VinaLLaMA: LLaMA-based Vietnamese Foundation Model

Anonymous ACL submission

Abstract

In this paper, we present VinaLLaMA, an open-weight, state-of-the-art (SOTA) Large Language Model for the Vietnamese language, built upon LLaMA-2 with an additional 800 billion trained tokens. VinaLLaMA not only demonstrates fluency in Vietnamese but also exhibits a profound understanding of Vietnamese culture. VinaLLaMA-7B-chat, trained on 1 million high-quality synthetic samples, achieves SOTA results on key benchmarks, including VLSP, VMLU, and Vicuna Vietnamese Benchmark, marking a significant advancement in the Vietnamese AI landscape and offering a versatile resource for various applications.

1 Introduction

The surge in Large Language Models (LLMs) such as ChatGPT and GPT-4 has significantly advanced the field of artificial intelligence (AI), particularly in language processing. Although Vietnamese is a low-resource language, Vietnam's AI sector witnessed a notable development with the introduction of several Vietnamese LLMs, including BLOOMZ's Vietcuna, URA-LLaMA, PhoGPT, and dama-2. Inspire by this, we introduce VinaLLaMA, a foundational LLM designed specifically for the Vietnamese language. VinaLLaMA, built on top of LLaMA-2, which is promisingly a contribution to the raise of Vietnamese LLMs.

Embracing the spirit of collaboration and open innovation, we are pleased to announce VinaLLaMA, an open-weight Foundation Language Model and its chat variant. These models are now accessible on HuggingFace, ensuring compatibility with all 'transformers' framework-supported libraries. This not only contributes to the global AI research landscape but also provides a specialized tool for exploring and enhancing Vietnamese language processing.

In this work, we present and implement VinaLLaMA-2.7B and VinaLLaMA-7B, which have remarkable result compared to other Vietnamese large language models. Secondly, we introduce a new method to generate synthetic data for both pretraining and fine-tuning purpose. Furthermore, our models and code are available to public, which can be an effective options for Vietnamese LLM researchers and developers

2 Related Work

The development of large language models (LLMs) began with the Transformer (Vaswani et al., 2017), the foundation architecture that enabled the pretraining of BERT (Cesar et al., 2023) and GPT (Radford et al.) using large-scale unsupervised data, and later are RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020).

GPT-3 (Brown et al., 2020) demonstrated capabilities in few-shot and zero-shot learning. This was followed by significant advancements with ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), which have made substantial contributions. These models exhibit a diverse range of skills and produce high-quality outputs, leading to the rise of newer LLMs such as Llama (Raffel et al., 2020), Llama-2 (Touvron et al., 2023), Bloom (Workshop et al., 2022), Falcon (Almazrouei et al., 2023), Qwen (Bai et al., 2023), and Mistral (Jiang et al., 2023).

The Vietnamese LLM landscape has seen significant advancements with models like Vietcuna (Nguyen et al., 2023d), which underwent further pre-training on BLOOMZ (Workshop et al., 2022), followed by fine-tuning with synthetic dataset. Another model, URA-Llama (Nguyen et al., 2023b), was developed through fine-tuning on a corpus of Vietnamese articles, building upon Llama-2's architecture. DopikAI's ViGPT™ (Nguyen et al., 2023c) also made strides by continuing pre-training and supervised fine-tuning with articles. Additionally, PhoGPT (Nguyen et al., 2023a) represents a new approach, which pre-trained on a 41GB text corpus.

3 VinaLLaMA

3.1 Pretraining

LLaMA-2, a highly regarded pre-trained language model in English, shows a significant gap in handling Vietnamese-related content due to limited relevant tokens. Additionally, its original tokenizer falls short in multilingual applications. To address these issues, we compiled a new dataset combining public and synthetic in-house data. We also selected BKAI's LLaMA-2-chat tokenizer (Lab, 2023) for the tokenizer to enhance the Vietnamese processing.

3.1.1 Public Data

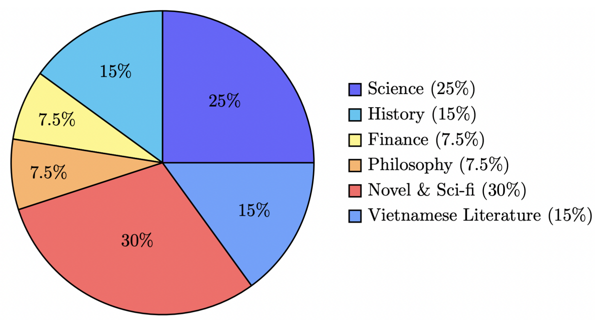

Figure 1: Distribution of Book Topics Used in VinaLLaMA training.

Books. We include nearly 250,000 volumes of Vietnamese literature in othe book dataset. This extensive collection spans various domains, including science, history, finance, and philosophy, as well as fiction genres like novels and science fiction, in addition to traditional Vietnamese literature as shown in figure 1.

Public News. The dataset is derived from two principal Vietnamese news sources, VnExpress [1] and BaoMoi [2]. . To align with ethical guidelines and content appropriateness standards, a filtering process was implemented, systematically excluding articles that contained keywords indicative of harmful or violent content.

CulturaX. Finally, we also include a subset of Vietnamese from CulturaX (Nguyen et al., 2023e) and an additional 100B tokens in English. The final public dataset has a total of roughly 330 billion tokens.

[1] VnExpress: https://vnexpress.net
[2] BaoMoi: https://baomoi.com

3.1.2 In-house Data

Influenced by the concept of synthetic textbooks for pretraining (Gunasekar et al., 2023), our approach incorporates a mechanism that selects random text segments from a publicly available dataset. These segments are then integrated into over 80 bespoke prompt templates to facilitate the rewriting task. The prompts, when fed into GPT-4, result in roughly 100,000 samples. We then use these samples to train with Vietcuna-3B-v2 to generate more than 500B synthetic tokens.

3.2 Supervised Instruction Fine-tuning

We employed proprietary methodologies to create 500,000 Vietnamese synthetic samples. These samples encompassed both instructional and conversational formats. To enhance the dataset, an additional 500,000 English samples were sourced from the OpenHermes-2.5 (teknium, 2023) and Capybara (Daniele and Suphavadeeprasit, 2023) datasets. The comprehensive final dataset encompasses an array of tasks, including reasoning, role-playing, poem writing, coding, function calling, and agent prompting. This diversity ensures a broad scope of capabilities in our model.

3.3 VinaLLaMA-2.7B

Adopting the structured pruning procedure outlined in (Xia et al., 2023), we successfully reduced our model to a more compact variant with 2.7 billion parameters. This process involved strategically pruning the network while retaining its core functional capabilities.

3.4 Training Details

For our pretraining phase, we utilized a cluster consisting of eight nodes, each equipped with 8x Intel Habana Gaudi2 Processors. This phase was completed over the one week. In contrast, the fine-tuning phase was conducted more rapidly, utilizing a single node of Google Cloud TPU v5e, and completed within a single day.

Additionally, our smaller 2.7 billion parameters variant, underwent an easier process. This variant was both continued pre-trained and fully fine-tuned over a period of five days. This process was carried out on a single node, which was provisioned with 8x NVIDIA A100 80GB GPUs.

Model	arc_vi (25-shot)	hellaswag_vi (10-shot)	mmlu_vi (0-shot)	truthfulqa_vi (0-shot)	Average
hoa-7b	0.2855	0.4329	0.2536	0.4542	0.3566
BLOOM-7B	0.3255	0.4830	0.2810	**0.4701**	0.3899
ViGPT™	0.2596	0.3877	0.2482	0.4612	0.3392
PhoGPT-7B5	0.2496	0.2577	0.2474	0.4677	0.3056
VinaLLaMA-2.7B	0.2906	0.4337	0.2490	0.4608	0.3585
VinaLLaMA-7B	**0.3350**	**0.4956**	**0.3168**	0.4552	**0.4007**

Table 1: VLSP Benchmark Scores of Pretrained Models

Model	arc_vi (25-shot)	hellaswag_vi (10-shot)	mmlu_vi (0-shot)	truthfulqa_vi (0-shot)	Average
URA-LLaMA-13B	0.3752	0.4830	**0.3973**	0.4574	0.4282
ViGPT™-170K	0.3651	0.4777	0.3412	0.4691	0.4133
PhoGPT-7B5-Instruct	0.2470	0.2578	0.2413	0.4759	0.3055
Vietcuna-7b-v3	0.3419	0.4939	0.3354	0.4807	0.4130
VinaLLaMA-2.7B-chat	0.3274	0.4814	0.3051	0.4972	0.4028
VinaLLaMA-7B-chat	**0.4239**	**0.5407**	0.3932	**0.5251**	**0.4707**

Table 2: VLSP Benchmark Scores of Supervised Fine-tuning Models

4 Evaluation

In our research, we utilized three distinct evaluation benchmarks: VLSP, VMLU, and the Vicuna Benchmark, with the latter being translated into Vietnamese by VinAI Research [3]. Specifically, we conducted separate experiments for pre-trained and instructional fine-tuning models. This methodology allowed us to compare the baseline capabilities of the pre-trained models against their performance post-instructional fine-tuning.

4.1 VLSP

The first benchmark employed is the VLSP-LLM 2023 (Cuong et al., 2023). It encompasses four distinct assessments: ARC Challenge, HellaSwag, MMLU, and TruthfulQA. This comprehensive benchmark suite allows for a robust evaluation of language models in understanding and generating Vietnamese text across various domains and complexities. The results are reported in Table 1 and Table 2.

In table 1, the VinaLLaMA-7B model demonstrated superior performance compared to other open-source Large Language Models (LLMs) that support Vietnamese. This was evident in its outperformance on three of the four benchmarks, leading to it achieving state-of-the-art results based on the average score across these benchmarks.

In table 2, VinaLLaMA-7B-chat, comparable in scale, astonishingly outperformed larger models, including those with 13 billion parameters, in terms of average score. This impressive achievement highlights the efficiency of the fine-tuning process and the model's ability at leveraging its training to deliver remarkable performance. Adding to this, the VinaLLaMA-2.7B, a significantly smaller model, showcased remarkable performance. It not only competed closely with larger 7B variants but also exceeded PhoGPT-7B5-Instruct. The result of VinaLLaMA-7B-chat, and the noteworthy performance of VinaLLaMA-2.7B, underscore the considerable potential of well-implemented fine-tuning strategies with synthetic data in elevating the capabilities of Large Language Models.

Model	(0-shot)	(5-shot)
ViGPT™	0.2379	0.2769
hoa-7b	0.2513	0.2903
BLOOM-7B	0.2527	0.2312
PhoGPT-7B5	0.2352	0.2325
VinaLLaMA-2.7B	0.2245	0.2688
VinaLLaMA-7B	**0.3199**	**0.3414**

Table 3: VMLU scores of pretrained models

4.2 VMLU

VMLU (AI et al., 2023), a benchmark suite tailored for evaluating foundation models in the Vietnamese language, comprises nearly 10000 MCQs across 58 subjects in domains like STEM, Humanities,

[3] https://www.vinai.io

Model	Coding	CommonSense	Counterfactual	Fermi	Generic	Knowledge	Math	Roleplay	Writing
PhoGPT-7B5-Instruct	0.000	3.600	3.100	0.000	3.500	3.900	0.000	3.800	3.900
ViGPT™-170K	0.500	2.500	1.600	2.400	3.500	3.300	0.000	2.900	2.900
URA-LLaMA-13B	0.714	0.600	0.200	0.000	0.500	0.200	1.333	0.000	0.000
URA-LLaMA-7B	0.000	0.200	1.100	0.000	0.200	0.300	0.000	0.000	0.600
ChatGPT	**4.000**	**4.000**	**4.000**	**3.700**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**
VinaLLaMA-7B-chat	2.714	**4.000**	**4.000**	3.500	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**

Table 4: Average Scores of Different Models Across Tasks in Vicuna Benchmark

Model	(0-shot)	(5-shot)
ViGPT™-170K	0.2419	0.2567
BLOOMZ-7B	0.3945	0.3831
PhoGPT-7B5-Instruct	0.2298	0.2379
Vietcuna-7b-v3	0.3441	0.3199
VinaLLaMA-2.7B-chat	0.2702	0.2567
VinaLLaMA-7B-chat	**0.4046**	**0.4140**

Table 5: VMLU scores of supervised fine-tuning models

and Social Sciences. Its diverse range of difficulty levels tests models from basic knowledge to advanced problem-solving. We conducted our experiments on the validation set of VMLU since the answers to the test set are not publicly available, URA-LLaMA-13B is also not being tested due to the lack of testing time. We reported the results under two few-shot settings: 0-shot and 5-shot, which can be viewed in Table 3 and 5.

In both table 3 and 5, VinaLLaMA-7B-chat achieve the highest score for the average of 58 subjects for both 0-shot and 5-shot. In addition, VinaLLaMA-2.7B-chat outperform ViGPT™-170K and PhoGPT-7B5-Instruct, which is approximately 3 times larger than our models. These promising results reflect the high ability of our model in question answering, logical reasoning and knowledge extraction.

4.3 Vicuna Benchmark Vietnamese

The Vicuna Benchmark (Zheng et al., 2023), translated into Vietnamese by VinAI. This comprehensive benchmark is composed of 80 distinct instructions spanning 9 diverse areas. Uniquely, the evaluation of the results from all participating models is conducted using GPT-4, which introduces an innovative approach to performance assessment. This methodology employs an ELO ranking system, traditionally used in chess and other competitive games, to rate the models.

In our Vicuna Benchmark evaluation, responses from models were assessed using a detailed five-point scale: 0 ('very bad'), 1 ('bad'), 2 ('ok'), 3 ('good'), and 4 ('very good'). This granular scoring system allows for an in-depth evaluation of the quality of each model's response. The final ELO score for each model is computed by aggregating these individual ratings, providing a holistic measure of a model's overall performance across the benchmark's varied tasks.

The benchmark results revealed that VinaLLaMA showcased commendable performance in Vietnamese, closely trailing behind ChatGPT-3.5-Turbo in some benchmarks. This indicates that VinaLLaMA is highly effective in Vietnamese language tasks, with only a slight margin separating it from the more advanced ChatGPT-3.5-Turbo, as shown in Table 4

5 Conclusion

In conclusion, the development of VinaLLaMA marks a significant milestone in the realm of language models, particularly in the context of Vietnamese language processing. Achieving state-of-the-art (SOTA) scores across all Vietnamese benchmarks, VinaLLaMA has demonstrated exceptional proficiency and adaptability. While its performance in English benchmarks was slightly less dominant, it still showed considerable competence, underscoring its effectiveness as a bilingual model.

A key factor in VinaLLaMA's success is the strategic use of carefully crafted synthetic data in its training process. VinaLLaMA's achievements not only set a new standard for language models in Vietnamese but also contribute valuable insights into the broader field of natural language processing.

6 Limitations

VinaLlama-7B-chat exhibits limitations that are widely recognized in the context of Large Language Models (LLMs), including the cessation of knowledge updates subsequent to the pretraining phase, the risk of generating content that may not be factual, such as unsubstantiated advice, and a propensity towards producing inaccuracies or "hallucinations."

This recognition highlights a critical vulnerability, which could potentially render the system susceptible to exploitation by malicious actors. In response to these challenges, forthcoming efforts will focus on enhancing the system's security and dependability. These enhancements will be achieved through the implementation of sophisticated reinforcement learning strategies, anticipated to substantially diminish risks and improve the accuracy of the system's output and decision-making capabilities.

References

Zalo AI, Japan Advanced Institute of Science, and Technology. 2023. Vmlu: A vietnamese multitask language understanding benchmark suite for large language models.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Llinet Benavides Cesar, Miguel-Ángel Manso-Callejo, and Calimanut-Ionut Cira. 2023. Bert (bidirectional encoder representations from transformers) for missing data imputation in solar irradiance time series. *Engineering Proceedings*, 39(1):26.

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

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-instruct: Synthetically generated diverse multi-turn conversations for effecient llm training. *arXiv preprint arXiv:(comming soon)*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

BKAI-HUST Foundation Models Lab. 2023. Llama-2-bk.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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

Duc Q. Nguyen, Sang T. Truong, Toan D. V. Nguyen, Dong D. Le, Nhi N. Truong, and Tho Quan. 2023b. Ura-llama: Universal adapted large language model for vietnamese.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023c. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764, Singapore. Association for Computational Linguistics.

Quan Nguyen, Huy Pham, and Dung Dao. 2023d. Vietcuna.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023e. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

teknium. 2023. teknium/openhermes-2.5-mistral-7b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A HuggingFace OpenLLM Leaderboard (English)

Since our approach in developing VinaLLaMA involved continued pretraining and fine-tuning with both English and Vietnamese data, aiming to establish it as a bilingual large language model. This strategy is key to enhancing its linguistic versatility and adaptability.

We have documented VinaLLaMA's performance metrics on the HuggingFace OpenLLM Leaderboard, providing a comparative analysis with other open-source models. These results, detailed in Table 6, offer insights into VinaLLaMA's standing in the realm of bilingual language models. This information is crucial for understanding its bilingual proficiency and for benchmarking its capabilities against existing models in the field.

The performance of both the VinaLLaMA-7B-chat and 2.7B-chat models was particularly noteworthy, not only achieving the highest overall scores but also showing remarkable strength in mathematical benchmarks, specifically the GSM8K. This highlights their capability in complex problem-solving and analytical reasoning, areas often challenging for language models.

Remarkably, VinaLLaMA-7B-chat exhibited exceptional performance, surpassing even the reinforcement learning-enhanced variants of Meta's LLaMA-2-chat. This achievement is significant, considering the advanced nature of reinforcement learning models and their typically strong performance in such tasks. The success of VinaLLaMA-7B-chat in this regard underscores its advanced capabilities and positions it as a leading model in the domain of mathematics and logic-based problem-solving.

B In-house data

In this section, we will explain more about our approach in generating high quality synthetic data. As explain in the Section 3.1.2, in the first stage, we split our data into different chunks and then feed into over 80 prompt templates. These prompts later are fed into GPT-4 to produce more than 100K synthetic data samples. The methodology and workflow of this process are illustrated in Figure 2.

At the second stage, we employed Vietcuna-3B-v2, our successor smaller-scale language model, to train on the synthetic samples generated in the pretraining step. This training process utilized the rewriting task-specific prompts, which is detailed in Figure 3. Subsequently, we replicated the procedure from Step 1 using this newly trained model. This iteration resulted in the generation of over 500 billion synthetic tokens The final result of this process is more than 500 billion of high-quality Vietnamese tokens ready to be used to continue pretraining on the base LLaMA 2 with the expanded tokenizer.

Model	**arc** (25-shot)	**hellaswag** (10-shot)	**mmlu** (5-shot)	**truthfulqa** (0-shot)	**winogrande** (5-shot)	**GSM8K** (5-shot)	**Average**
LLaMA-2 7B-chat-hf	0.5290	0.7855	0.4832	0.4557	0.7174	0.0735	0.5074
BLOOMZ-7B	0.4078	0.6209	0.3613	0.4522	0.6535	0.0008	0.4161
MPT-7B-chat	0.4625	0.7587	0.3762	0.4056	0.6843	0.0409	0.4547
Nous-Capybara-7B	**0.5520**	**0.7876**	**0.4880**	0.4907	**0.7340**	0.0690	0.5202
VinaLLaMA-2.7B-chat	0.3891	0.6556	0.3323	0.4838	0.5904	0.1350	0.4310
VinaLLaMA-7B-chat	0.5000	0.7293	0.4753	**0.4969**	0.6346	**0.3002**	**0.5227**

Table 6: HuggingFace OpenLLM Leaderboard Benchmark Scores of Supervised Fine-tuning Models

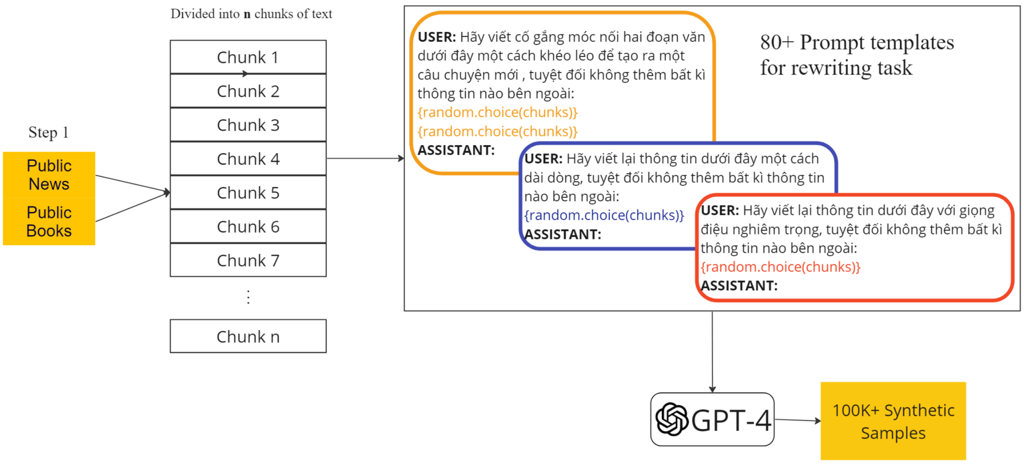

Figure 2: The first stage of generating synthetic data for pretraining

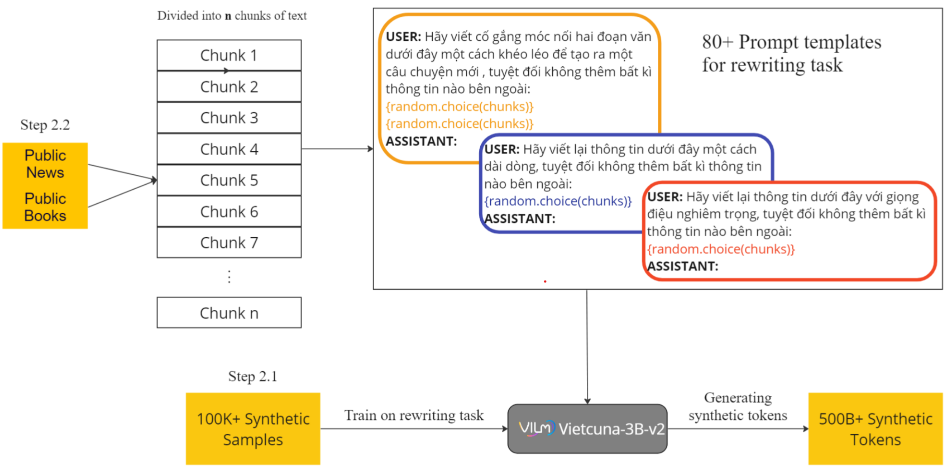

Figure 3: Our second stage of generating synthetic data for pretraining.