ii) We leverage Sentence Transformer-based models to ensure that the generated reasoning content maintain factual correctness and do not distort the final output. Experimental results on the Vietnamese dataset from the VLSP 2023 Challenge demonstrate that our model outperforms prior works and enhances linguistic consistency in its responses. Furthermore, we extend our evaluation to SeaExam — a multilingual mutiple-choices dataset, showing the effectiveness of our reasoning method compared to few-shot prompting techniques.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has transformed the approach to handling complex tasks such as question answering and multiple-choice problems. Many open-source LLMs have demonstrated impressive capabilities in natural language understanding. However, for tasks like question answering and multiple-choice reasoning, the act of users prompting models to produce direct answers only often fails to ensure accuracy. Meanwhile, at each generation step, models rely on the probability distribution over a list of candidate tokens to select the potential one by greedy or random sampling algorithms. Consequently, producing only a short sequence of tokens as the final output does not guarantee correctness, as these distributions are conditioned solely on the preceding input tokens. This implies that the models often lack the contextual understanding necessary for reasoning toward a correct answer. To address this issue, the CoT (Wei et al., 2022b) technique remains an effective approach to fully leverage the power of next token prediction. CoT encourages the model to articulate a sequence of intermediate reasoning steps, which facilitates the resolution of tasks that require multi-step logical thinking. To further enhance the reasoning capabilities of language models, a series of reinforcement learning-based methods have been proposed. Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) leveraged human-provided feedback to refine LLM outputs, ensuring that the reasoning steps generated by CoT align more closely with human-like judgment and reasoning. Proximal Policy Optimization (PPO) balanced exploration and exploitation by updating the reasoning policy using a clipped objective function, which helps avoid large, destabilizing changes while enhancing CoT reasoning across multiple steps.

In this study, we introduce **GreenMind-Medium-14B-R1**, a fine-tuned LLM model capable of reasoning for tasks within the Vietnamese community. Our model leverages the GRPO technique (Shao et al.; Guo et al., 2025), which has been shown to enhance reasoning effectiveness for the CoT method as well as reduce computational

*Corresponding author: viethq5@greennode.ai
[1]https://huggingface.co/GreenNode/GreenMind-Medium-14B-R1

costs. However, the limitation of this approach is its inability to control for linguistic bias (typically English and Chinese) inherent in the base models, which means that generated responses may contain characters from the language with the dominant training dataset. Additionally, the quality control of the reasoning process has not been addressed in the original work (Guo et al., 2025), which may lead to content distortion relative to the original query. To tackle these challenges, we augment the synthesized sequences of reasoning steps for each sample in the training dataset by utilizing a state-of-the-art LLM for reasoning tasks. We then re-check the data based on the labels of each sample. We design two reward functions: one for language check, which uses a banned letter dictionary, and another for reasoning content, which employs Sentence Transformer models to measure the semantic similarity of the generated response compared to the corresponding reasoning data.

Our contributions are described as follows:

- We propose algorithms and utilize our Vietnamese reasoning dataset to address the issue of language bias and ensure strict control over the reasoning content.
- We release a Vietnamese reasoning model with a medium size, specifically a 14.7 billion parameters, achieving a high overall accuracy of over 70% on multiple-choice datasets, including the VLSP 2023 Challenge (Le et al., 2024) and SeaExam (Li et al., 2024).
- We also conduct experiments across multiple languages and demonstrate that reasoning-based answers significantly improve compared to few-shot learning techniques.

2 Related Work

2.1 Chain-of-Thought

Chain-of-Thought (CoT) prompting (Wei et al., 2022a) was introduced to encourage models to "think step by step", providing a few exemplars with intermediate reasoning steps to improve multi-step inference. Empirical results show that CoT significantly boosts performance on arithmetic, commonsense, and symbolic reasoning benchmarks, with a 540-billion-parameter model achieving state-of-the-art accuracy on GSM8K using just eight CoT exemplars. Follow-up work on self-consistency decoding samples multiple reasoning paths and selects the most consistent answer, yielding substantial gains on GSM8K +17.9%, SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%), and ARC-challenge (+3.9%) (Wang et al., 2022). These studies reveal that structured intermediate reasoning can be an emergent capability in sufficiently large models.

2.2 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[2], Vietcuna-7B-v3[3], URA-LLaMA-7B[4], and URA-LLaMA-13B[5]. Vietcuna-3B and Vietcuna-7B-v3 were developed from the foundational models BLOOMZ-3B[6] and BLOOMZ-7B1[7] (Scao et al., 2022), respectively, and were further trained using 12GB of Vietnamese news texts for causal language modeling[8]. 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, PhoGPT (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.

2.3 Group Relative Policy Optimization

Reinforcement learning (RL) is a subfield of Machine Learning (ML) in which an agent learns to make decisions through interactions with its environment, aiming to maximize cumulative rewards. When applied to LLMs, RL helps fine-tune these models to better align with human preferences and improve their performance on specialized tasks that require complex reasoning processes. A key category of RL algorithms is policy optimization, which focuses on directly refining the policy—the decision-making strategy an agent follows based on different states. GRPO was introduced in

[2]https://huggingface.co/vilm/vietcuna-3b
[3]https://huggingface.co/vilm/vietcuna-7b-v3
[4]https://huggingface.co/ura-hcmut/ura-llama-7b
[5]https://huggingface.co/ura-hcmut/ura-llama-13b
[6]https://huggingface.co/bigscience/bloomz-3b
[7]https://huggingface.co/bigscience/bloomz-7b1
[8]https://www.vilm.org/research/how-did-we-train-vietcuna

DeepSeekMath (Shao et al.), with the aim of improving the reasoning abilities of LLMs, especially in mathematical problem-solving and code generation. The reward serves as the foundation for the training signal, guiding the optimization direction in reinforcement learning. To train DeepSeek-R1-Zero (Guo et al., 2025), authors implemented a rule-based reward mechanism comprising two primary reward types:

- Format rewards: This function is used to evaluate the model's ability to generate responses that adhere to the desired structure.
- Accuracy rewards: This function is used to evaluate whether the extracted result (obtained from the response using a heuristic or structure-based method) matches the ground truth.

3 Vietnamese Reasoning Dataset

3.1 Problem Definition

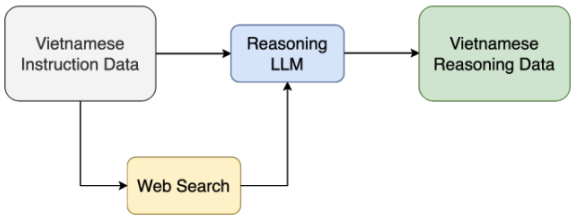

Figure 1: Reasoning Data Curation

In this section, we concentrate on curating high-quality Vietnamese reasoning tasks with verifiable answers. Each instance in the dataset consists of a pair of question-answer instruction $i \in I$ where I represents the space of the instruction problems. The objective is to generate both a final answer $a \in A$ and a corresponding reasoning chain $r \in R$. We define a reasoning chain r as a structured sequence of intermediate steps $\{s_1, s_2, ..., s_n\}$ where each step s_i constitutes a logical deduction that incrementally bridges the initial question to the final answer. To enrich factual correctness and coverage, we also retrieve supplementary context $c \in C$ from the web using Google Search[9]. This additional context serves to enhance both the precision and depth of the model's reasoning.

Formally, the reasoning process can be modeled as a function:

$$f : I \times C \rightarrow R \times A$$

[9]https://google.com

3.2 Instruction Selection

To ensure the robustness and generalizability of the reasoning capabilities of GreenMind, we design a multi-stage pipeline for selecting and curating instruction problems tailored for Vietnamese logical reasoning. Our selection process emphasizes linguistic diversity, logical depth, and cultural relevance. Specifically, we adopt the following criteria for instruction selection:

- **Task Type Diversity:** We include a broad range of reasoning tasks such as arithmetic word problems, commonsense inference, symbolic logic, deductive and inductive reasoning, multi-hop question answering, and ethical dilemma evaluation.
- **Linguistic Complexity:** Instructions are sampled across varying syntactic and lexical complexities to challenge the model's understanding of nuanced Vietnamese expressions.
- **Reasoning Depth:** We prioritize tasks that require multi-step deductions, analogical thinking, and counterfactual reasoning over those solvable with shallow pattern matching.
- **Verifiability:** Each instruction-answer pair is manually verified or derived from trusted Vietnamese educational and encyclopedic sources, ensuring factual accuracy and clarity in logical steps.

3.3 Reasoning Chain Generation

Beyond high-quality instructions, the generation of structured, verifiable reasoning chains is essential for training large language models capable of logical inference and multi-step deduction. To curate such high-quality solutions, we adopt a automated pipeline that incorporates web-scale retrieval to ensure factual correctness and logical coherence.

Given an instruction $i \in I$, we first retrieve supplementary context $c \in C$ from the web. This retrieved information often includes relevant definitions, background knowledge, or factual references that are not explicitly included in the instruction. The context c serves as external knowledge to support the reasoning process, especially in tasks requiring factual grounding or domain-specific expertise.

Subsequently, we generate a reasoning chain $r = s_1, s_2, ..., s_n$, where each step s_i is a logically valid and interpretable inference that incrementally

bridges the gap between the given question and the final answer $a \in A$. These steps are structured to reflect a natural flow of thought, ensuring that the reasoning path remains traceable, coherent, and grounded in both the instruction and the retrieved context.

To ensure the quality of the generated reasoning chains, we apply a multi-stage filtering and validation process:

- **Consistency Check:** We verify that the reasoning steps logically lead to the final answer and are internally consistent.
- **Redundancy Elimination:** Duplicate or unnecessary steps are pruned to maintain conciseness without sacrificing interpretability.
- **Format Conformity:** The reasoning chain must follow a step-by-step format to ensure compatibility with chain-of-thought (CoT) supervision.

Moreover, to promote generalization and robustness, we also include examples with multiple valid reasoning chains for the same instruction. This encourages the model to develop a flexible reasoning strategy rather than memorizing fixed templates.

By focusing on both correctness and interpretability, our approach to reasoning chain generation enables GreenMind to demonstrate superior performance in structured reasoning tasks, setting a strong foundation for Vietnamese LLMs with transparent and explainable outputs

4 GreenMind-Medium-14B-R1

In this section, we present the base architecture, provide statistics on the Vietnamese training data, and describe the optimization strategy we used to transform the pretrained model into a Vietnamese-focused reasoning model.

Base Model. We utilize `Qwen2.5-14B-Instruct` (Team, 2024) as a base model for fine-tuning process. `Qwen 2.5-14B-Instruct` is a dense, decoder-only Transformer language model comprising approximately 14.7 billion parameters. The architecture features 48 layers with a hidden state dimensionality of 5,120 and incorporates SwiGLU (Shazeer, 2020) feed-forward blocks alongside RMSNorm (Zhang and Sennrich, 2019) normalization. The model employs Gated-Query Attention (Dhingra et al., 2016) (GQA) with 40 query heads and 8 key-value heads, augmented by Rotary Position Embeddings (RoPE) (Su et al., 2024) combined with YaRN (Peng et al.) scaling to effectively support an extended context window of up to 128,000 tokens, enabling generation of sequences up to 8,192 tokens per request. Pretraining was conducted on an expansive multilingual corpus totaling 18 trillion tokens across more than 29 languages, including Vietnamese, representing a 2.5-fold increase over its version. Subsequently, the model underwent supervised fine-tuning on over one million high-quality instruction-response pairs, followed by staged reinforcement learning-based preference optimization. These design choices collectively enhance the model's capacity for long-context understanding, multilingual comprehension, and instruction-following capabilities, making it well-suited for complex natural language processing tasks including code generation, mathematical reasoning, and structured data interpretation. This instruct model demonstrates strong, state-of-the-art performance across a range of academic and practical benchmarks, often outperforming models of similar or even larger sizes in several key domains, followed by their report (Team, 2024).

Training data. We curated a high-quality Vietnamese instruction dataset with **55,418 samples**, each containing a question, a reasoning chain, and a final answer. To ensure broad generalization, instructions were drawn from diverse domains:

- **Mathematics:** Mathematical problems train the model in symbolic reasoning, structured logic, and step-by-step problem solving, which are foundational for strong STEM-related performance. (OLMo et al., 2024).
- **Cultural:** These instructions cover Vietnamese idioms, proverbs, traditional practices, historical events, and literary references. This domain strengthens the model's ability to interpret language with deep cultural semantics and regional specificity.
- **Legal and Civic Knowledge:** Focused on basic legal concepts and civic education, particularly relevant in localized Vietnamese contexts such as laws, public policy, and social norms.
- **Education and Exams:** Inspired by real-world school and university-level examination formats in Vietnam, fostering academic

problem-solving patterns.

Reward function 1 Format

Require: Completions $\mathcal{C}$, regex of sequence format rg_s, regex of answer format rg_a, list of candidate results l_{ans}, score of completion structure $score_c$, score of answering structure $score_a$, score of answering candidate structure $score_{ac}$

function FORMAT-REWARDS($\mathcal{C}$, rg_s, rg_a, l_{ans}, $score_c$, $score_a$, $score_{ac}$)
  $l_{scores} = []$ ▷ List of scores
  for $i = 1$ to $length(\mathcal{C})$ **do**
    $score \leftarrow 1.0$
    if $not\_match(\mathcal{C}_i, rg_s)$ **then**
      $score \leftarrow score - score_c$
    end if
    $\hat{p} \leftarrow find(\mathcal{C}_i, rg_a)$ ▷ Get predictions
    if $length(\hat{p}) == 1$ **then**
      if $\hat{p} \nsubseteq l_{ans}$ **then**
        $score \leftarrow score - score_{ac}$
      end if
    else
      $score \leftarrow score - score_a$
    end if
    Append $score$ to l_{scores}
  end for
  return l_{scores}
end function

Optimization with reward functions. We fine-tune the model to focus on tasks that require generating concise answers, which involve a step-by-step reasoning process. Following DeepSeek-R1 (Guo et al., 2025), we design two fundamental reward functions.

- Format rewards: Our objective is to ensure that reasoning chains are enclosed within the `<think>...</think>` tags, and that the final answer is enclosed within the `<answer>...</answer>` tags. Among these, we place greater emphasis on the structure of the final answer, as it remains the ultimate goal to be achieved. Details on how the format reward is computed are described in Algorithm 1. To enable smooth reward assignment, we recommend that:

$$\begin{cases} 0.0 < \text{score}_{ac} < \text{score}_a, \text{score}_c < 1.0 \\ \text{score}_c + \text{score}_a = 1.0 \end{cases}$$

- Accuracy rewards: This function is used to assign scores to the generated answers. A prerequisite is that the prediction must be extracted from the completion based on the expected answering structure. For tasks involving multiple choices, we still encourage assigning partial score to help the model capture the scope of possible outcomes, e.g. one of candidates $[A, B, C, D, E]$ (see Algorithm 2).

To the best of our knowledge, there is no specific statistical report on language data distribution mentioned in the papers or technical reports by the authors of the Qwen model family. Therefore, we identify potential biases toward specific languages through empirical observations. The results for base model indicate that the sampling process can effectively handle Vietnamese language, although Chinese characters occasionally appear. Additionally, the reasoning content should be tightly controlled and remain within the scope of the original topic to ensure alignment with the final answer or the list of possible answers. In summary, we propose the design of two additional reward functions to address the above challenges:

- Language rewards: We define a banned character list to penalize undesired language usage. Our goal is to guide the model to generate content in a single language that matches the input query. Therefore, a completion receives a reward if and only if it does not contain any characters from the banned list (see Algorithm 3).

- Semantic similarity rewards: Based on our proposed Vietnamese reasoning dataset, we measure the closeness of completions using a Sentence Transformers-based model. The selection model should be validated to ensure good performance on the specific monolingual setting. With Algorithm 4, the cosine score is preserved if it exceeds a predefined threshold; otherwise, it is set to zero to mitigate the risk of hallucination.

5 Experiment

5.1 Implementation Details

We fine-tune our model with all parameters (full fine-tuning) to ensure optimal performance, allowing the model to fully adapt to the downstream

Reward function 2 Answering

Require: Completions $\mathcal{C}$, regex of answering format rg_a, list of candidate results l_{ans}, score of answering candidates $score_{ac}$, ground truth gt

function ANSWERING-REWARDS($\mathcal{C}$, rg_a, l_{ans}, $score_{ac}$, gt)
  $l_{scores} = []$ ▷ List of scores
  for $i = 1$ to $length(\mathcal{C})$ **do**
    $score \leftarrow 0.0$
    $\hat{p} \leftarrow find(\mathcal{C}_i, rg_a)$ ▷ Get predictions
    if $length(\hat{p}) == 1$ **then**
      if $\hat{p} == gt$ **then**
        $score \leftarrow 1$
      else if $\hat{p} \subset l_{ans}$ **then**
        $score \leftarrow score_{ac}$ ▷ For multiple-choices tasks
      else
        $score \leftarrow 0$
      end if
    else
      $score \leftarrow 0$
    end if
    Append $score$ to l_{scores}
  end for
  return l_{scores}
end function

Reward function 3 Language

Require: Completions $\mathcal{C}$, List of banned letters l_{bl}

function LANGUAGE-REWARDS($\mathcal{C}$, l_{bl})
  $l_{scores} = []$ ▷ List of scores
  for $i = 1$ to $length(\mathcal{C})$ **do**
    $score \leftarrow 1.0$
    if $exist(\mathbf{c} \in \mathcal{C}_\mathbf{i}, l_{bl})$ **then**
      $score \leftarrow 0.0$
    end if
    Append $score$ to l_{scores}
  end for
  return l_{scores}
end function

task and leverage the capacity of all layers for improved generalization. We use DeepSpeed (Rajbhandari et al., 2020) framework to enable efficient large-scale model training by reducing memory footprint and accelerating training throughput. Specifically, we leverage ZeRO Stage 3 to partition optimizer states, gradients, and parameters across GPUs, which allows us to train models that would otherwise exceed device memory limitations. Additionally, mixed-precision training further improves computational efficiency without sacrificing model accuracy. Additionally, we report the configuration of the hyperparameters used during the fine-tuning process, as detailed in Table 1.

Reward function 4 Semantic Similarity of Reasoning Content

Require: Completions $\mathcal{C}$, Sentence Transformers model ST_model, reasoning data rs, similarity threshold ξ

function SS-REWARDS($\mathcal{C}$, ST_model, rs, ξ)
  $l_{scores} = []$ ▷ List of scores
  for $i = 1$ to $length(\mathcal{C})$ **do**
    $score \leftarrow ST\_model(\mathcal{C}_i, rs_i)$
    if $score < \xi$ **then**
      $score \leftarrow 0.0$
    end if
    Append $score$ to l_{scores}
  end for
  return l_{scores}
end function

We utilize 7 GPUs for model fine-tuning and, one GPU for inference by employing the vLLM framework (Kwon et al., 2023). Our objective is to produce completions that are creative while mitigating hallucinations. The hyperparameters for inferencing are presented in the Table 2. All experiments were conducted on 8 H100 GPUs.

5.2 Experimental Results

Finetuning results. We present some basic analysis after fine-tuning for approximately 4 epochs, as reported in Figure 2. The GRPO loss function starts at 0 and gradually increases. The reason is that the Kullback-Leibler divergence approaches infinity as the distributions of π_θ and π_{ref} become more different.

Quantitative Evaluation. Experimental results

Table 1: Training Hyperparameters

Hyperparameter	Value
epochs	4
per_device_train_batch_size	1
gradient_accumulation_steps	8
gradient_checkpointing	true
learning_rate	5.0e-7
lr_scheduler_type	cosine
warmup_ratio	0.03
beta	0.001
max_prompt_length	256
max_completion_length	1024
num_generations	4
use_vllm	true
vllm_gpu_memory_utilization	0.9

Table 2: vLLM Inferencing Hyperparameters

Hyperparameter	Value
Repetition Penalty	1.2
Temperature	0.6
Top-p (nucleus)	0.8
Top-k	4

on the SeaExam multiple-choice dataset (Li et al., 2024) show that our reasoning model outperforms most Southeast Asian languages, as well as the overall average across all languages, when compared to baseline models with significantly larger parameter sizes — including those with up to 70 billion parameters (see Table 3). Notably, these models were evaluated under few-shot prompting settings. On the VLSP 2023 Challenge dataset (Le et al., 2024), our model achieves superior performance over all previously reported models. In particular, `greennode-7b` and `greennode-14b` were trained using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2023), respectively. Regarding the VMLU[10] dataset, at the current time, the only accessible model is `DeepSeek-R1-Distill-Llama-70B`[11]. Results indicate that our model performs slightly better across most topics, including both mathematical and social science domains.

Qualitative results. The qualitative results are presented in the Table 6. These results demonstrate that the model's responses adhere to structural rules of reasoning and outcome formulation. We showcase reasoning chains across various topics, ranging from natural sciences to social sciences. The visualizations illustrate that the model performs a sequence of logical inferences before arriving at the final answer.

[10] https://vmlu.ai

[11] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B

Table 3: SeaExam performance compared to SOTA model

Model	SeaExam-ID	SeaExam-TH	SeaExam-VI	Avg
Meta-Llama-3.1-70B-Instruct	65.8	**70.6**	72.6	69.7
gemma3-27b-it	64.4	67.5	73.1	68.4
Qwen2.5-14B-Instruct	67.6	68.8	73.1	69.8
GreenMind-Medium-14B-R1	**74.36**	69.75	**74.44**	**72.79**

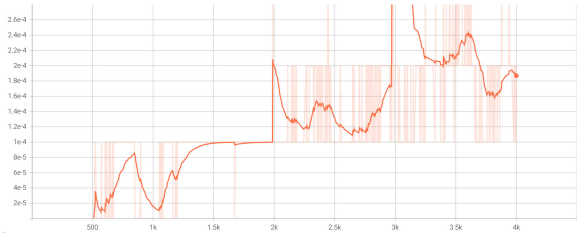

Figure 2: Training Loss.

6 Conclusion

We release **GreenMind-Medium-14B-R1**, a medium-sized Vietnamese language model capable of effectively addressing questions that require intermediate-level reasoning, such as general knowledge and social science topics. By leveraging the GRPO strategy for fine-tuning, we guide the model to generate logically coherent responses. This approach aims to provide users with informative answers, as well as intuitive explanations—valuable not only for end users but also for further research in improving data quality and sampling techniques.

Acknowledgments

We sincerely express our deep appreciation to **GreenNode.ai**[12], our affiliated organization, for their unwavering support throughout the course of this research. GreenNode.ai has played a pivotal role by providing essential resources—most notably, access to high-performance H100 GPUs—which significantly accelerated the fine-tuning process of our models. This generous support was instrumental in the successful development of a Vietnamese reasoning language model.

[12] https://greennode.ai/

Model	Access	STEM	Social Science	Humanities	Others	Avg
VNPTAI.IO-Medium-R1	Private	77.09	82.3	78.85	69.98	77.43
MISA-Llama3-v1.1	Private	77.5	80.75	76.62	71.6	76.87
BnK-AI-Medium-v2	Private	80.94	80.76	70.7	74.06	76.66
VNPTAI.IO-Large-v4	Private	78.05	79.05	75.39	70.37	76.21
GreenNode-xMedium-v1	Private	75.7	81.09	75.25	69.33	75.5
GreenMind-Medium-14B-R1 (Ours)	Weight	76.78	77.36	72.32	69.03	74.29
CakebyVPBank-Large	Private	77.75	78.11	70.38	67.82	73.99
DeepSeek-R1-Distill-Llama-70B	Weight	76.77	76.23	67.98	66.82	72.41

Table 4: VMLU performance compared to fine-tuned models

Model	ComprehensionQA-vi ↑	Exams-vi ↑	LAMBADA-vi ↓	WikiQA-vi ↑	MMLU-vi ↑
cpt-smartbot-13b	0.6633	0.3473	21.9864	0.4455	0.414
ura-llama-13b	0.6556	0.342	17.5614	0.438	0.3973
greennode-7b (prior work)	0.6122	0.2892	189.7782	0.3335	0.387
greennode-14b (prior work)	0.6711	0.3672	29.5967	0.468	0.5281
GreenMind-Medium-14B-R1 (our)	**0.8689**	**0.7796**	**10.7609**	**0.7915**	**0.7124**

Table 5: **VLSP 2023 Challenge.** The performance of our model outperforms most SOTA models.

References

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hoang-Quynh Le, Duy-Cat Can, Khanh-Vinh Nguyen, and Mai-Vu Tran. 2024. Overview of the vlsp 2023 – comom shared task: A data challenge for comparative opinion mining from vietnamese product reviews. *arXiv preprint arXiv:2402.13613*.

Yixuan Li, Xu Tan, Yichong Wang, Zihan Zhang, Longyue Wang, Shuo Wang, Xiaohua Liu, Rui Wang, Jingjing Liu, and Tie-Yan Liu. 2024. Seaexam: Benchmarking large language models for southeast asian languages with human exam questions. *arXiv preprint arXiv:2404.11086*.

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.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Miranda, V, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 OLMO 2 Furious. *arXiv (Cornell University)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. *URL https://arxiv.org/abs/2309.00071*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Com-*

puting, Networking, Storage and Analysis, pages 1–16. IEEE.

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*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-Consistency improves chain of thought reasoning in language models. *arXiv (Cornell University)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain-of-Thought prompting elicits reasoning in large language models. *arXiv (Cornell University)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Task:	**Mathematics**
Prompt:	Hình lập phương có cạnh là 5 m. Vậy thể tích hình đó là: A. $150m^3$ B. $25m^3$ C. $100m^3$ D. $125m^3$
Answer:	<think>Đầu tiên, chúng ta cần nhớ công thức tính thể tích của hình lập phương: Thể tích = a³ (trong đó a là độ dài cạnh của hình lập phương) Trong trường hợp này, cạnh của hình lập phương là 5m. Tiếp theo, áp dụng công thức vào dữ liệu cho sẵn: V = 5³ = 5 × 5 × 5 = 125 m³ Cuối cùng, so sánh kết quả với các lựa chọn được cung cấp: - A. 150 m³ - B. 25 m³ - C. 100 m³ - D. 125 m³ Kết quả đúng nhất chính là 125 m³.</think> <answer>D</answer>

Task:	**Natural Science**
Prompt:	Các cơ quan sinh dục cái và cơ quan sinh dục đực của thực vật có hoa là: A. Đài hoa và nhụy hoa. B. Cánh hoa và bao phấn. C. Nhụy và nhị. D. Cả 3 ý trên.
Answer:	<think>Đầu tiên, chúng ta cần xác định: - Cơ quan sinh dục cái (nữ) ở thực vật có hoa là gì? - Cơ quan sinh dục đực (nam) ở thực vật có hoa là gì? 1. Cơ quan sinh dục cái thường gồm phần noãn hay kén noãn nằm bên dưới. 2. Cơ quan sinh dục đực thường gồm nhị hoa sản xuất tinh trùng dạng phấn hoa. Tiếp theo, phân tích mỗi phương án: A - Đài hoa và nhụy hoa không chính xác vì đây chỉ là bộ phận bảo vệ hoặc phụ trợ. B - Cánh hoa và bao phấn cũng tương tự như trên. C - Nhụy và nhị đúng với mô tả phía trên. D - Tất cả ba đều sai khi xét đến vị trí cụ thể của cơ quan sinh dục. Vậy rõ ràng nhất từ góc nhìn cấu trúc chức năng thì C sẽ phù hợp.</think> <answer>C</answer>

Task:	**Social Science**
Prompt:	Chống bão ở nước ta phải luôn kết hợp với chống A. hạn hán. B. động đất. C. ngập lụt. D. sương muối.
Answer:	<think>Đầu tiên, chúng cần xác định vấn đề chính của câu hỏi: "chống bão". Tiếp theo, phân tích mỗi phương án: - Phương án A: Hạn hán - Đây không liên quan trực tiếp đến hậu quả thường gặp khi có bão. - Phương án B: Động đất - Cũng không liên quan nhiều tới tình hình thời tiết do bão gây nên. - Phương án C: Ngập lụt - Khi xảy ra bão, mưa lớn dẫn đến nguy cơ ngập lụt rất cao. - Phương án D: Sương muối - Thường xuất hiện vào mùa khô hoặc trên vùng biển nhất định, ít liên quan đến bão. Từ đó, rõ ràng phương án phù hợp nhất sẽ là sự kết hợp giữa chống bão và chống ngập lụt vì hai yếu tố này đi cùng nhau trong thực tế.</think> <answer>C</answer>

Table 6: Qualitative results on SeaExam with `GreenMind-Medium-14B-R1`.