VINASMOL: EFFICIENT EXTENSION OF AN ENGLISH-ONLY SLM TO VIETNAMESE
TECHNICAL REPORT (UNRELEASED)

Linh Vu Tu
École polytechnique, LINAGORA
first-name.lastname@polytechnique.edu

ABSTRACT

VinaSmol is a proof-of-concept project that aims to add Vietnamese language capabilities and knowledge to SmolLM2-360M.

Our approach aims to demonstrate that efficient pretraining methods on an open, medium-sized Vietnamese dataset can effectively integrate a new language into an LLM that was not trained on it.

***K*eywords** Small Language Model · Vocabulary extension · Vietnamese

1 Introduction

The advent of Large Language Models (LLMs) and technologies such as Retrieval-Augmented Generation (RAG) have fueled an increasing number of software integrations, such as in search engines and productivity office suites. Given the predominance of English content on the web, most of state-of-the-art LLMs rely heavily on English-first datasets. This data imbalance has significant impacts on model performance and cultural bias.

An increasing number of multilingual LLMs have emerged as an alternative to tackle such issues. For the South-East Asian languages, models such as Sailor [9], SeaLLMs [39, 65] and SEA-LION [33] have settled as strong alternatives to English-first LLMs. Vietnamese has seen similar initiatives with language models including Vietcuna [42], PhoGPT [35] and VinaLlama [36].

However, the majority of these models are based on English-first but multilingual models such as Llama [55, 14], Qwen [54, 59, 43, 60] and Gemma [52, 53]. Such derived LLMs are not open-source under the OSI definition [18] partly due to their derived licenses (Llama [30, 28], Gemma [12]) and because their base models are trained on undisclosed pretraining datasets. This impacts their transparency and their usability. Furthermore, there is a lack of recent Vietnamese foundation models. To the best of our knowledge, PhoGPT, published in 2023, is the most recent Vietnamese foundation model.

In this paper, we conduct a short survey of open-source Vietnamese LLMs in order to better understand the key factors that improve their multilingual capabilities. Furthermore, we introduce VinaSmol, a fully open-source Vietnamese SLM, built on top of SmolLM2-360M [3] using several state-of-the-art techniques for vocabulary extension. VinaSmol starts from zero Vietnamese pretraining tokens, meaning that our vocabulary extension methods may be adapted to other open-source models such as Lucie 7B [13].

2 Related work

2.1 Language specialization

A common method for specializing a model to another language is to continue pretraining and finetune that model on the new language [23]. In order to avoid catastrophic forgetting, a common methodology is to continue pretraining on a mixture of English and of the target languages [9, 33]. Other mitigations include merging monolingual models via

language-specific neuron training as in SeaLLMs [49, 65]. Finally, low-compute approaches use parameter-efficient finetuning (PEFT) such as LoRa [17] and QLoRa [8] as in GemSUra [56].

2.2 Vocabulary extension

Most Vietnamese LLMs are based on a foundation model which already has Vietnamese vocabulary.

However, if the base model is monolingual, the lack of vocabulary of the target language in the base model's tokenizer introduces multiple challenges. First, the tokenized representation of the new language would be inferior, which hurts downstream performance [2]. Second, the tokenization of the new language would be inefficient due to high fertility, which significantly increases the number of training tokens in the corpus.

In such a case, vocabulary extension is crucial in order to improve performance [46] and to cut down on pretraining costs. This brings the question of a good initialization for the new token embeddings. Various initialization strategies have been introduced and surveyed [32].

One method called EEVE [20] focuses on the adaptation of the new token embeddings by emcompassing both tokenizer extension and training.

With VinaSmol, we push vocabulary extension to its limits by choosing SmolLM2-360M-Instruct, a base SLM pretrained on English-only corpora [3]. We test our methods on the most difficult setting in order to prove their efficacy, rather than starting from a multilingual model for easier language specialization.

3 Evaluation of Vietnamese LLMs

There are several synthetic benchmarks for SEA languages such as XQuad, M3Exam [64], SEA-HELM [48] and the ViLLM suite [56, 57], VMLU [63] for Vietnamese. However, the only human evaluation of Vietnamese LLMs we are aware of is a study of LLM performance on Vietnamese literature by neurond [38]. Therefore, we chose to conduct a more empirical evaluation based on real-word LLM usage and human feedback, complementing existing synthetic benchmarks.

3.1 Methodology

We evaluated 5 Vietnamese-speaking LLMs with Arena evaluation [67] on an OpenWebUI [4] frontend. All of the models range between 7B and 9B parameters and were quantized with `Q6_K`.

An Arena model generates two anonymous responses and the user upvotes the preferred answer. This result is incorporated into a list of feedbacks which is used to compute a leaderboard based on ELO scores of each model.

The evaluation involved 15 participants from Linagora Vietnam with 39 feedbacks in total.

3.2 Results

Model	Rating	Won	Lost
Gemma2-9B-IT	**1074**	14	8
Gemma-SEA-LION-v3-9B-IT	1017	9	8
SeaLLM-v3-7B-Chat	983	3	4
Sailor2-8B-Chat	981	16	18
VinaLlama-7B-Chat	929	7	12

Table 1: OpenWebUI leaderboard for 5 Vietnamese LLMs.

Table 1 shows the evaluation results of the selected Vietnamese LLMs. These results can be interpreted in several ways. First, recently-released models perform better than older ones. Furthermore, recent models have more recent base models that are better multilingual abilities. For example, VinaLlama is based on Llama 2 [55], whereas Gemma-SEA-LION-v3 is based on Gemma 2 [52] which exhibits strong multilingual capabilities.

We hypothesize that the initial multilingual abilities of the base model mostly determines the performance of the end model, and that the subsequent continued pretraining mostly unlocks hidden multilingual abilities of the base model.

3.3 Limitations

Due to the small number of participants and feedbacks, and bias due to the considered population, we acknowledge that our results are not statistically significant. However, we believe that this evaluation helped confirm that the synthetic benchmarks results do translate in differences with real-word appreciation of the models.

4 Datasets

In this section, we present the Vietnamese datasets and data preparation methods used for continued pretraining of VinaSmol. Due to limited time and resources, the pretraining dataset of VinaSmol amounts to 2B tokens in total.

4.1 Sources

4.1.1 English datasets

Adding English datasets to the continued pretraining mixture avoids catastrophic forgetting of the acquired language proficiency, capabilities and knowledge of the base model.

For the initial pretraining phase, we use:

- Cosmopedia v2 from the SmolLM Corpus [3]
- FineWebEdu [27]
- Gutenberg-en, only retaining the works in public domain [13]
- Wikipedia-en

Cosmopedia v2 and FineWebEdu are both part of SmolLM's training corpus. Due to the large size of these datasets and the small size of SmolLM2-360M, we inferred that the risk of overfitting was low enough to include them as replay datasets.

4.1.2 Code datasets

Similar to English, we add code datasets in order to preserve SmolLM's performance on programming tasks. As a strategic choice due to limited resources, we choose to focus on Python in order to limit the number of programming languages in the corpus.

Specifically, we use the Python subset of Starcoderdata [25], retaining the examples with quality scores generated by HuggingFace's `python-edu-scorer` higher than 3.

4.1.3 Vietnamese datasets

- Wikipedia-vi
- epfml/FineWeb2-HQ [29]: General, high-quality web dataset.
- CulturaX [37]: General web dataset.
- madlad-400_vi: General web dataset. Clean Vietnamese subset of MADLAD-400 [22]
- Binhvq News Corpus [5]: large Vietnamese news dataset.
- phongmt184172/mtet [34]: Parallel Vietnamese/English pairs.
- doanhieung/vbpl [31]: Vietnam's official law texts.

4.2 Data preparation

Our data preparation pipeline uses datatrove [40]. We take inspiration from Sailcraft [9] for Vietnamese-specific filters.

4.2.1 Formatting

For Wikipedia, we performed superficial formatting by removing some boilerplate sections with regexes.

4.2.2 Normalization

Pretraining web corpora can come with significant noise due to encoding errors and remaining HTML artifacts, which can hurt downstream performance. We fixed text encoding with ftfy [47] and normalized some Unicode punctuation with their canonical equivalents.

4.2.3 Deduplication

Web corpora often contain near document-level duplicates due to overlapping datasets, multiple versions of a same web page. Paragraph-level duplicates can occur very frequently in the form of copyright notices and plagiarized paragraphs. Data deduplication avoids amplifying such content and improves performance [24].

For the document-level deduplication, we use MinHashLSH [6] implemented by Rensa.

We also perform exact substring deduplication, using the deduplicate-text-datasets repository.

4.2.4 Filtering

Data filtering is a necessary step in order to remove low-quality and inadequate content from pretraining datasets. However, the specific parameters and settings depend on the dataset language. Filtering Vietnamese datasets was a surprisingly challenging task due to the lack of appropriate resources for the Vietnamese language.

Due to the specificies of each language, we wrote two distinct `datatrove` pipelines in order to process our Vietnamese and English corpora separately.

Language filtering We use the GlotLID [19] language classifier in order to filter non-English or non-Vietnamese content in the respective datasets.

URL filtering In order to remove most adult content from the datasets, we use URL-based filtering with the default blacklist and domain name-based heuristics of `datatrove` [40]. However, filtering web pages based on certain banned subwords can remove informative content that covers sensitive topics. Therefore, we supply a domain whitelist to avoid filtering on informative websites such as Wikipedia.

Quality filtering We use quality filters from C4 [45], Gopher [44] and FineWeb [41].

Token counting uses the Sailor 2 8B tokenizer, which includes both English and Vietnamese.

Using KenLM [15] models pretrained on OSCAR and Wikipedia, We also remove 90% of documents with a perplexity lower than 20% percentile or higher than 80% percentile.

Explicit content filtering In order to filter adult content, we use a custom list of flagged words from Sailcraft and discard documents with more than 1% of flagged words.

4.2.5 Anonymization

As a preliminary effort of removing personally-identifiable information, we match public IPs and emails with regular expressions and redact them with a list of placeholders.

Due to a lack of accurate PII detectors for Vietnamese, we decided not to use more advanced tools in order to avoid too many false positives.

5 VinaSmol-360M

In this section, we expand on the training of VinaSmol, a Vietnamese SLM based on SmolLM2-360M-Instruct. Our methodology can be divided in four major steps:

- Extend SmolLM2's tokenizer with new Vietnamese tokens.
- Continued pre-training of the model on Vietnamese.
- Instruction fine-tuning.
- Model merging to combine the finetuned checkpoints.

We combine several techniques used for LLMs targeting South-East Asian languages, prioritizing efficiency and results within reasonable compute resources. The goal is to enable replication of the VinaSmol training process on a single node and in reasonable time.

5.1 Tokenizer

As the base SmolLM2 tokenizer has no Vietnamese tokens, vocabulary extension is an important step that reduces pretraining costs. Vietnamese tokenization differs from Latin-centric tokenization since space is not word separator in Vietnamese.

In order to extend the base vocabulary, we train a new tokenizer on our Vietnamese corpora starting with the Vietnamese alphabet, and extend the base SmolLM2 tokenizer with the new tokens. We use a SentencePiece tokenizer, which contrary to BPE does not apply space pre-tokenization.

In order to further refine the list of added tokens, we removed tokens containing symbols, discarded tokens with leading or trailing spaces, and filtered unaccentuated words not present in an 11,000-word Vietnamese wordlist in order to avoid adding new English tokens. Reducing the added vocabulary makes it easier for the model to adapt to the new embeddings.

After vocabulary extension, the tokenizer vocabulary grew from 49152 to 55936. We observed that the resulting tokenizer mostly splits sentences into monosyllabic tokens. This is a significant improvement from the performanc of the base BPE tokenizer, which splits many Vietnamese characters into two bytes.

Caveats SmolLM2's tokenizer is a byte-level BPE but we train a SentencePiece tokenizer on Vietnamese data and merge new new vocabulary into the BPE tokenizer. Therefore, some new multisyllabic tokens are wasted due to space pre-tokenization. Furthermore, adding the merge pairs into the base tokenizer is not trivial and we have not succeeded in doing so.

Therefore, we add the new tokens as added tokens, not as part of the regular vocabulary. Even if we did our best to avoid impacting the original vocabulary, this method can result in slightly worse tokenization of English sequences since some Vietnamese tokens are subwords of English tokens.

5.2 Embedding initialization

After extending the tokenizer with new vocabulary, we resize the embedding layer of SmolLM2-360M and initialize the token embeddings with the average embeddings of their subword tokens [16]. Whenever a meaningful translation is available, we use a convex combination of the initialized embedding and the embedding of their translation using EnViT5-base [34]. A possible improvement would be to transplant the embeddings of another Vietnamese language model using Orthogonal Matching Pursuit [10].

We differ from Kim, Choi, and Jeong [20] since SmolLM2-360M has tied embeddings and no Vietnamese completion capabilities.

5.3 Continued Pretraining

After vocabulary extension, we continued the pretraining of SmolLM2-360M-Instruct on our Vietnamese-English dataset mixture on a single GPU. We chose the `litgpt` [1] for its ease of use and configuration in single-node environments. We refer to the final checkpoint as VinaSmol-base.

5.3.1 Standard training

In our first experiments, we performed normal continued pretraining as a baseline to compare the EEVE pipeline.

Model	Total tokens	Sequence length	Batch size	Learning rate	Warmup steps
SmolLM2-360M	3T	2048*	1024	3e-3	5000
VinaSmol-360M	2B	2048	32	2e-4	50

Table 2: Hyperparameter comparison between SmolLM2 and VinaSmol.

Results. Following the end of continued pretraining, VinaSmol achieves a perplexity of 10 on its dataset. Although such a number may be unsatisfactory, it could be due to the quality of our corpus.

5.3.2 EEVE

We decided to experiment with the multi-stage training from EEVE [20]. Since SmolLM2-360M has tied embeddings, we follow a simplified approach by only performing stages 3, 4, 6, 7 from the original EEVE pipeline. These four steps involve freezing different subsets of the model parameters and are summarized below:

1. Train the new Vietnamese token embeddings.
2. Train all of the embedding layer.
3. Train all of the parameters.
4. Train only the transformer layers.

5.4 Finetuning

We finetune VinaSmol on the following Vietnamese instruction datasets:

1. A Vietnamese translation of Alpaca [50]
2. Alpaca Multi-turn dialogue
3. MURI-IT [21] (Vietnamese)

5.5 Model merging

We merge the Vietnamese fine-tunes with the VinaSmol-base checkpoint and the initial SmolLM2-360M-Instruct with `mergekit` [11]. This approach mitigates catastrophic forgetting and reduces friction between tasks. We perform multi-stage merging with DARE-TIES [62] for the Vietnamese finetunes and DELLA-Linear for the Vietnamese merged finetune and the base models [7].

Caveats. Finetunes and continually pretrained models exhibit different amplitudes of weight alteration [61], which can prevent effective capability transfer without weight disentanglement [61].

Furthermore, the impact of vocabulary extension on the task vectors is an aspect we haven't explored yet.

6 Future work

6.1 Benchmarks

Since our instruct model was unsatisfactory, we did not run benchmarks. However, evaluation is planned in the upcoming weeks.

6.2 Validation of our approach

An ablation study without vocabulary extension would be useful for proving the benefits of vocabulary extension when the base model has no capabilities in the target language, as opposed to [66].

6.2.1 Context length extension

We continued the pretraining of SmolLM2 on sequence lengths of 2048 to save costs. However, this caused the model to lose its context length extension of 8192. Therefore we could rerun context length extension for the final VinaSmol model.

6.2.2 Annealing

Following the approach used by SmolLM2, we should add high-quality content, technical content, textbooks, medium-sized (and instruction) datasets during the annealing phase in order to maximize their impact.

6.2.3 Overcome temporal misalignment

The age difference between pretraining and finetuning datasets results in degraded performance [26] due to misaligned language use and knowledge.

For our cases, the Binhvq news corpus, CulturaX and MADLAD-400 date back from 2023 or older, FineWeb2-HQ was compiled in 2025. A good metric is to report the temporal distribution of the training dataset.

6.2.4 Document-level code-switching

Using a Vietnamese-to-English dictionary for translation, code-switching could help the model align its word representations between English and Vietnamese [9].

Extending our approach to larger models. We hope that our approach could be used to extend larger models to Vietnamese, such as Lucie 7B [13].

7 Conclusion

We conducted a survey of Vietnamese LLMs in order to understand the key drivers of language performance. We hypothesize that the base model has the most impact on downstream performance.

We introduced VinaSmol, a bilingual English-Vietnamese SLM built on top of SmolLM2-360M-Instruct. We hope that VinaSmol serves as an example of SLM vocabulary extension, paving the way for training open-source Vietnamese models with low resources.

References

[1] Lightning AI. *LitGPT*. https://github.com/Lightning-AI/litgpt. 2023.

[2] Mehdi Ali et al. *Tokenizer Choice For LLM Training: Negligible or Crucial?* 2024. arXiv: 2310.08754 [cs.LG]. URL: https://arxiv.org/abs/2310.08754.

[3] Loubna Ben Allal et al. *SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model*. 2025. arXiv: 2502.02737 [cs.CL]. URL: https://arxiv.org/abs/2502.02737.

[4] Anonymous Author(s). *Designing an open-source LLM interface and social platforms for collectively driven LLM evaluation and auditing*. 2024. URL: https://openwebui.com/assets/files/whitepaper.pdf (visited on 12/31/2024).

[5] Vuong Quoc Binh. *Binhvq News Corpus*. 2020. URL: https://github.com/binhvq/news-corpus (visited on 2020).

[6] Andrei Z. Broder. "On the resemblance and containment of documents". In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* (1997), pp. 21–29. URL: https://api.semanticscholar.org/CorpusID:11748509.

[7] Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. *DELLA-Merging: Reducing Interference in Model Merging through Magnitude-Based Sampling*. 2024. arXiv: 2406.11617 [cs.CL]. URL: https://arxiv.org/abs/2406.11617.

[8] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314.

[9] Longxu Dou et al. *Sailor: Open Language Models for South-East Asia*. 2024. arXiv: 2404.03608 [cs.CL]. URL: https://arxiv.org/abs/2404.03608.

[10] Charles Goddard and Fernando Fernandes Neto. *Training-Free Tokenizer Transplantation via Orthogonal Matching Pursuit*. 2025. arXiv: 2506.06607 [cs.CL]. URL: https://arxiv.org/abs/2506.06607.

[11] Charles Goddard et al. "Arcee's MergeKit: A Toolkit for Merging Large Language Models". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Ed. by Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 477–485. DOI: 10.18653/v1/2024.emnlp-industry.36. URL: https://aclanthology.org/2024.emnlp-industry.36.

[12] Google. *Gemma Terms of Use*. 2025. URL: https://ai.google.dev/gemma/terms (visited on 03/24/2025).

[13] Olivier Gouvert et al. *The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation*. 2025. arXiv: 2503.12294 [cs.CL]. URL: https://arxiv.org/abs/2503.12294.

[14] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[15] Kenneth Heafield. "KenLM: Faster and Smaller Language Model Queries". In: *WMT@EMNLP*. 2011. URL: https://api.semanticscholar.org/CorpusID:8313873.

[16] John Hewitt. *Initializing New Word Embeddings for Pretrained Language Models*. Dec. 6, 2021. URL: https://www.cs.columbia.edu/~johnhew//vocab-expansion.html.

[17] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[18] open source initiative. *Open Source AI Definition*. 2024. URL: https://opensource.org/ai/open-source-ai-definition (visited on 10/31/2024).

[19] Amir Kargaran et al. "GlotLID: Language Identification for Low-Resource Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, pp. 6155–6218. DOI: 10.18653/v1/2023.findings-emnlp.410. URL: http://dx.doi.org/10.18653/v1/2023.findings-emnlp.410.

[20] Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. *Efficient and Effective Vocabulary Expansion Towards Multilingual Large Language Models*. 2024. arXiv: 2402.14714 [cs.CL]. URL: https://arxiv.org/abs/2402.14714.

[21] Abdullatif Köksal et al. *MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions*. 2024. arXiv: 2409.12958 [cs.CL]. URL: https://arxiv.org/abs/2409.12958.

[22] Sneha Kudugunta et al. *MADLAD-400: A Multilingual And Document-Level Large Audited Dataset*. 2023. arXiv: 2309.04662 [cs.CL]. URL: https://arxiv.org/abs/2309.04662.

[23] Guillaume Lample and Alexis Conneau. *Cross-lingual Language Model Pretraining*. 2019. arXiv: 1901.07291 [cs.CL]. URL: https://arxiv.org/abs/1901.07291.

[24] Katherine Lee et al. *Deduplicating Training Data Makes Language Models Better*. 2022. arXiv: 2107.06499 [cs.CL]. URL: https://arxiv.org/abs/2107.06499.

[25] Raymond Li et al. *StarCoder: may the source be with you!* 2023. arXiv: 2305.06161 [cs.CL]. URL: https://arxiv.org/abs/2305.06161.

[26] Shayne Longpre et al. *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*. 2023. arXiv: 2305.13169 [cs.CL]. URL: https://arxiv.org/abs/2305.13169.

[27] Anton Lozhkov et al. *FineWeb-Edu: the Finest Collection of Educational Content*. 2024. DOI: 10.57967/hf/2497. URL: https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

[28] Jordan Maris. *Meta's LLaMa license is still not Open Source*. 2025. URL: https://opensource.org/blog/metas-llama-license-is-still-not-open-source (visited on 02/18/2025).

[29] Bettina Messmer, Vinko Sabolcec, and Martin Jaggi. "Enhancing Multilingual LLM Pretraining with Model-Based Data Selection". In: *ArXiv* abs/2502.10361 (2025). URL: https://api.semanticscholar.org/CorpusID:276394897.

[30] Meta. *META LLAMA 3 COMMUNITY LICENSE AGREEMENT*. 2024. URL: https://www.llama.com/llama3/license (visited on 04/18/2024).

[31] Vietnam Ministry of Justice, ed. *The National Database of Legal Documents*. URL: https://vbpl.vn.

[32] Nandini Mundra et al. *An Empirical Comparison of Vocabulary Expansion and Initialization Approaches for Language Models*. 2024. arXiv: 2407.05841 [cs.CL]. URL: https://arxiv.org/abs/2407.05841.

[33] Raymond Ng et al. *SEA-LION: Southeast Asian Languages in One Network*. 2025. arXiv: 2504.05747 [cs.CL]. URL: https://arxiv.org/abs/2504.05747.

[34] Chinh Ngo et al. *MTet: Multi-domain Translation for English and Vietnamese*. 2022. arXiv: 2210.05610 [cs.CL]. URL: https://arxiv.org/abs/2210.05610.

[35] Dat Quoc Nguyen et al. *PhoGPT: Generative Pre-training for Vietnamese*. 2024. arXiv: 2311.02945 [cs.CL]. URL: https://arxiv.org/abs/2311.02945.

[36] Quan Nguyen, Huy Pham, and Dung Dao. *VinaLLaMA: LLaMA-based Vietnamese Foundation Model*. 2023. arXiv: 2312.11011 [cs.CL]. URL: https://arxiv.org/abs/2312.11011.

[37] Thuat Nguyen et al. *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*. 2023. arXiv: 2309.09400 [cs.CL]. URL: https://arxiv.org/abs/2309.09400.

[38] Trinh Nguyen. *A Deep Dive Research to Vietnamese LLMs*. 2024. URL: https://www.neurond.com/blog/vietnamese-llm.

[39] Xuan-Phi Nguyen et al. *SeaLLMs – Large Language Models for Southeast Asia*. 2024. arXiv: 2312.00738 [cs.CL]. URL: https://arxiv.org/abs/2312.00738.

[40] Guilherme Penedo et al. *DataTrove: large scale data processing*. 2024. URL: https://github.com/huggingface/datatrove.

[41] Guilherme Penedo et al. *The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale*. 2024. arXiv: 2406.17557 [cs.CL]. URL: https://arxiv.org/abs/2406.17557.

[42] Huy Pham Quan Nguyen and Dung Dao. *Vietcuna*. 2023. URL: https://huggingface.co/vilm/vietcuna-7b-v3 (visited on 2023).

[43] Qwen et al. *Qwen2.5 Technical Report*. 2025. arXiv: 2412.15115 [cs.CL]. URL: https://arxiv.org/abs/2412.15115.

[44] Jack W. Rae et al. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. 2022. arXiv: 2112.11446 [cs.CL]. URL: https://arxiv.org/abs/2112.11446.

[45] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: 1910.10683 [cs.LG]. URL: https://arxiv.org/abs/1910.10683.

[46] Jean Seo et al. *How does a Language-Specific Tokenizer affect LLMs?* 2025. arXiv: 2502.12560 [cs.CL]. URL: https://arxiv.org/abs/2502.12560.

[47] Robyn Speer. *ftfy*. Zenodo. Version 5.5. 2019. DOI: 10.5281/zenodo.2591652. URL: https://doi.org/10.5281/zenodo.2591652.

[48] Yosephine Susanto et al. *SEA-HELM: Southeast Asian Holistic Evaluation of Language Models*. 2025. arXiv: 2502.14301 [cs.CL]. URL: https://arxiv.org/abs/2502.14301.

[49] Tianyi Tang et al. *Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models*. 2024. arXiv: 2402.16438 [cs.CL]. URL: https://arxiv.org/abs/2402.16438.

[50] Rohan Taori* et al. *Alpaca: A Strong, Replicable Instruction-Following Model*. 2023. URL: https://crfm.stanford.edu/2023/03/13/alpaca.html.

[51] Deep Search Team. *Docling Technical Report*. Tech. rep. Version 1.0.0. Aug. 2024. DOI: 10.48550/arXiv.2408.09869. eprint: 2408.09869. URL: https://arxiv.org/abs/2408.09869.

[52] Gemma Team et al. *Gemma 2: Improving Open Language Models at a Practical Size*. 2024. arXiv: 2408.00118 [cs.CL]. URL: https://arxiv.org/abs/2408.00118.

[53] Gemma Team et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.

[54] Qwen Team. *Introducing Qwen1.5*. 2024. URL: https://qwenlm.github.io/blog/qwen1.5 (visited on 02/04/2023).

[55] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288.

[56] Sang T. Truong et al. *Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models*. 2024. arXiv: 2403.02715 [cs.CL]. URL: https://arxiv.org/abs/2403.02715.

[57] Sang T. Truong et al. *Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models*. 2024. URL: https://ai.stanford.edu/~sttruong/villm (visited on 2024).

[58] *Vietnam Journals OnLine*. URL: https://vjol.info.vn.

[59] An Yang et al. *Qwen2 Technical Report*. 2024. arXiv: 2407.10671 [cs.CL]. URL: https://arxiv.org/abs/2407.10671.

[60] An Yang et al. *Qwen3 Technical Report*. 2025. arXiv: 2505.09388 [cs.CL]. URL: https://arxiv.org/abs/2505.09388.

[61] Le Yu et al. *Extend Model Merging from Fine-Tuned to Pre-Trained Large Language Models via Weight Disentanglement*. 2024. arXiv: 2408.03092 [cs.CL]. URL: https://arxiv.org/abs/2408.03092.

[62] Le Yu et al. *Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch*. 2024. arXiv: 2311.03099 [cs.CL]. URL: https://arxiv.org/abs/2311.03099.

[63] Jaist ZaloAI. *VMLU - A Vietnamese Multitask Language Understanding Benchmark Suite for Large Language Model*. 2023. URL: https://vmlu.ai (visited on 2024).

[64] Wenxuan Zhang et al. *M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models*. 2023. arXiv: 2306.05179 [cs.CL]. URL: https://arxiv.org/abs/2306.05179.

[65] Wenxuan Zhang et al. *SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages*. 2024. arXiv: 2407.19672 [cs.CL]. URL: https://arxiv.org/abs/2407.19672.

[66] Jun Zhao et al. *LLaMA Beyond English: An Empirical Study on Language Capability Transfer*. 2024. arXiv: 2401.01055 [cs.CL]. URL: https://arxiv.org/abs/2401.01055.

[67] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL]. URL: https://arxiv.org/abs/2306.05685.

A Annealing

Due to time constraints, we have not done the annealing phase. This section aims to present some future work for improving our current VinaSmol LLM.

A.1 Annealing datasets

For the annealing phase, we keep 60% of the mixture with the proportions in the initial continued pretraining stage. The remaining 40% consists of the following datasets:

- Wikipedia (English, Vietnamese)
- Binhvq News Corpus
- CCVJ, an in-house dataset of Vietnamese academic papers
- Gutenberg-en
- FineMath 4+
- StackMathQA

A.1.1 CCVJ

We present the CreativeCommons Vietnamese Journals Dataset (CCVJ). We compile CCVJ by downloading 11000+ academic papers from 13 permissively-licensed journals referenced by DOAJ and/or VJOL [58] and processing them with Docling [51].

We used the OAI-PMH endpoints to get a list of referenced papers, including their metadata and their PDF download URLs.

We filtered papers that did not have a PDF and removed papers with incompatible licenses. We ended up with around 10GB of PDF downloads.

Licenses We did our best to find the individual licenses for every paper in the dataset. All of the works in CCVJ satisfy the following conditions:

- The publishing journal operates under a compatible license.
- The paper page specifies a compatible license, if provided.

We refer as "compatible" licenses the following CreativeCommons licenses:

- CC BY 4.0
- CC BY-SA 4.0
- CC BY-NC 4.0
- CC BY-NC-SA 4.0

We refer as "incompatible" the CreativeCommons licenses that disallow derivative works, namely:

- CC BY-ND 4.0
- CC BY-NC-ND 4.0